SAMPLING WITH UNEQUAL PROBABILITIES AND WITHOUT REPLACEMENT - A REJECTIVE METHOD

G.H. Choudhry and M.P. Singh[1]

An alternative to the direct selection of sample is suggested, which while retaining the efficiency at the same level simplifies the selection and variance estimation processes in a wide variety of situations. If n^* is the largest feasible ΠPS sample size that can be drawn from a given population of size N, then the proposed method entails selection of m ($=N-n^*$) units using a ΠPS scheme and rejecting these units from the population such that the remainder is a ΠPS sample of n^* units; the final sample of n units is then selected as a subsample from the remainder set. This method for selecting the ΠPS sample can be seen as an analogue of SRS where it is well known that the "unsampled" part of the population as well as any subsample from this part are also SRS from the entire population when SRS is the procedure used. The method is very practical for situations where m is less than the actual sample size n. Moreover, the method has the additional advantage in the context of continuing surveys, eg. Canadian Labour Force Survey (LFS), where the number of primary sampling units (PSU's) may have to be increased (or decreased) subsequent to the initial selection of the sample. The method also has advantages in the case of sample rotation. Main features of the proposed scheme and its limitations are given. Efficiency of the method is also evaluated empirically.

1. INTRODUCTION

Selection of primary sampling units (PSU's in a multi-stage sampling scheme) with unequal probabilities has found wide applications in large scale surveys. However, in many cases either with replacement or one PSU per stratum sampling is used because of their simplicity. The more efficient without replacement sampling schemes (e.g. Fellegi [1963],

---

[1] G.H. Choudhry and M.P. Singh, Census and Household Survey Methods Division, Statistics Canada.

Hartley and Rao [1962], etc.) become quite complex either in terms of the selection procedure or variance estimation, even for moderately large sample sizes. In this article we propose an alternative approach (a rejective method) of selecting the sample which retains the efficiency of direct selection method using any sampling scheme for the purpose of rejecting m units. This method simplifies the sample selection and usually the variance estimation in a wide variety of situations. In addition, the suggested approach has several other operational advantages in the context of large scale continuing surveys such as:

a) changes in the sample, both in terms of increase or decrease in the number of PSU's depending upon the need of the time, can be achieved just as easily as in the case of simple random sampling (SRS), and

b) large scale survey frames are often used as main sources of selecting samples for ad hoc surveys from time to time (see Drew, Choudhry and Gray [1978]).

In such cases, following this approach, unequal probability samples can be selected for the ad hoc surveys with the same ease as SRS and without conflicting with the main continuous survey, once the samples has been selected for the main survey.

c) A feature often required for continuous surveys is the rotation of PSU's after a certain period of time. Again, this can easily be achieved in this approach irrespective of the complexity of the selection scheme used (for rejecting m units).

This rejective method used for selecting ΠPS samples is very practical and is recommended especially for those situations where the value of the parameter m is less than the sample size n, since the computation of Π_{ij}, the probability that both the units i and j are in the sample, is simplified to a great extent. It should be emphasized that if m is greater than the sample size n, then one should use the direct ΠPS method for selecting the sample unless there are over-riding considerations such as those mentioned earlier in the context of continuing surveys.

Sections 2 and 3 describe the actual selection mechanism and calculations of the inclusion probabilities and the joint inclusion probabilities. In section 4, suitability of this approach for selecting PSU's in Non Self-Representing (NSR) areas in the Canadian Labour Force Survey (LFS) is demonstrated. Results of empirical study are presented in the last section comparing the efficiencies of the Horvitz-Thompson [1952] estimator.

2. SELECTION PROCEDURE

The given finite population U consists of N units, $\{u_1, u_2, \ldots, u_N\}$ and a known "size measure" x_i is associated with the unit u_i; i=1, 2, ..., N. It is required to draw a sample of n distinct units from the population in such a way that the probability of unit u_i being in the sample is proportional to its size x_i (ΠPS) for each i=1, 2, ..., N.

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

$$p_i = \frac{x_i}{\sum_{i=1}^{N} x_i} \quad ; \; i=1,\, 2, \ldots N. \tag{2.1}$$

A sample of n distinct units will be selected from the population such that the probability Π_i; (i=1, 2, ..., N) for the ith unit to be in the sample is np_i. Since Π_i; i=1, 2, ..., N are the probabilities and hence necessarily less than or equal to one, therefore, the largest possible value of n (say n^*) is given by

$$n^* = [\frac{1}{p_{(N)}}], \tag{2.2}$$

where $p_{(N)} = \text{Max}\ (p_1, p_2, \ldots, p_N)$ and $[\cdot]$ is the integer function, i.e. the function gives the largest integer less than or equal to the argument.

The first step in the proposed method is to select $m(=N-n^*)$ units from the population using any ΠPS scheme and reject these units such that the remainder is a ΠPS sample of size n^* from the given population. A simple random sample without replacement (SRSWOR) of n out of n^* retained units is then selected. In order to show that this final sample is a ΠPS sample of size n from the entire population, we define

$$p_i^* = \frac{1-n^* p_i}{N-n^*} ; \quad i=1,2, \ldots, N . \tag{2.3}$$

It can be readily checked that p_i^*, $(i=1,2, \ldots, N)$ are the probabilities since from (2.3) $p_i^* > 0$ for all i and also from (2.3) $\sum_{i=1}^{N} p_i^* = 1$, therefore, $0 < p_i^* < 1$, $i=1, 2, \ldots, N$.

Since the sample of size m is selected using any given ΠPS scheme, the probability that the unit i is in the sample is mp_i^*; that is:

$$Pr(i \varepsilon R) = mp_i^* ; \qquad i=1, 2, \ldots, N, \tag{2.4}$$

where $R = \{R_1, R_2, \ldots, R_m\}$ is the set of m units selected with sizes p_i^*, $i=1, 2, \ldots, N$ using a ΠPS scheme. Let S^* be the set of n^* units not in the R, i.e. $S^*=U-R$. Now the final sample is a simple random sample of n units out of n^* units in S^*. We denote by S the set of n units selected in the second stage of sampling, i.e. SRS of n out of n^* units in S^*. That S is a ΠPS sample of size n from the population U of size N can be shown as follows:

$$\Pr(i\varepsilon S^*) = \Pr(i\notin R)$$

$$= 1-mp_i^* \;; \qquad i=1, 2, \ldots, N. \qquad (2.5)$$

Substituting $m=N-n^*$ and p_i^* from (2.2) in equation (2.5) gives

$$\Pr(i\varepsilon S^*) = n^* p_i \;; \qquad i=1, 2, \ldots, N. \qquad (2.6)$$

Thus, S^* is a ΠPS sample of size n^* (largest permissible sample size) from U. Since S is SRS of n out of n^* in S^*, therefore, we have

$$\Pr(i\varepsilon S) = \Pr(i\varepsilon S^*) \times \Pr(i\varepsilon S | i\varepsilon S^*)$$

$$= (n^* p_i) \times (\frac{n}{n^*})$$

$$= np_i \;; \qquad i=1, 2, \ldots, N. \qquad (2.7)$$

Denoting by Π_i, the probability that the unit i is in the set S we write

$$\Pi_i = np_i \;; \qquad i=1, 2, \ldots, N, \qquad (2.8)$$

i.e. S is a ΠPS sample of size n from the population U of size N.

When m<<n, the sample of m units can be selected using, for example, Fellegi's method [1963] or randomized PPS systematic method due to Hartley and Rao [1962] depending on the value of the parameter m and the population size N. Sinha [1973] has also suggested a rejective sampling scheme which achieves pre-specified inclusion probabilities of first two orders. Suitability of other methods may also be investigated for various situations.

3. CALCULATION OF Π_{ij}

The formula for Π_{ij}, the probability that both the units i and j are in the sample denoted by S in the previous section, will be derived. Denoting by Π^*_{ij}, the probability that both the units i and j are in S^*, we immediately have:

$$\Pi_{ij} = \frac{n(n-1)}{n^*(n^*-1)} \times \Pi^*_{ij}; \quad \begin{array}{l} i=1,\ 2,\ \ldots,\ N-1, \\ j=i+1,\ i+2,\ \ldots\ N. \end{array} \tag{3.1}$$

In order to find Π^*_{ij}, we define the following four mutually exclusive and exhaustive events for the units i and j with respect to the sets R and S^*.

Event E1: Both the units i and j in R.
Event E2: Unit i in R and unit j in S^*.
Event E3: Unit i in S^* and unit j in R.
Event E4: Both the units i and j in S^*.

Then we have:

$$\Pi^*_{ij} = \Pr(E4)$$

$$= 1-\{\Pr(E1) + \Pr(E2) + \Pr(E3)\} . \tag{3.2}$$

But $\Pr(E1) + \Pr(E2) = \Pr(i\varepsilon R)$

$$= mp^*_i \quad . \tag{3.3}$$

Similarly

$$\Pr(E1) + \Pr(E3) = \Pr(j\varepsilon R)$$

$$= mp^*_j \ . \tag{3.4}$$

Adding (3.3) and (3.4) gives

$$Pr(E1) + Pr(E1) + Pr(E2) + Pr(E3) = m(p_i^* + p_j^*). \tag{3.5}$$

Substituting in (3.2) for Pr(E1) + Pr(E2) + Pr(E3) from (3.5) we obtain

$$\begin{aligned} \Pi_{ij}^* &= 1 - m(p_i^* + p_i^*) + Pr(E1) \\ &= 1 - m(p_i^* + p_i^*) + \delta_{ij} \end{aligned} \tag{3.6}$$

where

$$\begin{aligned} \delta_{ij} &= Pr(E1) \\ &= Pr(\text{Both } i,\ j \varepsilon R). \end{aligned}$$

Substituting for Π_{ij}^* from (3.6) above in (3.1) gives Π_{ij}, the probability that both i, j are in the set S, i.e.

$$\Pi_{ij} = \frac{n(n-1)}{n^*(n^*-1)} \times [1 - m(p_i^* + p_j^*) + \delta_{ij}]\ ; \tag{3.7}$$

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

where $m=N-n^*$ and p_i^*, $(i=1,\ 2,\ \ldots,\ N)$ is defined in (2.3).

In conventional (or direct) sampling, as n increases, the computation of Π_{ij} becomes complicated, but in this case, the complexity lies only in the computation of δ_{ij}, which depends on the value of the parameter m. Since m is a population parameter and does not depend on the size of the sample, the complexity of Π_{ij} will not increase with the size of the sample. It may be noted that for the special case when the value of the parameter m is equal to 1, $\delta_{ij} = 0$, thus

$$\Pi^*_{ij} = 1-(p^*_i+p^*_i)$$

$$= (N-1)(p_i+p_j)-1 \tag{3.8}$$

and

$$\Pi_{ij} = \frac{n(n-1)}{(N-1)(N-2)} [(N-1)(p_i+p_j)-1]. \tag{3.9}$$

Since in this case only one unit (m=1) will be rejected with PPS, therefore, no special sampling scheme is required with this rejective method.

4. APPLICATION TO CANADIAN LABOUR FORCE SURVEY

The Canadian Labour Force Survey (LFS) follows a stratified multi-stage sampling design [see Statistics Canada, Catalogue No. 71-526]. In the non self-representing (NSR) areas, comprised of rural areas and small urban centers, a ΠPS sample of PSU's is selected from each stratum, where the "size measure" is the total population of the PSU from the previous census. In the earlier design, prior to the 1971 redesign, Fellegi's [1963] method was used to select PSU's where two PSU's were selected from each stratum. The method, though quite efficient, becomes very complicated when the number of PSU's to be selected is large, usually more than three. During the redesign of the LFS following the 1971 Census, one of the criterion for the choice of selection procedure at the design stage was that the procedure should be flexible enough to allow expansion of the sample in terms of the number of PSU's as well as rotation of PSU's. The randomized PPS (probability proportional to size) systematic sampling method [Hartley and Rao, 1962] was adopted for selecting PSU's, since it meets the necessary requirements [Gray, 1973]. However, computation of Π_{ij} becomes complicated for large values of n and N. This proposed sampling scheme, while equally efficient as well, has the additional advantage of simplicity for sample expansion and computation of Π_{ij}'s.

Following the redesign of the LFS, the sample was expanded in terms of number of PSU's in the NSR areas and the current number of selected PSU's per stratum varies from province to province. The actual number of PSU's selected from each stratum within a province and the number of cases with $m<n$ by province are given in Table A1 (Appendix A). In the context of simplifying the calculations of the joint inclusion probabilities for variance estimation, one would see that out of 127 NSR strata across Canada, 108 would result in simpler calculations, 16 with equal difficulty and only 3 with greater difficulty. Since the proposed scheme has the advantage that it permits the sample increase by simply selecting additional PSU's with SRS, thus the procedure among others may be considered for selecting PSU's in NSR areas during the redesign of LFS following the 1981 Census. Moreover, after the sample increase, the new Π_{ij}'s will be reconstructed from the previous ones by simply multiplying with the appropriate constant factor.

5. EMPIRICAL STUDY

We have chosen 4 populations to empirically evaluate the rejective scheme of sampling for samples of size 2, 3 and 4. The description of these populations is given in Table 1, where x is the known value of the size measure, and y is the value of the characteristic of interest which is unknown but measurable. The value of the parameter m is also given in Table 1 for each of the populations.

Table 1: Description of the Populations for Empirical Study

Pop.No.	Source	N	m	CV(x)[1]	CV(y)	ρ_{xy}
1	Fellegi [1963]	6	2	0.25	0.64	0.93
2	Gray [1971 b]	10	2	0.08	0.07	0.93
3	Cochran [1963, P. 204]	10	3	0.17	0.19	0.97
4	Cochran [1963, P. 225]	10	2	0.14	0.15	0.65

1 $$CV(x) = \frac{\sqrt{\sum_{i=1}^{N} (x_i - \bar{x})^2/N}}{\bar{x}}$$

CV(y) is defined in the same fashion.

In this study, the rejective scheme of sampling is accomplished by rejecting m units with Fellegi's [1963] PPS method, and also by rejecting m units with Randomized PPS systematic method of Hartley and Rao [1962]. The Horvitz and Thompson [1952] estimator $\hat{Y}_{HT} = \frac{1}{n} \sum_{i \varepsilon S} y_i/p_i$, where $\sum_{i \varepsilon S}$ denotes the sum over the n units in the sample, for estimating the unknown population total $Y = \sum_{i=1}^{N} y_i$ is considered. The variance of $\hat{Y}_{HT}$ as given by Yates and Grundy [1953], i.e.

$$V(\hat{Y}_{HT}) = \frac{1}{n^2} \sum_{i<j}\sum (\Pi_i \Pi_j - \Pi_{ij}) \left(\frac{y_i}{p_i} - \frac{y_j}{p_j}\right)^2 \qquad (5.1)$$

was computed for the rejective scheme when m units are rejected with Fellegi's PPS method and also when m units are rejected with Randomized PPS Systematic method, and these variances were then compared with those when sampling directly with Fellegi's PPS method and the Randomized PPS

Systematic method respectively. In order to compute the joint probabilities Π_{ij}'s for Fellegi's PPS method, the "working probabilities" of the method were computed by an iterative procedure and the Π_{ij}'s were constructed by summing the probabilities of all those samples that contain both the units i and j. For sample sizes greater than 4, the procedure becomes more complicated and involves a great deal of tedious calculations. For the Randomized PPS Systematic method, the Π_{ij}'s were computed using a FORTRAN subroutine by Hidiroglou and Gray [1979]. The algorithm used by the above authors is a modification of Connor's [1966] formula and is due to Gray [1971a]. Variances for the rejective method using Fellegi's method to reject m units and for Fellegi's method for selecting the sample for samples of size 2, 3, and 4 are given in Table B1 (Appendix B). Similar comparison is made by replacing Fellegi's method by Randomized PPS Systematic method both for rejecting m units in the rejective method and for selecting the sample, and these results are given in Table B2 (Appendix B).

From the two tables in Appendix B, it is seen that the rejective method has the same level of efficiency as the PPS method used in the rejective method to reject the m units, i.e. Fellegi's PPS method and the Randomized PPS Systematic method. Moreover, a comparison between Tables B1 and B2 shows that Fellegi's PPS method and the Randomized PPS Systematic method have the same variances and therefore are equally efficient. Since the efficiency of the rejective method is the same as the PPS method used in the rejective method to reject m units, there will be an advantage in using the rejective method for those populations where the value of the parameter m is less than or equal to the actual sample size n.

ACKNOWLEDGMENT

The authors wish to thank the referee for some useful comments.

RESUME

On propose en remplacement de la sélection directe de l'échantillon une autre solution qui, tout en maintenant l'efficacité au même niveau, simplifie les processus de sélection et d'estimation des variances dans un grand nombre de cas. Si n* représente la plus grande taille possible de l'échantillon prélevé selon une méthode qui donne à chaque unité une probabilité d'inclusion proportionnelle à sa taille (ΠPT) à partir d'une population donnée de taille N, la méthode proposée suppose alors la sélection des unités m (=N-n*) en utilisant le schéma ΠPT et en retirant ces unités de la population de manière à ce que le reste soit un échantillon ΠPT d'unités n*; l'échantillon définitif des unités n est ensuite prélevé comme sous-échantillon à partir de l'ensemble restant. Cette méthode de sélection de l'échantillon ΠPT peut être considérée comme l'équivalent de l'EAS dans lequel il est bien connu que la partie 'non échantillonnée' de la population et tout sous-échantillon de cette partie constituent également l'EAS de l'ensemble de la population, si l'on applique la procédure EAS. La méthode est très pratique dans les cas où m est inférieur à la taille réelle n de l'échantillon. De plus, elle présente un autre avantage pour les enquêtes permanentes, par exemple l'enquête sur la population active du Canada (EPA) où il faut augmenter (ou diminuer) le nombre des unités primaires d'échantillonnage (UPE) après la sélection initiale de l'échantillon. La méthode est également intéressante dans le cas du renouvellement de l'échantillon. Le document présente les avantages et inconvénients du plan proposé. L'efficacité de la méthode y est aussi évaluée de façon empirique.

REFERENCES

[1] Cochran, W.G., (1963), Sampling Techniques, Second Edition, New York: John Wiley & Sons.

[2] Connor, W.S. (1966), "An Exact Formula for the Probability that Two Specified Sampling Units Will Occur in a Sample Drawn with Unequal Probabilities and Without Replacement", Journal of the American Statistical Association, 61, 384-390.

[3] Drew, J.D., Choudhry, G.H. and Gray, G.B. (1978), "Some Methods For Updating Sample Survey Frames and Their Effects on Estimation", Survey Methodology, Statistics Canada, 4, 225-263.

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

[5] Gray, G.B. (1971), "Joint Probabilities of Selection of Units in Systematic Samples", Proceedings of the American Statistical.

[6] Gray, G.B. (1971), "Variance Components and Variance Functions", Proceedings of 'Statistics Canada' 71, 119-126.

[7] Gray, G.B. (1973), "On Increasing the Sample Size (number of PSU's)", Internal Technical Memorandum, Census and Household Survey Methods Division, Statistics Canada.

[8] Hartley, H.O. and Rao, J.N.K. (1962), "Sampling With Unequal Probabilities and Without Replacement", Annals of Mathematical Statistics, 33, 350-374.

[9] Hidiroglou, M.A. and Gray, G.B. (1979), "Construction of Joint Probability of Seleciton for Systematic PPS Sampling", Journal of the Royal Statistical Society.

[10] Horvitz, D.G. and Thompson, D.J. (1952), "A Generalization Of Sampling Without Replacement from a Finite Universe", Journal of the American Statistical Association, 47, 663-685.

[11] Sinha, B.K. (1973), "On Sampling Schemes to Realize Pre-assigned Sets of Inclusion Probabilities of First Two Orders", Calcutta Statistical Association Bulletin, 22, 89-110.

[12] Statistics Canada (1977), "Methodology of the Canadian Labour Force Survey (1976)", Catalogue No. 71-526.

[13] Yates, F. and Grundy, P.M. (1953), "Selection Without Replacement from Within Strata With Probability Proportional to Size", Journal of the Royal Statistical Society, Series B, 15, 235-261.

Appendix A

Table A1: Number of Strata With Rejective Sample of Units Compared With Actual Number Selected From NSR Strata in The Canadian Labour Force Survey

Province	n	No. of Strata with m<, = or > n			No. of Strata with m equal to n			
		m<n	m=n	m>n	1	2	3	4
Newfoundland	4	11	1	0	6	5	2	1
Prince Edward Island	6	3	0	0	0	3	0	0
Nova Scotia	3	9	5	1	3	6	5	1
New Brunswick	3	10	4	0	2	8	4	0
Quebec	3	18	1	2	3	15	1	2
Ontario	3	15	5	0	3	12	5	0
Manitoba	6	7	0	0	4	3	0	0
Saskatchewan	6	11	0	0	2	5	4	0
Alberta	4	14	0	0	9	5	0	0
British Columbia	4	10	0	0	4	5	1	0
Canada		108 less[1] difficulty	16 equal difficulty	3 greater difficulty	34	67	22	4

[1] Less, equal or greater difficulty refers to calculations of joint inclusion probabilities.

Appendix B

Table B1: Variances for the Rejective Scheme Using Fellegi's Method For Rejecting m Units (Scheme 1) and for Fellegi's Method For Selecting the Sample (Scheme 2)

Pop.No.	Sampling Schene	Sample Size 2	3	4
1	Scheme 1	8.0161	3.6782	1.5092
	Scheme 2	8.1672	3.8258	1.5269
	Efficiency (1 vs 2)	101.88%	104.01%	101.17%
2	Scheme 1	3.4922	2.0475	1.3251
	Scheme 2	3.4948	2.0509	1.3287
	Efficiency (1 vs 2)	100.07%	100.17%	100.27%
3	Scheme 1	276.04	161.67	104.48
	Scheme 2	276.15	161.81	104.63
	Efficiency (1 vs 2)	100.04%	100.09%	100.14%
4	Scheme 1	6375.5	3756.6	2447.1
	Scheme 2	6373.2	3753.7	2444.1
	Efficiency (1 vs 2)	99.96%	99.92%	99.88%

Table B2: Variances for the Rejective Scheme Using Randomized PPS Systematic Method for Rejecting m Units (Scheme 3) and for Randomized PPS Systematic Method for Selecting the Sample (Scheme 4)

Pop.No.	Sampling Scheme	Sample Size 2	3	4
1	Scheme 3	8.0261	3.6915	1.5242
	Scheme 4	8.5073	4.3927	1.5242
	Efficiency (3 vs 4)	106.00%	119.00%	100.00%
2	Scheme 3	3.4922	2.0475	1.3251
	Scheme 4	3.5114	2.0423	1.3309
	Efficiency (3 vs 4)	100.55%	99.75%	100.44%
3	Scheme 3	276.41	162.16	105.03
	Scheme 4	276.20	160.23	103.65
	Efficiency (3 vs 4)	99.92%	98.81%	98.69%
4	Scheme 3	6376.6	3758.0	2448.7
	Scheme 4	6373.7	3750.7	2446.2
	Efficiency (3 vs 4)	99.95%	99.81%	99.90%