SELECTING A SAMPLE OF SIZE n WITH PPSWOR FROM A FINITE POPULATION

G.H. Choudhry[1]

Let $U = \{1, 2, ..., i, ..., N\}$ be a finite population of N identifiable units. A known "size measure" x_i is associated with unit i; $i = 1, 2, ..., N$. A sampling procedure for selecting a sample of size n $(2<n<N)$ with probability proportional to size (PPS) and without replacement (WOR) from the population is proposed. With this method, the inclusion probability is proportional to size (IPPS) for each unit in the population.

1. INTRODUCTION

Yates and Grundy [4] have considered a selection procedure for selecting n units with PPSWOR where at the first draw one unit is selected with probability proportional to size, and at the second draw one unit is selected with probability proportional to the size from the remaining units and so on. But with this procedure the overall probability of including a unit in the sample is not proportional to its size. Fellegi [1] has proposed a method whereby the probability of selecting a unit is proportional to its size at each of the n successive draws. This is achieved by determining n sets of selection probabilities called "Working Probabilities". Thus the inclusion probability is proportional to size for each of the N units in the population. This method, however, requires cumbersome evaluation of "Working Probabilities" at each draw except the first one.

---

1 G.H. Choudhry, Census and Household Survey Methods Division, Statistics Canada.

In the present method, selection is made by Yates and Grundy [4] scheme at the first (n-1) draws and a set of "Working Probabilities" is determined for selecting a unit at the nth draw, so that the overall probability of including a unit in the sample becomes proportional to the size for each unit in the population. Empirical results show that the efficiency of this method is the same as that of Fellegi's [1] method.

2. SAMPLING PROCEDURE

Define the "Normalized Sizes" p_i proportional to x_i; $i = 1, 2, \ldots, N$ such that $\sum_{i=1}^{N} p_i = 1$, i.e.

$$p_i = \frac{x_i}{\sum_{i=1}^{N} x_i} \qquad i = 1, 2, \ldots, N. \qquad (2.1)$$

Let Π_i denote the probability that unit i is in the sample; then it can be shown that

$$\sum_{i=1}^{N} \Pi_i = n \quad . \qquad (2.2)$$

It is required that the inclusion probability Π_i be proportional to p_i; $i = 1, 2, \ldots, N$. This condition along with (2.1) and (2.2) imply

$$\Pi_i = np_i \qquad i = 1, 2, \ldots, N \qquad (2.3)$$

At the first draw one unit is selected with probability proportional to size, and at the second draw one unit is selected with probability proportional to the size from the remaining ones and so on up to (n-1)th draw. The probability of i_1th unit being selected at the first draw is p_{i_1}, the conditional probability of i_2th unit being selected at the second draw (given that i_1th unit was already selected at the first draw) is equal to

$$\frac{p_{i_2}}{1 - p_{i_1}}$$

etc. The conditional probability of i_{n-1}th unit being selected at (n-1)th draw (given that the units i_1, i_2, ..., i_{n-2} were previously selected) is equal to

$$\frac{p_{i_{n-1}}}{1 - p_{i_1} - p_{i_2} \cdots - p_{i_{n-2}}} .$$

Let q_i; i = 1, 2, ..., N be the set of "Working Probabilities" for selection at the nth draw, then the conditional probability of i_nth unit being selected at the nth draw (given that i_1, i_2, ... i_{n-1} were previously selected) is equal to

$$\frac{q_{i_n}}{1 - q_{i_1} - q_{i_2} \cdots - q_{i_{n-1}}} .$$

Then the overall probability, $\delta_i(k)$, of selecting ith unit at the kth draw is

$$\delta_i(k) = \sum_{(k-1,i)} p_{i_1} \times \frac{p_{i_2}}{1-p_{i_1}} \times \frac{p_{i_3}}{1-p_{i_1}-p_{i_2}} \cdots \times \frac{p_{i_{k-1}}}{1-p_{i_1}-p_{i_2}-\cdots-p_{i_{k-2}}} \times$$

$$\frac{p_i}{1-p_{i_1}-p_{i_2}\cdots-p_{i_{k-1}}} \qquad k = 1, 2, \ldots, n-1 \qquad (2.4)$$

and

$$\delta_i(n) = \sum_{(n-1,i)} p_{i_1} \times \frac{p_{i_2}}{1-p_{i_1}} \times \frac{p_{i_3}}{1-p_{i_1}-p_{i_2}} \times \dots \times \frac{p_{i_{n-1}}}{1-p_{i_1}-p_{i_2}\dots-p_{i_{n-2}}} \times \frac{q_i}{1-q_{i_1}-q_{i_2}\dots-q_{i_{n-1}}} \tag{2.5}$$

where $\sum_{(k-1,i)}$ (as in Fellegi's [1] paper) denotes the summation over all possible ordered (k-1) tuples of $(i_1, i_2, \dots, i_{k-1})$ such that $i_1, i_2, \dots, i_{k-1}$ are different integers between 1 and N, and none of them is equal to i. Then Π_i, the probability that the ith unit is in the sample, is given by

$$\Pi_i = \sum_{k=1}^{n} \delta_i(k)$$

$$= \sum_{k=1}^{n-1} \delta_i(k) + \delta_i(n)$$

where $\delta_i(k)$ for $k = 1, 2, \dots, n-1$ is given by (2.4) and $\delta_i(n)$ is given by (2.5). In the expression for $\delta_i(n)$, q_i; $i = 1, 2, \dots, N$ must be determined so that the condition $\Pi_i = np_i$; $i = 1, 2, \dots N$ is satisfied, i.e.

$$np_i = \sum_{k=1}^{n-1} \delta_i(k) + \sum_{(n-1,i)} p_{i_1} \times \frac{p_{i_2}}{1-p_{i_1}} \times \dots \times \frac{p_{i_{n-1}}}{1-p_{i_1}-p_{i_2}\dots-p_{i_{n-2}}} \times \frac{q_i}{1-q_{i_1}-q_{i_2}\dots-q_{i_{n-1}}} \qquad i = 1, 2, \dots, N$$

$$q_i = \frac{np_i - \sum_{k=1}^{n-1} \delta_i(k)}{\sum_{(n-1,i)} \frac{p_{i_1} p_{i_2} \cdots p_{i_{n-1}}}{(1-p_{i_1})(1-p_{i_2}) \cdots (1-p_{i_1}-p_{i_2} \cdots -p_{i_{n-2}})(1-q_{i_1}-q_{i_2} \cdots -q_{i_{n-1}})}}$$

$$i = 1, 2, \ldots N. \qquad (2.6)$$

The set of "Working Probabilities" q_i; $i = 1, 2, \ldots N$ can be obtained by solving the set of simultaneous non-linear equations given in (2.6), by iterative procedure where the initial value for q_i can be taken as p_i; $i = 1, 2, \ldots, N$.

For n=2, the method is the same as the one given by Fellegi [1], but for $n \geq 3$, Π_i can be made equal to np_i for all i by evaluating only one set of Working Probabilities instead of (n-1) sets of Working Probabilities as in Fellegi's method.

3. CALCULATION OF Π_{ij}

The joint probability of including both the units i and j in the sample; Π_{ij}, $i = 1, 2, \ldots N-1$; $j = i+1, \ldots N$ can be calculated as follows:

Let $\delta_{ij}(k,\ell)$ denote the probability that the unit i was selected at kth draw and the unit j was selected at the ℓth draw, where $k<\ell$. Then $\delta_{ij}(k,\ell)$ is given by

$$\delta_{ij}(k,\ell) = \sum_{(\ell-2,i,j)} p_{i_1} \times \frac{p_{i_1}}{1-p_{i_1}} \times \frac{p_{i_3}}{1-p_{i_1}-p_{i_2}} \times \cdots \times \frac{p_{i_{k-1}}}{1-p_{i_1}-p_{i_2}\cdots-p_{i_{k-2}}} \times$$

$$\frac{p_i}{1-p_{i_1}-p_{i_2}\cdots-p_{i_{k-1}}} \times \frac{p_{i_{k+1}}}{1-p_{i_1}-p_{i_2}\cdots-p_{i_{k-1}}-p_i} \times$$

$$\cdots \times \frac{p_{i_{\ell-1}}}{1-p_{i_1}-p_{i_2}\cdots-p_{i_{k-1}}-p_i-p_{i_{k+1}}\cdots-p_{i_{\ell-2}}} \times$$

$$\cdots \times \frac{p_j}{1-p_{i_1}-p_{i_2}\cdots-p_{i_{k-1}}-p_i-p_{i_{k+1}}\cdots-p_{i_{\ell-1}}} .$$

$k = 1, 2, \ldots, n-2$

$\ell = k+1, \ldots, n-1$

and $\delta_{ij}(k,n)$ is given by

$$\delta_{ij}(k,n) = \sum_{(n-2;i,j)} p_{i_1} \times \frac{p_{i_2}}{1-p_{i_1}} \times \frac{p_{i_3}}{1-p_{i_1}-p_{i_2}} \times \cdots \times \frac{p_{i_{k-1}}}{1-p_{i_1}-p_{i_2}\cdots-p_{i_{k-2}}} \times$$

$$\frac{p_i}{1-p_{i_1}-p_{i_2}\cdots-p_{i_{k-1}}} \times \frac{p_{i_{k+1}}}{1-p_{i_1}-p_{i_2}\cdots-p_{i_{k-1}}-p_i} \times$$

$$\cdots \times \frac{p_{i_{n-1}}}{1-p_{i_1}-p_{i_2}\cdots-p_{i_{k-1}}-p_i-p_{i_{k+1}}\cdots-p_{i_{n-2}}} \times$$

$$\cdots \times \frac{q_j}{1-q_{i_1}-q_{i_2}\cdots-q_{i_{k-1}}-q_i-q_{i_{k+1}}\cdots-q_{i_{n-1}}}$$

$k = 1, 2, \ldots n-1,$

where $\sum_{(\ell-2;i,j)}$ denotes the summation over all possible ordered $(\ell-2)$-tuples of $(i_1, i_2, \ldots i_{k-1}, i_{k+1}, \ldots i_{\ell-1})$ such that $i_1, i_2, \ldots i_{k-1}, i_{k+1}, \ldots, i_{\ell-1}$ are different integers between 1 and N and none of them is equal to i or j. Then Π_{ij}, the probability that the units i and j are both in the sample is given by

$$\Pi_{ij} = \sum_{k=1}^{n-2} \sum_{\ell=k+1}^{n-1} [\delta_{ij}(k,\ell) + \delta_{ji}(k,\ell)] + \sum_{k=1}^{n-1} [\delta_{ij}(k,n) + \delta_{ji}(k,n)]$$

$$= \sum_{k=1}^{n-1} \sum_{\ell=k+1}^{n} [\delta_{ij}(k,\ell) + \delta_{ji}(k,\ell)] \tag{3.1}$$

$i = 1, 2, \ldots, N-1$

$j = i+1, \ldots, N.$

4. ROTATING SAMPLE

Suppose that in a stratum we want to conduct the survey in m_o first stage units (f.s.u.'s) for some specified period of time. This period could be fixed pre-specified or may occur as and when one or more f.s.u.'s get exhausted. In order to accommodate such a rotation scheme, we initially select $n = \sum_{t=0}^{T} m_t$ f.s.u.'s where m_o is as defined above, and m_t are the number of f.s.u.'s needed for rotation at the time period t; $t = 1,2,3, \ldots, T$. At time $t = 0$, take a simple random sample of m_o out of n f.s.u.'s and for the purpose of rotation at time period t, a simple random sample of m_t units is selected from the remaining $n - (m_o + m_1 + \ldots + m_{t-1})$ out of the n initially selected units. Since the original probability of selecting unit i is $\Pi_i = np_i$, and at any given time the conditional probability of a unit being selected (given that it was originally selected in the first stage of sampling) is equal to m_o/n, therefore the unconditional probability, Π_i', that the unit i is in the final sample is

$$\Pi_i' = \Pi_i \times \frac{m_o}{n}$$
$$= np_i \times \frac{m_o}{n}$$
$$= m_o p_i \qquad i = 1, 2, \ldots, N$$

as required.

Similarly, the unconditional probability, Π_{ij}' , that the unit i and j are both in the sample is given by

$$\Pi_{ij}' = \frac{m_o(m_o-1)}{n(n-1)} \quad \Pi_{ij}$$

$$i = 1, 2, \ldots, N-1$$
$$j = i+1, i+2, \ldots, N.$$

where Π_{ij}, is given by (3.1).

In Fellegi's [1] scheme, since the probability of selecting a unit is proportional to the size at each of the successive draws, therefore for a rotating sample, additional f.s.u.'s are selected at the time of rotation.

5. ESTIMATOR FOR THE POPULATION TOTAL AND ITS VARIANCE

Let $s = \{i_1, i_2, \ldots, i_n\}$ denote the n sampled units and y_i be the value of study variable y for unit i in the population; i = 1, 2, ..., N. The unknowr population total $Y = \sum_{i=1}^{N} y_i$ is to be estimated from the observations y_i for $i \varepsilon s$. Horvitz and Thompson [3] estimator for the population total Y is

$$\hat{Y} = \frac{1}{n} \sum_{i \varepsilon s} \frac{y_i}{P_i} \tag{5.1}$$

and the variance of $\hat{Y}$ as given by Yates and Grundy [4] is

$$V(\hat{Y}) = \frac{1}{n^2} \sum_{i<j}\sum (\Pi_i \Pi_j - \Pi_{ij}) \left(\frac{y_i}{p_i} - \frac{y_j}{p_j}\right)^2 \tag{5.2}$$

where Π_i is the probability that the unit i is in the sample and Π_{ij} is the probability that both the units i and j are in the sample.

An unbiased estimator of $V(\hat{Y})$ is

$$\hat{V} = \frac{1}{n^2} \sum_{\substack{i<j \\ (i,j\varepsilon s)}}\sum \left(\frac{\Pi_i \Pi_j - \Pi_{ij}}{\Pi_{ij}}\right) \left(\frac{y_i}{p_i} - \frac{y_j}{p_j}\right)^2 . \tag{5.3}$$

In the following section, the results of an empirical study using data from Fellegi [1] and Gray [2] have been presented. The Π_{ij} values have been tabulated for Fellegi's [1] method and the proposed method for samples of size 3 and 4. The non-negativity of the variance estimator can be checked from the tabulated Π_{ij} values, i.e. $\Pi_{ij} < \Pi_i \Pi_j$ for all (i,j) pairs in the population. Variances of $\hat{Y}$ and variances of $\hat{V}$ have been computed for the two methods for samples of size 3 and 4 using the two sets of data, i.e. Fellegi [1] and Gray [2].

6. EMPIRICAL RESULTS

6.1 Example 1 Data from Fellegi [1].

The population consists of six primary sampling units. The p_i and y_i values are given in Table (6.1.1).

Table (6.1.1): p_i and y_i Values for Example (1).

i	1	2	3	4	5	6
p_i	0.10	0.14	0.17	0.18	0.19	0.22
y_i	0.60	0.98	1.53	2.16	2.85	4.18

$$Y = \sum_{i=1}^{N} y_i = 12.30$$

The "Working probabilities" for selecting a sample of size 3 are given in Table (6.1.2) for the two schemes. The "Number of iterations" column is the number of iterations it took to obtain the convergence* to the solution. Note that $p_i(k)$; $i = 1, 2, \ldots, N$ are the "Working probabilities" at the kth draw; $k = 1, 2, \ldots, n$ for Fellegi's [1] scheme, where $p_i(1) = p_i$; $i = 1, 2, \ldots, N$. Further q_i; $i = 1, 2, \ldots, N$ are the "Working probabilities" at the nth draw for the proposed scheme. Recall that p_i; $i = 1, 2, \ldots, N$ are the "Working probabilities" at each of the first n-1 draws for the proposed scheme.

---

* The iteration procedure was terminated when the change in the value of each of the elements of the probability vector was less than or equal to 1.0E-8 in magnitude.

Table (6.1.2): "Working Probabilities" for Selecting 3 units.

No. of iterations	i	1	2	3	4	5	6
-	p_i	0.100000	0.140000	0.170000	0.180000	0.190000	0.220000
-	$p_i(1)$	0.100000	0.140000	0.170000	0.180000	0.190000	0.220000
6	$p_i(2)$	0.090190	0.132410	0.167577	0.180157	0.193252	0.236414
9	$p_i(3)$	0.076367	0.119154	0.160684	0.177404	0.196167	0.270224
9	q_i	0.068868	0.113368	0.158820	0.177494	0.198613	0.282837

From the above table we notice that for Fellegi's [1] scheme, the "Working probabilities" for units 1, 2, and 3, which are the three smallest units in the population, decrease during successive draws, whereas for units 4, 5, and 6, which are the three largest units in the population, the "Working probabilities" increase during successive draws. Since for the proposed scheme, the "Working probabilities" at the first $n-1$ draws remain unchanged; therefore at the nth draw (3rd draw in this case) the "Working probabilities" for units 1, 2, and 3 i.e. the three smallest units, are smaller than the corresponding "Working probabilities" for Fellegi's [1] scheme, and for units 4, 5, and 6 i.e. three largest units, the "Working probabilities" are larger than the corresponding "Working probabilities" for Fellegi's [1] scheme.

The following table exhibits the values of Π_{ij} for the two schemes for sample size 3. The values above the main diagonal correspond to Fellegi's [1] scheme and those under the main diagonal correspond to the proposed scheme.

Table (6.1.3): Π_{ij} Values for Sample Size 3.

i j	1	2	3	4	5	6
1	X	0.086165	0.109898	0.118557	0.127675	0.157705
2	0.086163	X	0.161733	0.174315	0.187513	0.230274
3	0.109902	0.161742	X	0.221121	0.237550	0.289698
4	0.118557	0.174316	0.221111	X	0.255473	0.310534
5	0.127674	0.187510	0.237547	0.255479	X	0.331790
6	0.157704	0.230269	0.289699	0.310538	0.331791	X

From the above table it is seen that the Π_{ij} values for the two schemes do not differ up to 4 decimals, and since the variance is a function of Π_{ij} values, therefore, the two schemes will be equally efficient as seen from the following table.

Table (6.1.4): Variance of $\hat{Y}$ and Variance of $\hat{V}$ for the Two Schemes for Sample Size 3.

Selection Scheme	$V(\hat{Y})$	$V(\hat{V})$
Fellegi's Scheme	3.8258	4.6166
Proposed Scheme	3.8259	4.6171

Similarly for a sample of size 4, the following tables give the "Working Probabilities", the Π_{ij} values, and variance of $\hat{Y}$ and variance of $\hat{V}$ for the two schemes.

Table (6.1.5): "Working Probabilities" for Selecting 4 units.

No. of iterations	i	1	2	3	4	5	6
-	p_i	0.100000	0.140000	0.170000	0.180000	0.190000	0.220000
-	$p_i(1)$	0.100000	0.140000	0.170000	0.180000	0.190000	0.220000
6	$p_i(2)$	0.090190	0.132410	0.167577	0.180157	0.193252	0.236414
9	$p_i(3)$	0.076367	0.119154	0.160684	0.177404	0.196167	0.270224
16	$p_i(4)$	0.051667	0.086649	0.130509	0.153692	0.184616	0.392867
17	q_i	0.033017	0.070222	0.121849	0.150012	0.187892	0.437008

Table (6.1.6): Π_{ij} Values for Sample Size 4.

i j	1	2	3	4	5	6
1	X	0.166245	0.216290	0.235994	0.256805	0.324666
2	0.167197	X	0.320554	0.348706	0.377519	0.466976
3	0.216261	0.320123	X	0.445971	0.478668	0.578517
4	0.235761	0.348406	0.446103	X	0.513248	0.616081
5	0.256432	0.377327	0.478869	0.513522	X	0.653760
6	0.324349	0.466948	0.578645	0.616208	0.653850	X

Table (6.1.7): Variance of $\hat{Y}$ and Variance of $\hat{V}$ for the Two Schemes for Sample Size 4.

Selection Scheme	$V(\hat{Y})$	$V(\hat{V})$
Fellegi's Scheme	1.5323	0.4672
Proposed Scheme	1.5269	0.4553

6.2 Example 2: Data from Gray [2].

The population in this example is a stratum in Nova Scotia in LFS with dummy characteristics. The stratum consists of ten primary sampling units. The p_i and y_i values are given in Table (6.2.1).

Table (6.2.1): p_i and y_i Values for Example (2).

i	1	2	3	4	5	6	7	8	9	10
p_i	0.0957	0.1043	0.1043	0.1006	0.0896	0.0881	0.0986	0.1055	0.1149	0.0984
y_i	10.06	10.35	10.38	9.57	9.30	8.96	10.00	10.50	11.33	9.55

$$Y = \sum_{i=1}^{N} y_i = 100.00$$

As in example (1), the "Working probabilities", the Π_{ij} values, and variance of $\hat{Y}$ and variance of $\hat{V}$ for the two schemes for samples of size 3 and 4 were computed from the data in table (6.2.1) above. The behaviour of the "Working probabilities" was similar to those in example (1), and the Π_{ij} values for the two schemes were identical to 5 decimals both for samples of size 3 and 4. Due to space tables of "Working probabilities" and those of Π_{ij} values are not given. In the following table "Number of iterations" required to obtain the "Working probabilities", and variance of $\hat{Y}$ and variance of $\hat{V}$ for samples of size 3 and 4 are given.

Table (6.2.2) "Number of iterations" to obtain "Working Probabilities" and Variance of $\hat{Y}$ and Variance of $\hat{V}$ for the two Schemes for samples of size 3 and 4.

Selection Scheme	Sample Size	No. of iterations at draw 1	2	3	4	$V(\hat{Y})$	$V(\hat{V})$
Fellegi's Scheme	3	-	5	6		2.0509	2.7418
Proposed Scheme	3	-	-	6		2.0508	2.7418
Fellegi's Scheme	4	-	5	6	7	1.3287	0.6647
Proposed Scheme	4	-	-	-	7	1.3287	0.6647

For the two numerical examples in this study, it is observed that the Π_{ij} values for the two selection schemes, ie., Fellegi's [1] scheme and the proposed scheme are almost identical. Although it seems that the underlying design for the two selection schemes is the same, choice between the two should be made on operational convenience. Since Fellegi's scheme requires the evaluation of "Working Probabilities" at each draw except the first one, whereas the scheme proposed in this paper requires the evaluation of "Working Probabilities" at the last draw only, this results in considerable reduction in computing.

ACKNOWLEDGEMENT

The author wishes to thank Mr. G.B. Gray and Dr. M.P. Singh, Census and Household Survey Methods Division, Statistics Canada, and the referee for helpful comments.

RESUME

Soit U = {1, 2, 3, ..., i, ..., N} une population finie de N unités indentifiables. Une "mesure de la taille" connue x_i est associée à l'unité i, i = 1, 2, ..., N.

L'auteur propose une méthode d'échantillonnage pour choisir une taille d'échantillon n (2<n<N) dont la probabilité est proportionnelle à la taille et sans remise. De cette façon la probabilité d'inclusion est proportionnelle à la taille pour chaque unité de la population.

REFERENCES

[1] Fellegi, I.P., "Sampling With Varying Probabilities Without Replacement: Rotating and Non-Rotating Samples", Journal of the American Statistical Association, Vol. 58 (1963), pp 183-201.

[2] Gray, G.B., "Variance Components and Variance Function", Proceedings of 'Statistics 71' Canada, pp 119-26.

[3] Horvitz, D.G. and Thompson, D.J., "A generalization of Sampling Without Replacement from a Finite Universe", Journal of the American Statistical Association, Vol. 47 (1952), pp 663-85.

[4] Yates, F. and Grundy, P.M., "Selection Without Replacement from Within Strata With Probability Proportional to Size", Journal of the Royal Statistical Society, Series B, Vol. 15 (1953), pp 235-61.