THE USE OF MONTE CARLO SIMULATIONS IN TEACHING SURVEY SAMPLING

Camelia Goga[1] and Anne Ruiz-Gazen[2]
[1]IMB, Université de Bourgogne, France
[2]Toulouse School of Economics (Gremaq, IMT), France
ruiz@cict.fr

Our objective is to illustrate the use of simulations in the teaching of a graduate course on survey sampling theory. Students are from a master's degree in statistics and have a strong mathematical background. The course consists essentially in theoretical lectures and exercises but it contains also some computer based training which appears to be very helpful for students to understand the theoretical concepts taught. For the computer based training, real populations from official French surveys data base are used and students are asked to carry out some Monte Carlo simulations. Simulations consist in generating a large number of samples according to different sample designs and estimate some finite population parameters according to different estimation methods. Many properties of the sample designs and of the estimation methods can be recovered by using simulations and this point will be illustrated in more detail.

INTRODUCTION

The purpose of the present paper is to explain in detail how we use simulations in our courses on survey sampling for students from master's degrees in statistics (University of Toulouse and University of Franche-Comté). For most of our students, this course is the first and the only one on survey sampling and using computer based training is very helpful to make them better understand the different theoretical concepts. In survey sampling, the focus is on finite population parameters and we adopt a design-based or randomization theory approach in the sense that the statistical inference is based on the sampling randomness while the measured variables are considered as fixed. On the one hand, students are not used to such an approach to inference and they may encounter difficulties at the beginning of the course. On the other hand, we believe that it may be easier to explain a Monte-Carlo experience in the design-based context than in the usual inference context. In the design-based context, the Monte-Carlo experience consists in sampling a part of a given population. Even if we don't have time to enter into the details of the sampling algorithms (which may be complex), students understand more easily the design-based inference when they manipulate data sets on the computer. From our experience it is more difficult to make students understand that a computer can generate values from a Gaussian random variable for instance. Once the design-based approach to sampling inference is better understood by students, we use simulations to illustrate different parts of the survey sampling theory.

In the proposed activities, we make use of two data sets from the INSEE (Institut National de la Statistique et des Etudes Economiques). Both data sets are considered as target populations and the students draw hundreds of samples from these populations. The simulations are made with SAS and R. Since the 8.2 version, SAS incorporates several procedures that are aimed at taking into account complex survey designs and we use these procedures exclusively. In R, we use two packages that have been recently developed: the *sampling* and the *survey* packages. In the following, we propose to use simulations in SAS and R to illustrate the asymptotic properties of the Horvitz-Thompson estimator and compare several sampling designs and several estimation procedures. We also give another possible illustration which we usually don't have time to implement in our courses but which is of interest. Finally, we highlight some limitations of our approach.

THE DATA SETS

Let us give some details concerning the two data sets we use below. The first data set consists in 554 "communes" from the Haute-Garonne department in the south of France. The measured variables for each commune are the total number of housings and the total number of empty housings collected at the 1999 French census. The finite population parameter of interest is the total number of empty housings in the 554 communes (equal to 10,768) and we consider the number of housings as an auxiliary variable.

In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society. Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS8, July, 2010), Ljubljana, Slovenia*. Voorburg, The Netherlands: International Statistical Institute. www.stat.auckland.ac.nz/~iase/publications.php [© 2010 ISI/IASE]

The second data set is from the 2001 French Business survey. For 7104 workers, we observe a dummy variable equal to one when the person works in agriculture, forestry or fishing and equal to zero otherwise. The objective is to estimate the proportion of people working in agriculture, forestry or fishing which is equal to 4.2%. We assume that the category of the communes where the workers live is known. There are five categories of communes: rural, urban with less than 20,000 inhabitants, urban between 20,000 and 200,000 inhabitants, urban with more than 200,000 inhabitants except Paris and Paris. This variable is considered as an auxiliary variable and taken into account in the stratified samplings.

The two data sets enable us to illustrate different properties according to the finite population parameter of interest. For the first data set, the focus is on the estimation of a finite population total for a quantitative variable with large variance while for the second data set, the parameter is a proportion corresponding to the mean of a dummy variable with small variance.

ILLUSTRATION OF ASYMPTOTICS

The first properties we illustrate via simulations are the unbiasedness and the asymptotic normal distribution of the Horvitz-Thompson estimator (HT) under the simple random sampling without replacement (SI). Under the design-based inference, this estimator is unbiased because the average of its values for all possible samples is equal to the finite population parameter. When doing simulations, we generate many possible samples and, thanks to the law of large numbers, we can illustrate the unbiasedness of the HT estimator by averaging the different estimations and compare with the true parameter. On Figure 1, we consider the HT estimator for the total number of empty housings in the Haute-Garonne example for the SI design of size 70. We plot the number of simulations on the x-axis and the corresponding average value of the HT total estimations on the y-axis. The plot illustrates clearly the convergence of the average of the HT estimations to the value 10,768 (horizontal line) which is the total number of empty housings in the 554 communes of Haute-Garonne. On Figure 2, we consider 1,000 samples from a SI design of size 70 and plot the histogram of the HT estimations of the total number of empty housings in the Haute-Garonne example. The students can visualize easily that the sampling distribution of the HT estimator is symmetric and unimodal and looks like a Gaussian distribution which is the assumption we make when defining confidence intervals.

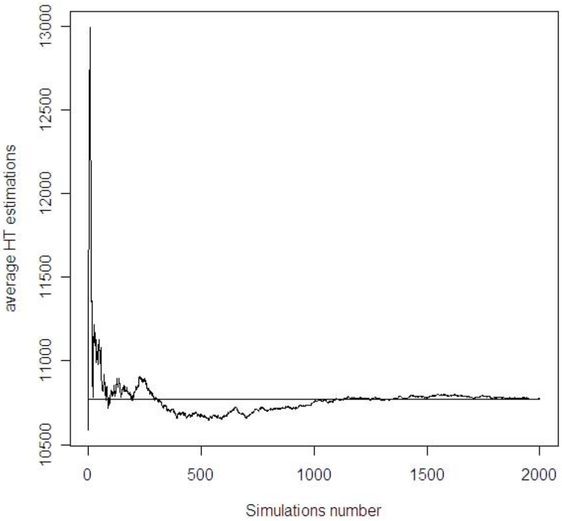

Figure 1. Convergence of the average HT estimations to the true total number of empty housings (horizontal line) for the SI design of size 70 on the Haute-Garonne example

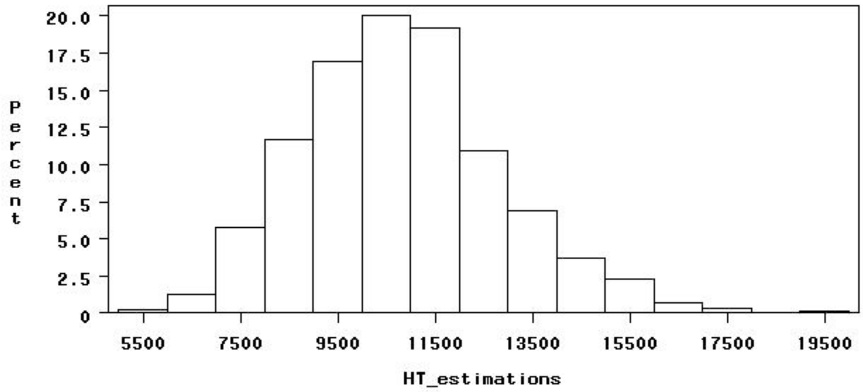

Figure 2. Histogram of 1,000 HT estimations for the SI design of size 70 on the Haute-Garonne example

SURVEY DESIGNS COMPARISON

In order to compare several survey designs, we focus on the HT estimator. We propose to generate numerous samples according to different sampling designs and calculate the average or the empirical mean of the HT estimations together with the coefficients of variation obtained as the ratio of the empirical variance to the empirical mean of the HT estimations. Design effects are also very useful to compare different survey designs in terms of precision of the HT estimator. The design effect is obtained by dividing the empirical variance of the HT estimations under a given design with the empirical variance of the HT estimations under the SI design. As described in Aragon, Haziza and Ruiz-Gazen (2004), other Monte-Carlo indicators may be of interest such as the relative bias in percentage of the true parameter value. Note that in order to compare the different designs, the sample sizes are all taken equal.

The sampling designs we consider are the ones we study in the theoretical lectures, namely, the SI sampling, the stratified SI sampling according to different methods for allocating observations to strata, the proportional to size and with replacement sampling and the one-stage cluster sampling with the SI sampling at the first stage. For the stratified SI sampling and the Haute-Garonne example, several methods for allocating the observations to the strata are considered. The strata are based on the auxiliary variable (the total number of housings) and we have defined four strata (communes with less than 100 housings, between 100 and 300 housings, between 300 and 1000 housings, more than 1000 housings). The simplest way for allocating observations to strata is the proportional allocation where the number of sampled units in each stratum is proportional to the size of the stratum. The design effect of this first stratified design is equal to 70%. The gain over the SI design is more substantial when the allocation of observations is proportional to the total of housings in each stratum with a design effect equal to 40%. The use of an optimal allocation (in the Neyman sense) does not improve much further the results. We conclude that the gain in precision for stratified designs over the SI design can be important when estimating population totals essentially because the variance of the number of empty housings is large. This is no longer true when considering the example on workers from the French Business data set and the estimation of a proportion in general because the variance of a dummy variable in always less than ¼ (see Cochran, p. 109-110 for more details on the gains in precision in stratified sampling for proportions). In the Haute-Garonne example, when sampling proportionally to the size (given by the number of housings) with replacement, the gain in precision over the SI design is similar to the gain obtained by stratifying and allocating observations proportionally to the number of housings (40%). In some sense, the proportional to size design is the continuous version (continuous sampling weights) of the stratified design where the sampling weights are discrete (four strata lead to four different weights). The one-stage cluster sampling is also implemented on the Haute-Garonne example where the primary units are the “Bassins de vie quotidienne” (groups of close communes defined by INSEE). This design which may be of interest for cost reasons provides less precision than the SI design with a design effect equal to 170%.

SAS implementation

In SAS, all the sampling methods we consider are implemented in the recent *proc surveyselect*. The choice between the different possibilities is made easily by changing the option “method=”. Interestingly, there is no need in SAS to define a loop in order to draw several independent samples. We just need to precise the number of replications by the option “rep=”. Interested readers may consult Aragon and Ruiz-Gazen (2004) for more details on the implementation and syntax examples with the Haute-Garonne example.

R implementation

All the previous sampling designs are also implementing in the R package *sampling* (Tillé & Matéi, 2009) which contains many other possible designs. Up to our knowledge, there is no option to obtain several replications of samples.

ESTIMATORS COMPARISON

In order to compare several estimators, we consider the SI design and calculate empirical coefficients of variation for all the estimators. We compare the HT estimator, which does not incorporate any auxiliary information, the poststratified estimator, the ratio and the simple regression estimators which incorporate auxiliary information. From the obtained results (gain of precision when using auxiliary information) the students realize the importance of incorporating auxiliary information at the estimation stage especially if it has not been taken into account at the design stage. The effect of considering an adapted model in the model-assisted approach is also analysed.

SAS implementation

In SAS, the HT estimator is calculated by using the *proc surveymeans* and the option “total” when estimating a total. Concerning estimators that take into account auxiliary information, they are not directly available but they can be obtained via the *proc surveymeans* and *surveyreg* with some manipulations of the “predict” option. Interested readers may consult Aragon and Ruiz-Gazen (2004) for the syntax on the Haute-Garonne example and for more details. This SAS procedure, together with the proc *surveyfreq* and the *proc surveylogistic* that we don’t have time to use in our computer training classes, is adapted for stratified and cluster sampling designs.

R implementation

The R package *survey* (Lumley, 2010) is a very complete package for estimating finite population parameters. In our computer training classes, the population parameters we focus on are totals and means but the *survey* package is also adapted to the estimation of ratios, quantiles,… The functions are called *svytotal, svymean, svyratio, svyquantile*... Moreover and contrary to SAS, the *survey* package contains procedures aimed at calculating estimators that take into account auxiliary information. The functions are called *postStratify* and *calibrate* and they can deal with very general survey designs (through the definition of a survey design object obtained by the *svydesign* function).

ANOTHER ILLUSTRATION

Among other possible illustrations of using simulations in a graduate course on survey sampling, we cite the comparison of several imputation methods as in Haziza and Kuromi (2007). In SAS, Haziza (2002) has proposed a system for imputation simulations that is extremely friendly and useful to illustrate the impact of imputation on estimation for different nonresponse mechanisms. Up to our knowledge, similar functionalities do not exist in R.

LIMITATIONS

The main limitation of simulations is that they may be computationally intensive, especially in R. The reason may be that our programs are not optimized and incorporate loops which are very much time consuming in R, in particular for the calibration procedure.

From our experience, SAS is faster than R but as already noticed, SAS is not really adapted to a survey sampling course. The *proc surveyselect* incorporates the possibility to generate several independent samples very fast by using the option “rep=”. But SAS has several drawbacks in the context of simulations. For example, in order to take into account the stratified sampling in the *proc surveymean*, the sizes of the strata have to be replicated in a file as many times as the number of replications. More generally, the *proc surveyreg* is not aimed at calculating ratio or regression estimators and the way we calculate such estimators with SAS is not easily understandable by students.

CONCLUSION

We have illustrated how simulations on real data sets can help students to understand the theoretical properties of a graduate course on survey sampling. As stated in Stokes (2002), a survey sampling course often has a repetitive “design-estimator-variance” form and students are fond of additional activities similar to the ones we propose in the present paper. But other possible activities such as the ones detailed in Stokes (2002) or Chang, Lohr and McLaren (1992) are also of interest at the graduate level.

REFERENCES

Aragon, Y., Haziza, D., & Ruiz-Gazen, A. (2004). Les simulations dans l’enseignement des sondages avec le logiciel Genesis sous SAS et la bibliothèque Sondages sous R. *La Revue Modulad, 32*, 86-91.

Aragon, Y., & Ruiz-Gazen, A. (2004). Utilisation des procédures SAS dans l’enseignement des sondages. In P. Ardilly (Ed.), *Echantillonnage et méthodes d'enquêtes*. Dunod.

Chang, T., Lohr, S., & McLaren, G. (1992). Teaching survey sampling using simulation. *The American Statistician, 46*, 232–237.

Haziza, D. (2002). GENESIS: Generalized system for imputation simulation. *The Imputation Bulletin, 2(2)*.

Haziza, D., & Kuromi, G. (2007). Handling Item Non Response in Surveys. *CS-BIGS, 1*(2), 102-118.

Lumley, T. (2010). The package “survey”. R contributed package.

Stokes, L. (2002). Use of Mini-Projects in the Teaching of Survey Sampling. *Proceedings of the Sixth International Conference on the Teaching of Statistics (ICOTS6).*

Tillé, Y., & Matei, A. (2009). The package “sampling”, R contributed package.