THE GENERALIZED SIMULATION SYSTEM (GENESIS): A PEDAGOGICAL AND METHODOLOGICAL TOOL

David Haziza, Statistics Canada
Household Survey Methods Division, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6

Key Words: Estimation, Full response, Imputation, Imputation classes, Nonresponse, Simulation studies, System.

1. INTRODUCTION

The Generalized Simulation system (GENESIS) version 1.2, built at Statistics Canada, is a menu driven system based on SAS Release 8. The system consists in SAS macros linked to menus using SAS/AF. GENESIS is simple to use and relatively efficient system in terms of execution time.

The use of GENESIS may be justified from two different angles. First, GENESIS may be used in a survey sampling course by instructors and students in order to illustrate some theoretical concepts in survey sampling such as the choice of a sampling design, choice of an estimator or choice of an allocation method for stratified random sampling. Also, GENESIS may be used by survey practitioners in statistical agencies to help them decide on an imputation strategy for a given survey. For example, it may be useful to answer questions such as: what imputation method to use? How to form the imputation classes? How many classes should we use?

GENESIS contains three main modules:

(1) Full response module
(2) Imputation module
(3) Class module

The first step for the user is to provide the system with a data file in SAS format. This file represents the "population" that will be used as the starting point for the simulations. In GENESIS important results are stored in SAS tables. This gives the user more flexibility in processing the results. For example, the user can easily calculate Monte Carlo measures other than those calculated by default by GENESIS.

In section 2, we describe the full response module and give an example of application in which we compare the Horvitz-Thompson ratio and regression estimators in the case of simple random sampling without replacement. In section 3 we describe the imputation module and give an example in which we compare different imputation method in terms of relative bias of the imputed estimator. The class module is described in section 4. An example illustrating the formation of imputation classes is also presented. Finally, in section 5, we present some work we are planning to do in the future.

2. FULL RESPONSE MODULE

This module is likely to be used in a survey sampling course. Simulation studies are increasingly used in statistical courses as a pedagogical tool. However, in a survey sampling course, it may be difficult for students to build programs for complex sampling designs. GENESIS allows students to investigate several aspects of sampling/estimation for such designs.

Consider a finite population U of size N. The objective is to estimate a population total $Y = \sum_{i \in U} y_i$ of a variable of interest y or a population mean $\bar{Y} = Y/N$. To that end, we select a random sample, s, of size n, according to a given sampling design. In the full response module, several sampling designs are available: simple random sampling without replacement, proportional-to-size sampling with and without replacement, stratified random sampling, Poisson sampling, one-stage and two-stage cluster sampling, two-phase sampling and the Rao-Hartley-Cochran method. In the case of stratified random sampling, it is worth to note that the user may use one of the following allocation methods: optimal allocation, Neyman allocation, proportional allocation and manual allocation. Based on the sampled observations, GENESIS calculates point estimators and their associated variance estimator. For some sampling designs (simple random without replacement, proportional-to-size sampling, stratified random sampling and Poisson sampling), GENESIS computes the Horvitz-Thompson, ratio and regression estimators for population totals and means. In the case of stratified random sampling, GENESIS calculates the separate ratio and regression estimators as well as the combined ratio and regression estimators.

The Horvitz-Thompson estimator of the population total Y is given by

$$\hat{Y}_{HT} = \sum_{i \in s} w_i y_i,$$

where $w_i = 1/\pi_i$ is the survey weight attached to unit i and $\pi_i = P(i \in s)$ is its first-order inclusion

probability. For ratio estimation, we assume that an auxiliary variable z is available for all the sampled units and that the population total $Z = \sum_{i \in U} z_i$ is known. The ratio estimator is then given by

$$\hat{Y}_{ra} = \frac{\hat{Y}_{HT}}{\hat{Z}_{HT}} Z.$$

For regression estimation, we assume that a vector of q auxiliary variables $\mathbf{z} = (z_1, ..., z_q)'$ is available for all the sampled units and that the population total $\mathbf{Z} = \sum_{i \in U} \mathbf{z}_i$ is known. The regression estimator is then given by

$$\hat{Y}_{greg} = \hat{Y}_{HT} + \left(\mathbf{Z} - \hat{\mathbf{Z}}_{HT}\right)' \hat{\mathbf{B}},$$

where $\hat{\mathbf{B}} = \left(\sum_{i \in s} w_i \mathbf{z}_i \mathbf{z}_i' / \sigma_i^2 \right)^{-1} \left(\sum_{i \in s} w_i \mathbf{z}_i y_i / \sigma_i^2 \right)$ and $\hat{\mathbf{Z}}_{HT} = \sum_{i \in s} w_i \mathbf{z}_i$.

After the estimation step, GENESIS calculates several useful Monte Carlo measures such as the relative bias of point and variance estimators, their mean square error and the coverage probability of confidence intervals. Also, GENESIS displays several useful graphics that facilitates the comparisons between the estimators.

Example 1: Suppose we have a population of size N = 1000 containing 2 variables, y and x_1. The correlation between these two variables is approximately equal to 0.78. This population will serve as a basis for the simulation study. The results of a regression analysis involving y and x_1 are shown in Table 1. From this population, we select R = 10000 simple random samples without replacement of size n = 100. In each selected sample, The Horvitz-Thompson, ratio and regression estimators are computed. Finally, GENESIS computes the following Monte Carlo measures:

i) The relative bias (RB) of an estimator $\hat{\theta}$ is given by

$$RB\left(\hat{\theta}\right) = \frac{1}{R} \sum_{i=1}^{R} \frac{\left(\hat{\theta}^{(i)} - \theta\right)}{\theta} \times 100,$$

where $\hat{\theta}^{(i)}$ is a version of $\hat{\theta}$ for the i^{th} replicate, i = 1,…, R.

ii) The mean square error (MSE) of $\hat{\theta}$ is given by

$$MSE\left(\hat{\theta}\right) = \frac{1}{R} \sum_{i=1}^{R} \left(\hat{\theta}^{(i)} - \theta\right)^2.$$

From Table 2, it is clear that all the estimators have a negligible bias. However, both the ratio and regression estimators reduce substantially the MSE in comparison with the Horvitz-Thompson estimator since the variable y is highly correlated with x_1. Also, in this case, the ratio and regression estimators perform almost equally in terms of MSE. GENESIS also displays the distribution of the relative error (figure 1).

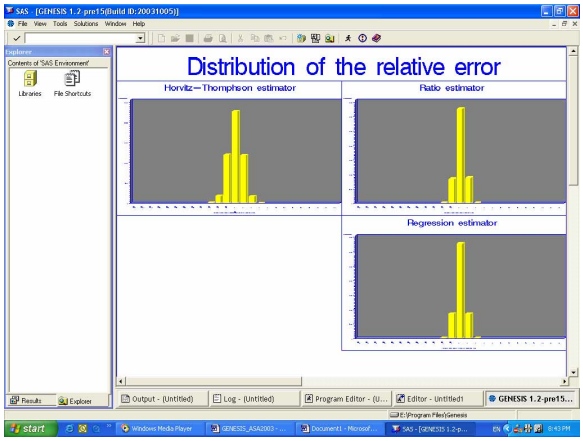
Table 1 : Regression analysis

Variable	Parameter estimates	Standard error	t	Pr > |t|
Intercept	3.30	0.748	4.42	<.0001
x_1	0.90	0.02	37.3	<.0001

Table 2 : Relative bias (in %) and MSE of the estimators

	Horvitz-Thompson	Ratio	Regression
RB (%)	-0.001	-0.005	-0.017
MSE	212456	85210	84451

Figure 1: Distribution of the relative error



3. IMPUTATION MODULE

In the imputation module, the system makes it possible to carry out simulation studies to test the performance of imputed estimators (and, in some cases, variance estimators) under different scenarios. Once again, the user must provide a population that will serve as a basis or the simulation. GENESIS performs the following steps:

(i) From the population, GENESIS draws simple random samples without replacement .

(ii) In each selected sample, GENESIS generates nonresponse to the variable of

interest according to one of the following three response mechanisms:

- MCAR (Missing Completely At Random): the probability of response is constant
- MAR (Missing At Random): the probability of response depends on one or more auxiliary variables
- NMAR (Not Missing At Random): the probability of response depends on the variable of interest

The user must specify the desired response rate. In the case of the MAR and NMAR mechanisms, the user can also choose to generate the nonresponse so that the probability of response increases or decreases with a function of the auxiliary variables or with the variable of interest.

(iii) To fill the hole, GENESIS performs imputation according to one of the following imputation methods:

- Previous value (or historical) imputation
- Mean imputation
- Ratio imputation
- Regression imputation
- Random hot deck imputation
- Nearest neighbour imputation (for which the user may specify the choice of distance)

Finally, GENESIS calculates an imputed estimator, denoted $\bar{y}_I$, of the population mean $\bar{Y}$, and given by

$$\bar{y}_I = \frac{1}{n}\left[\sum_{i\in s_r} y_i + \sum_{i\in s_m} y_i^*\right], \qquad (1)$$

where s_r is the set of r units that responded to item y, s_m is the set of m units that did not respond to item y ($r+m=n$), and y_i^* is the imputed value created in order to "fill the hole" for the missing value y_i.

(iv) For some imputation methods (mean, ratio, ratio and random hot-deck imputation), GENESIS estimates the variance of the imputed estimator $\bar{y}_I$ using one of the following methods:

- The two-phase approach under the MCAR mechanism (Rao, 1990)
- The two-phase approach based on a model (Särndal, 1992)
- The reverse approach under the MCAR mechanism (Shao and Steel, 1999)
- The reverse approach based on a model (Shao and Steel, 1999)

After the estimation steps, GENESIS calculates several useful Monte Carlo measures such as the relative bias of point and variance estimators and their mean square error.

Example 2: We have generated a population of size $N = 1000$ with two variables y and x_1. The results of a regression analysis involving y and x_1 are shown in Table 3. From this population, GENESIS selects $R = 10000$ simple random samples without replacement, each of size $n = 100$. In each selected sample, GENESIS generates nonresponse to variable y according to the logistic function

$$p_i = \frac{\exp(\gamma_0 + \gamma_1 x_{1i})}{1+\exp(\gamma_0 + \gamma_1 x_{1i})} \ .$$

The parameters γ_0 et γ_1 are chosen for the overall response rate to be approximately equal to 70%. To replace the missing values to variable y, mean, ratio (using x_1) and simple linear regression (using x_1) have been used. In each replicate, GENESIS calculates the imputed estimator (1) . The results are shown in Table 4. It is clear that in the case of mean imputation, the bias of the imputed estimator is not negligible (approximately 6.6%). This result is not surprising since both the probability of response and the variable of interest are correlated with the variable x_1. However, mean imputation does not use x_1 for constructing the imputed values. In other words, the appropriate auxiliary information was not included in the imputation model. In the case of ratio imputation, the relative bias of the imputed estimator is even more substantial (approximately -14%). Although, ratio imputation uses the variable x_1 in the imputation model, it forces the regression line to go through the origin. However, the intercept is highly significant (see Table 3) and is 'naturally discarded' by ratio imputation. The inclusion of the intercept in the imputation model (i.e., simple linear regression imputation) reduces the relative bias to virtually zero.

Table 3 : Regression analysis

Variable	Parameter estimates	Standard error	t	Pr > |t|
Intercept	22.5	0.20	112.2	<.0001
x_1	1.3	0.03	42.3	<.0001

Table 4 : Relative bias (in %) and MSE of the imputed estimator

	Mean	Ratio	Regression
RB (%)	3.99	0.038	-0.098
MSE	1.94	0.31	0.32

4. CLASS MODULE

In practice, it is customary to first form classes and then impute within each class. The primary objective of forming classes is to reduce nonresponse bias. Instead of forming classes, one could impute values directly using a regression model. However, there are at least two reasons for using classes: (1) it is more practical when it is a matter of imputing a number of variables at once, and (2) classes provide a degree of robustness as compared with the use of regression imputation.

We begin by giving a theoretical justification for the formation of imputation classes. Let us consider a finite population of size N and let y be a variable of interest. The objective is to estimate the population mean $\bar{Y}$. To that end, we draw a random sample without replacement of size n. Suppose that the units respond to item y independently of one another such that the response probability for unit i is p_i, $i = 1,...,n$. Mean imputation uses the imputed value $y_i^* = \bar{y}_r = \frac{1}{r}\sum_{i \in s_r} y_i$. In this case, the imputed estimator (1) is biased and the bias is given by

$$Bias(\bar{y}_I) = E(\bar{y}_I) - \bar{Y} \approx \frac{1}{N\bar{P}} \sum_{i \in U} (p_i - \bar{P})(y_i - \bar{Y}), \quad (2)$$

where $\bar{P} = \frac{1}{N}\sum_{i \in U} p_i$ is the mean of the response probabilities in the population. Expression (2) of the bias is equal to 0 if the covariance in the population between variables p and y is zero, which is the case, for example, if all units in the population have the same probability of responding (uniform response mechanism) and/or if the value of the variable of interest is the same for all units in the population. Obviously, these two requirements are very rarely met in practice. For this reason, we will call $\bar{y}_I$ a "non-adjusted" estimator. To reduce nonresponse bias, it is common practice to divide the population into C disjoint imputation classes U_c of size N_c $\left(\bigcup_{c=1}^{C} U_c = U, \sum_{c=1}^{C} N_c = N\right)$, which leads to a corresponding partition of the sample s into classes $s_c = s \cap U_c$ of size n_c $\left(\bigcup_{c=1}^{C} s_c = s, \sum_{c=1}^{C} n_c = n\right)$. We then impute independently within each class, which leads to the "adjusted" imputed estimator based on C classes

$$\bar{y}_{I,c} = \sum_{c=1}^{C} w_c' \bar{y}_c \quad (3)$$

where $w_c' = \frac{n_c}{n}$ is a measure of the relative size of class c and

$$\bar{y}_c = \frac{1}{n_c}\left[\sum_{i \in s_{r_c}} y_i + \sum_{i \in s_{m_c}} y_i^*\right] \quad (4)$$

denotes the imputed estimator for class c, $c = 1,\ldots, C$. In the case of mean imputation within classes, the bias for the adjusted estimator is given by

$$\text{Bias}(\bar{y}_{I,c}) \approx \frac{1}{N}\sum_{c=1}^{C} \bar{P}_c^{-1} \sum_{i \in U_c} (p_i - \bar{P}_c)(y_i - \bar{Y}_c), \quad (5)$$

where $\bar{P}_c = \frac{1}{N_c}\sum_{i \in U_c} p_i$ and $\bar{Y}_c = \frac{1}{N_c}\sum_{i \in U_c} y_i$. The bias in (5) is equal to zero if the covariance between the variables p and y is zero in each class. In practice, it may be possible to meet this requirement by forming imputation classes that are homogeneous with respect to the response probabilities p_i's and/or to the variable of interest y.

Various methods are used in practice to form imputation classes. GENESIS allows the user to test the behaviour of two methods for constructing imputation classes: the method by cross-classification and the score method.

Method by cross-classification: Under this method, the classes are formed through combinations of several categorical variables that are believed to explain well the variable of interest (i.e., the variable being imputed). To ensure stability of the imputed estimator, it is customary to specify the following two constraints:

1. The minimal number of donors in a given class is fixed to k.

2. In a given class, the number of donors must be greater than the number of recipients.

Mean or random hot-deck imputation is then performed in all the classes satisfying the above constraints. Since the initial number of classes is potentially very large, it is likely that a large number of classes will not satisfy the constraints. Hence, some form of collapsing is needed. In GENESIS, the least significant variable is dropped from the list and the classes are defined by combinations of the remaining variables. Once again, mean or random hot-deck imputation is performed in all classes that satisfy the constraints. If, there are still some classes that do not satisfy the constraints, the least significant variable is dropped from the list and so on. This process continues until each recipient has found a donor or the maximum number of variables has been dropped.

Score method: This methods leads to a partition of the sample in such a way that, within classes, units (respondents and nonrespondents) are homogeneous with respect to one of the two scores, response

probabilities p_i or item values y_i. The steps for the formation of classes may then be described as follows:

1. Using auxiliary information available on all sampled units, estimate the response probabilities p_i and predict the variable of interest y_i. Two scores, $\hat{p}_i$ and $\hat{y}_i$, are then available for all units in the sample (respondents and nonrespondents).

2. Choose one or both scores. Using the chosen score(s) partition the sample using either

(i) an equal-quantile method: In this case, the class boundaries are determined by the j/k quantiles of the $\hat{p}_i$ or $\hat{y}_i$ populations, $j = 1,...,k-1$.

or

(ii) a classification algorithm. In GENESIS, the procedure FASTCLUST is used.

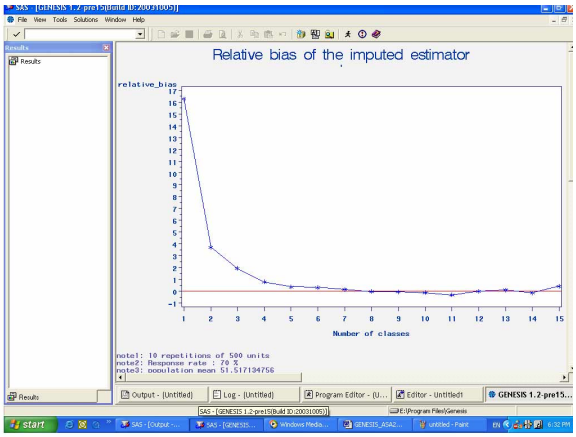
3. Within each class, perform random hot-deck imputation and compute the within-class imputed estimator $\bar{y}_{Ic}$ given in (4).

4. Combine these estimators to get (3).

For both methods, GENESIS provides Monte Carlo measures, such as the relative bias of the imputed estimator or the relative root mean square error (RMSE). For the scores method, GENESIS also provides graphics showing the behaviour of the relative bias and the RMSE when 1, 2,…, C imputation classes are used.

Example 3: We have a population of size $N = 2000$. From this population, GENESIS selects $R = 50$ simple random samples without replacement, each of size $n = 500$. In each selected sample, nonresponse to variable y is generated in a similar fashion to example 2. The overall response rate was set to 70 %. In each sample (with respondents and nonrespondents), GENESIS forms imputation classes according to the score method described using the equal-quantile method and the score $\hat{y}_i$. Note that the score $\hat{y}_i$ is obtained by regression the variable y on and x_1. In each imputation class, mean imputation is used. In each sample, GENESIS forms $C = 1, 2,….., 15$ imputation classes. GENESIS calculates the imputed estimator (3). Figure 2 displays the behaviour of the relative bias of the imputed estimator (3) versus the number of classes. It is clear for figure 1 that as the number of classes increases, the relative bias decreases toward zero. Also, note that the relative bias stabilizes after 5 imputation classes.

Figure 2: Behaviour of the relative bias



5. FURURE WORK

In the full response module, we are studying adding the following options/functionalities:

i) Stratified two-stage design.
ii) The jackknife and bootstrap for variance estimation.

In the imputation module, we are studying adding the following options/functionalities:

i) GENESIS considers only the estimation of population means. The estimation of population totals and domain mean is presently under investigation.
ii) The computation of correct confidence intervals that take the nonresponse variance into account.
iii) GENESIS selects only simple random samples without replacement. It would be interesting to add unequal probability sampling design such as proportional-to-size sampling and stratified random sampling. In this case, it would be interesting to have the option of weighted/unweighted imputation for all the imputation methods.
iv) In the variance estimation methods, the use of the Rao-Shao jackknife (Rao and Shao, 1992) and the 'naïve' jackknife is under investigation.

In the class module, we are studying adding the following options/functionalities:

i) GENESIS considers only the estimation of population means. The estimation of population totals and domain mean is presently under investigation.

ii) The estimation of the variance is under investigation.

iii) GENESIS selects only simple random samples without replacement. Once again, it would be interesting to add unequal probability sampling design such as proportional-to-size sampling.

iv) GENESIS forms imputation classes. The formation of reweighing classes is presently under investigation.

REFERENCES

Rao, J. N. K., (1990). Variance estimation under imputation for missing data, *Technical report, Statistics Canada, Ottawa.*

Rao, J. N. K., Shao, J., (1992). Jackknife variance estimation with survey data under hot-deck imputation, *Biometrika*, 79, pp 811-822.

Särndal, C.-E. (1992), "Methods for Estimating the Precision of Survey Estimates when Imputation has been Used", *Survey Methodology*, 18, pp. 241-252.

Shao, J., Steel, P., (1999). Variance estimation for survey data with composite imputation and nonnegligible sampling fractions, *Journal of the American Statistical Association*, 94, pp 254-265.