A SAS® MACRO for Generating a Set of All Possible Samples with Unequal Probabilities without Replacement

Alan Silva, Igor Nascimento
University of Brasilia, Brasilia, Distrito Federal, Brazil.

Abstract

This paper considers listing all possible samples of size *n* with unequal probabilities without replacement in order to find the sample distribution. The main application of that is to estimate the Horvitz-Thompson (**HT**) estimator and possibly to know the shape of its sample distribution to construct confidence intervals. The algorithm computes all possible samples of the population, in contrast with **PROC SURVEYSELECT** which generates any samples of size *n*, but not all possible samples.

Introduction

The **HT** Estimator is the best-known general estimate of the population total for unequal probability sampling without replacement (Cochran, 1977). It is used mainly in finite populations and with small samples. For that, an auxiliary variable as a measure of size is required to permit an efficient estimation.

The main purpose of this paper is to present a SAS® Macro to compute all possible samples of size *n* of a population of size *N* with unequal probabilities without replacement, allowing showing properties as expectation and variance of the estimators and its sample distribution. It is possible to use the **PROC SURVEYSELECT** to select units without replacement and with the probability proportional to size (PPS Method) and then use the **PROC SURVEYMEANS** to estimate the population total, but this method is a little bit different of the proposed by Horvitz and Thompson (1952). Beyond, this work can help the sampling teaching because there are many calculations to compute **HT** estimates and using the algorithm presented here, this job become easier.

The Horvitz-Thompson Estimator

The Horvitz-Thompson estimator (Horvitz and Thompson, 1952) is used in sampling without replacement and it is a generalization of Hansen-Hurwitz (**HH**) estimator. The unbiased **HT** estimator for the population total is:

$$\hat{t}_{HT} = \sum_{i=1}^{n} \frac{y_i}{\pi_i} \qquad (1)$$

where π_i is the probability of the *i*th unit is in the sample. To find π_i is necessary to compute the conditional probability. Let ψ_i= P(select unit *i* on the first draw). In general (Lohr, 1999), P(unit *i* chosen first, unit *j* chosen second) = P(*i*)P(*j* | *i*).

$$P(j)P(i \mid j) = \psi_j \frac{\psi_i}{1-\psi_j} \qquad (2)$$

Note that in sampling without replacement,

$$\psi_i \frac{\psi_j}{1-\psi_i} \neq \psi_j \frac{\psi_i}{1-\psi_j} \qquad (3)$$

and for *n* = 2, π_{ij} is given by

$$\pi_{ij} = \psi_i \frac{\psi_j}{1-\psi_i} + \psi_j \frac{\psi_i}{1-\psi_j} \qquad (4)$$

So, π_i is computed by (Lohr, 1999).

$$\pi_i = \sum_{j=1, j\neq i}^{N} \pi_{ij} \big/ (n-1) \qquad (5)$$

Also, note that $\sum_{i=1}^{N} \pi_i = n$. The variance of **HT** estimator (Horvitz and Thompson, 1952) is given by:

$$V(\hat{t}_\pi) = \sum_{i=1}^{N} \frac{1-\pi_i}{\pi_i} y_i^2 + \sum_{i=1}^{N} \sum_{j\neq i}^{N} \frac{(\pi_{ij} - \pi_i\pi_j)}{\pi_i\pi_j} y_i y_j \qquad (6)$$

where π_{ij} is the probability of *i*th and *j*th units are both in the sample. The second part of (6) is about the covariance between the populations units . Depending of population size *N* much computational resource is required because are necessary C_2^N combinations. The sample estimate of variance is given by (Horvitz and Thompson, 1952):

$$\hat{V}_1(\hat{t}_\pi) = \sum_{i=1}^{N} \frac{1-\pi_i}{\pi_i^2} y_i^2 + \sum_{i=1}^{N} \sum_{j\neq i}^{N} \frac{(\pi_{ij} - \pi_i\pi_j)}{\pi_{ij}\pi_i\pi_j} y_i y_j \qquad (7)$$

However, this estimator can result in negative estimates. Another sample estimator of variance is given by Yates and Grundy (1953) as:

$$\hat{V}_2(\hat{t}_\pi) = \sum_{i=1}^{N} \sum_{j>i}^{N} \frac{(\pi_i\pi_j - \pi_{ij})}{\pi_{ij}} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 \qquad (8)$$

The above formulas for the Horvitz-Thompson estimator look very similar to the formulas that are applying under equal probability sampling. However, the variance of the HT estimator for the population total looks a little bit different and it is not calculated by **PROC SURVEYMEANS**.

SAS® Macro

The SAS® Macro basically uses the **IML** (Interactive Matrix Language) procedure and the parameters of the algorithms are:

```
%HT_General(tab=,*the name of database*
var=,*the name of the variable to be analyzed*
aux=,*the name of the variable to be used in the
probability selection*,
n=*the sample size*,
str=*the name of the variable that constitute a
stratum in a stratified design*)
```

If it will not be specified any value for the parameter "str" then the Simple Random Sampling will be done. The parameter "n" is used to draw a sample of size *n* in all strata.

The macro begins with the generation of all permutations of *n* elements using the program developed by Wicklin (2010). This step is the only outside of **IML** environment, because it was not possible to use recursive functions to generate the permutations inside **IML**. These samples are the key of the **HT** estimator, because as it was seen before, the sample order is very important.

```
data perm&n (drop=i);
 array a{&n};
 do i = 1 to &n; a[i]=i; end;
 do i = 1 to fact(&n);
  call allperm(i, of a[*]);
  output;
 end;
run;
```

The dataset "perm&n", where &n is the sample size and it is resolved to perm1, perm2 and so forth is created with the positions (indices) of the samples. The real samples are created in the dataset "estimates" together with the **HT** estimator, **SRS** estimator, the Probability of selected sample and the variances.

Macro and Function Call

The results presented in the next section were created using the call bellow.

```
/*** GENERAL SAS ****/
data database;
input id sale size srs;
cards;
1 11 100 1
2 20 200 1
3 24 300 1
4 245 1000 1
;
/* Using HT Estimator*/
%HT_general(tab=database,var=sale,aux=size,n=2,str=);
/* Using SRS Estimator*/
%HT_general(tab=database,var=sale,aux=srs,n=2,str=);
```

Illustration

To illustrate the potential of the algorithm, we consider the example presented by Lohr (1999) about supermarkets. A town has four supermarkets, ranging in size from 100 to 1000 m². We want to estimate the total amount of sales in the four stores for last month by sampling two stores. The data are in Table 1.

Store	Size m²	Sales (in Thousands)
A	100	11
B	200	20
C	300	24
D	1000	245
Total	1600	300

Table 1- Sales by Supermarkets Size.

We can see that the Total amount of sales is 300 and considering a sample of size *n* = 2, the joint probability, π_{ij}, and the stores probability π_i are in Table 2. The probability of Stores A and B is in the sample is computed by $\frac{100}{1600}\frac{200}{1500} + \frac{200}{1600}\frac{100}{1500} = 0.0172619$. Also, note that $\sum_{i=1}^{4} \pi_i = 2$ and all possible samples are shown in Table 3. The expected value, which is equal to 300 (an unbiased estimate) is computed by summing the product of the variables "**HT**" and "**Prob**". It is interesting to note that the total number of samples 12 is computed by $C_2^N n! = C_2^4 2!$ and that the probability of Sample 1 is given by $\frac{100}{1600}\frac{200}{1500} = 0.0083$. The variance of **HT** estimator using Equation (6) and for *n* = 2 is equal to 4383.5622. In the same way, the expectation of **VarHT1** and **VarHT2** (Table 3) are both equal to 4383.5622. Again, unbiased estimates. To evaluate the efficiency of **HT** estimator in relation to **SRS** we calculate the design effect, *deff* = 8.51%, knowing that the **SRS** variance also for *n* = 2 is equal to 51496, showing the great efficiency of **HT** estimator.

	A	B	C	D	
A	0	0.01726	0.02692	0.14583	0.19001
B	0.01726	0	0.05563	0.2976	0.37051
C	0.02692	0.05563	0	0.45673	0.53928
D	0.14583	0.2976	0.45673	0	0.90018
	0.19001	0.37051	0.53928	0.90018	2

Table 2: Joint Probability Selection.

Sample	Selection1	Selection2	HT	Prob	VarHT2	VarHT1
1	11	20	111.86	0.00833	47.06	-14691.48
2	20	11	111.86	0.00893	47.06	-14691.48
3	11	24	102.39	0.01250	502.81	-10832.07
4	24	11	102.39	0.01442	502.81	-10832.07
5	11	245	330.05	0.04167	7939.75	4659.30
6	245	11	330.05	0.10417	7939.75	4659.30
7	20	24	98.48	0.02679	232.72	-9705.15
8	24	20	98.48	0.02885	232.72	-9705.15
9	20	245	326.14	0.08929	5744.06	5682.80
10	245	20	326.14	0.20833	5744.06	5682.80
11	24	245	316.67	0.14423	3259.78	6782.82
12	245	24	316.67	0.31250	3259.78	6782.82

Table 3: All Possible Samples

If one desire to compute **SRS** estimates from **HT** estimator, it is enough to let the Size variable all equal to 1. The results are in Tables 4 and 5. Now, the probability of Stores A and B (or any store) is in sample of size *n* = 2 is computed by $\frac{1}{4}\frac{1}{3} + \frac{1}{4}\frac{1}{3} = 0.1666667$, and the probability of any store is in the sample is $\frac{n}{N} = 0.5$. As the probability of the *i*th element drawn is the same, then the probability of any sample is $\frac{1}{4}\frac{1}{3} = 0.0833$ .The variance of **HT** estimator using Equation (6), for n = 2, and the expected values for **VarHT1** and **VarHT2** are all equal to 51496, the same variance of **SRS** estimator shown above.

	A	B	C	D	
A	0	0.01726	0.02692	0.14583	0.19001
B	0.01726	0	0.05563	0.2976	0.37051
C	0.02692	0.05563	0	0.45673	0.53928
D	0.14583	0.2976	0.45673	0	0.90018
	0.19001	0.37051	0.53928	0.90018	2

Table 4: Joint Probability Selection

Sample	Selection1	Selection2	HT	Prob	VarHT2	VarHT1
1	11	20	62	0.083333	162	162
2	20	11	62	0.083333	162	162
3	11	24	70	0.083333	338	338
4	24	11	70	0.083333	338	338
5	11	245	512	0.083333	109512	109512
6	245	11	512	0.083333	109512	109512
7	20	24	88	0.083333	32	32
8	24	20	88	0.083333	32	32
9	20	245	530	0.083333	101250	101250
10	245	20	530	0.083333	101250	101250
11	24	245	538	0.083333	97682	97682
12	245	24	538	0.083333	97682	97682

Table 5: All Possible Samples

Conclusions

The HT estimator reduces considerably the variance of the total estimate in relation to Simple Random Sampling or Stratified Random Sampling. However, to compute its variance is necessary many calculations because the $[Cov(y_i, y_j)]$ and because of the sampling without replacement, which considers the order of the drawn. The algorithm presented here is important beyond to generate all possible samples to verify, computationally, properties as expectation and population variance $V(\hat{t}_\pi)$ instead $\hat{V}(\hat{t}_\pi)$, become easier the sampling teaching, in contrast with the **PROC SURVEYSELECT** which generates any samples of size n. Additionally, it was possible to compute the exact confidence interval for the HT estimator and we have as a first approximation for the probability distribution for the HT estimator, the skew normal.

It is important to note that the HT Estimator is the general case and only it needs to be computationally implemented. However, the macro presented here has a limitation to handle a population only of size less than 170 because it is not possible (lack of memory) to compute the factorial of numbers greater than 170.

References

Cochran, W.G. (1977). *Sampling Techniques*. John Wiley & Sons, 3rd edition.

Durbin, J. (1953). *Some Results in Sampling when the units are selected with unequal probabilities*. Journal of the Royal Statistical Society, B, 15, 262-269.

Hansen, M.H. and Hurwitz, W.N. (1943). *On the theory of sampling from finite population*. The Annals of Mathematical Statistics, pp. 333-362.

Horvitz, D.G. and Thompson, D.J. (1952). *A Generalization of Sampling Without Replacement From a Finite Universe*. Journal of the American Statistical Association, pp. 663-685.

Lohr, S.L. (1999). Sampling: Design and Analysis. Duxbury Press.

Sarndal, C.E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling.* Springer Series in Statistics, 2th edition.

Wicklin, R. (2010). *Generating All Permutations*. http://blogs.sas.com/iml/index.php?/archives/26-Generating-All-Permutations.html.

Yates, F., Grundy, P.M. (1953). *Selection without replacement from within strata with probability proportinal to size*. Journal of the Royal Statistical Society, pp. 253-261.

Contact Information

Name: **Alan Silva**
Company: **University Of Brasilia**
E-mail: **alansilva@unb.br**

Name: **Igor Nascimento**
Company: **FUNCEF**
E-mail: **igorn@funcef.br**