Survey Methodology, June 1987
Vol. 13, No. 1, pp. 85-92
Statistics Canada

A Sampling Procedure with Inclusion Probabilities Proportional to Size

A. DEY and A.K. SRIVASTAVA[1]

ABSTRACT

A new unequal probability sampling scheme for selecting n (> 2) units without replacement from a finite population is proposed. This scheme ensures that the inclusion probabilities are proportional to sizes. It has the advantage of simplicity in selection and estimation and also provides a non-negative variance estimator. The variance of the Horvitz-Thompson (H-T) estimator under the proposed scheme is shown to be smaller than that of the customary estimator in probability proportional to size sampling with replacement. The proposed scheme also compares favourably with the without replacement scheme suggested by Sampford (1967) in an empirical study on a few natural populations.

KEY WORDS: Unequal probability sampling; Horvitz-Thompson estimator.

1. INTRODUCTION

In unequal probability sampling of n units without replacement from a finite population containing N units, if π_i denotes the inclusion probability of the i-th unit in the sample $i = 1, 2, \ldots, N$, the Horvitz and Thompson (1952) estimator (H-T estimator) of Y, the population total of the study variable y, is given by

$$\hat{Y} = \sum_{i \in s} (y_i/\pi_i), \tag{1.1}$$

where y_i is the y-value for the i-th unit and the summation extends over the units included in the sample. The variance of $\hat{Y}$ is

$$Var\ (\hat{Y}) = \sum_{i=1}^{N} \sum_{j>i}^{N} (\pi_i\pi_j - \pi_{ij})\ (y_i/\pi_i - y_j/\pi_j)^2 \tag{1.2}$$

where π_{ij} denotes the joint inclusion probability of the i-th and j-th units in the sample $(i \neq j, i, j = 1, 2, \ldots, N)$.

Considerable reduction in the variance of $\hat{Y}$ can be expected if the sampling scheme ensures that π_i are proportional to a given measure of size, say, x_i for $i = 1, 2, \ldots, N$, where it is assumed that x_i are nearly proportional to y_i. Sampling schemes in which π_i are proportional to x_i are termed Inclusion Probability Proportional to Size (IPPS) schemes. For a comprehensive account of unequal probability sampling procedures, including IPPS sampling schemes, the reader is referred to the monograph of Brewer and Hanif (1983).

Some desirable properties of an unequal probability scheme without replacement in general, and IPPS schemes in particular, are simplicity in selection and estimation, availability of a non-negative variance estimator, and better efficiency than with the probability proportional to size (PPS) with replacement strategy. Unfortunately, for sample size greater than two, not many of the available procedures meet these requirements fully.

[1] A. Dey and A.K. Srivastava, Indian Agricultural Statistics Research Institute, Library Avenue, New Delhi 110012, India.

In this paper, an IPPS sampling scheme is suggested for arbitrary sample sizes, $n > 2$. The procedure is rather simple both in sample selection and at the estimation stage since compact expressions for π_{ij} are available. It has also been possible to provide a positive estimator of variance of the *H-T* estimator of Y. The performance of the *H-T* estimator under the proposed scheme is compared with the PPS with replacement strategy and a simple sufficient condition is derived under which the performance of the former strategy is superior to that of the latter. An empirical study on a few natural populations indicates that the proposed scheme compares favourably with that suggested by Sampford (1967).

2. THE SAMPLING PROCEDURE

Consider a population of N units with y as the study variable and x, an auxiliary variable, as the size. It is assumed that x-values are known for all the population units. A sample of size $n\,(>2)$ is to be selected. To start with, it is assumed that n is *even*.

Divide the population into $m\,(> n/2)$ groups such that the i-th group contains $N_i\,(>2)$ units $(i = 1, 2, \ldots, m)$ and, for each group,

$$X_i/X > (n-2)/[n(m-1)], \tag{2.1}$$

where

$$X_i = \sum_{u=1}^{N_i} x_{i_u},$$

x_{i_u} is the value of x for the u-th unit in the i-th group and $X = X_1 + X_2 + \ldots + X_m$.

Equation (2.1) is satisfied if the X_i $(i = 1, 2, \ldots, m)$ are made nearly equal. It has been seen in actual populations, considered by Rao and Bayless (1969) and others, that this condition is satisfied for quite a few values of m if the groups are so formed that their sizes, X_i, are nearly equal. Rao and Lanke (1984) suggested a grouping procedure in which N units are grouped into R groups such that group totals, X_i, are nearly equal and group sizes are either $[N/R]$ or $[N/R]+1$, where $[x]$ is the largest integer contained in x. For the formation of groups, the Rao-Lanke procedure may also be tried.

Having formed the m groups, the suggested sampling procedure consists of the following steps:

Step 1. Select $n/2$ groups out of the m groups using Midzuno's (1951) sampling procedure with probabilities $\{P_i'\}$, that is, select one group with probability

$$P_i' = [n(m-1)P_i - (n-2)]/(2m-n), \text{ with } P_i = X_i/X,$$

and the remaining $(n/2) - 1$ groups with equal probabilities without replacement.

Step 2. From each of the selected groups, select two units by any IPPS procedure, say by Durbin's (1967) procedure, that is, in the i-th selected group $(i = 1, 2, \ldots, n/2)$ select one unit with probability

$$p_{i_u|i} = x_{i_u}/X_i,$$

and the second unit with revised probability

$$p_{i_u|i_v} = x_{i_v}\,[1/(X_i - 2x_{i_v}) + 1/(X_i - 2x_{i_u})]\,/D_i,$$

where

$$D_i = [1 + \sum_{u=1}^{N_i} x_{i_u}/(X_i - 2x_{i_u})].$$

For this sampling procedure, the inclusion probability for the i_u-th unit is evidently given by

$$\pi_{i_u} = n\,p_{i_u} \tag{2.2}$$

where

$$p_{i_u} = x_{i_u}/X.$$

Also, the joint inclusion probabilities for a pair of units are given by

$$\pi_{i_u i_v} = \frac{n\,p_{i_u}\,p_{i_v}\,(P_i - p_{i_u} - p_{i_v})}{D_i(P_i - 2\,p_{i_u})\,(P_i - 2\,p_{i_v})} \tag{2.3}$$

and

$$\pi_{i_u j_v} = \frac{n\,(n-2)\,p_{i_u}p_{j_v}}{(m-1)(m-2)\,P_iP_j}\;[(m-1)(P_i + P_j) - 1], \tag{2.4}$$

$$i \neq j,\ i, j = 1, 2, \ldots, m.$$

Thus we see that the proposed scheme is indeed an IPPS scheme.

As mentioned earlier, at step 2 of the proposed procedure, any IPPS scheme for selecting two units can be used. Since the procedure of Durbin (1967), which is equivalent to those of Rao (1963) and Brewer (1963), generally performs well, it has been adopted at step 2.

3. A VARIANCE ESTIMATOR

Two well-known unbiased estimators of $Var(\hat{Y})$ are due to Horvitz and Thompson (1952) and Yates and Grundy (1953). Both these estimators, however, suffer from the drawback that they sometimes assume negative values. In this section, a positive estimator of variance is proposed that utilizes the two-stage nature of the proposed sampling scheme.

Using a result due to Des Raj (1966), an unbiased estimator of $Var(\hat{Y})$ is given by

$$\hat{V}(\hat{Y}) = \sum_{i=1}^{n/2} \pi_i^{-1} \sum_{u<v}\sum \left[\frac{\pi_{i_u|i}\,\pi_{i_v|i}}{\pi_{i_u i_v|i}} - 1\right]\left[\frac{y_{i_u}}{\pi_{i_u|i}} - \frac{y_{i_v}}{\pi_{i_v|i}}\right]^2$$

$$+ \sum_{i<j}^{n/2}\sum^{n/2} \left(\frac{\pi_i\pi_j}{\pi_{ij}} - 1\right)\left[\frac{\hat{Y}_i}{\pi_i} - \frac{\hat{Y}_j}{\pi_j}\right]^2, \tag{3.1}$$

where

$$\pi_i = nP_i/2,$$

$$\pi_{ij} = \frac{n(n-2)}{4(m-2)}\{(P_i + P_j) - 1/(m-1)\},$$

$$\pi_{i_u|i} = 2p_{i_u}/P_i,$$

$$\pi_{i_u i_v|i} = \frac{2p_{i_u}p_{i_v}(P_i - p_{i_u} - p_{i_v})}{D_i P_i (P_i - 2p_{i_u})(P_i - 2p_{i_v})},$$

and

$$\hat{Y}_i = \sum_{u=1}^{2} y_{i_u}/\pi_{i_u|i}, \tag{3.2}$$

y_{i_u} being the y-value of the u-th unit in the i-th group.

The two terms in the right side of (3.1) correspond to the Yates-Grundy variance estimator in Durbin's and Midzuno's procedures. Since under these two sampling procedures the Yates-Grundy estimator of variance is always positive, it follows that the variance estimator given by (3.1) is also positive. However, the estimator in (3.1) is neither the Horvitz-Thompson nor the Yates-Grundy variance estimator.

4. COMPARISON WITH PPS WITH REPLACEMENT STRAGEGY

In this section, we compare the efficiencies of the following two strategies:

Strategy 1. The proposed sampling scheme in conjunction with the Horvitz–Thompson estimator.

Strategy 2. PPS sampling with replacement in conjunction with the customary estimator.

Strategy 1 is more efficient than Strategy 2 if and only if

$$\sum_{i=1}^{m}\sum_{u}\sum_{\neq v}^{N_i}\pi_{i_u i_v}(y_{i_u}/p_{i_u} - Y)(y_{i_v}/p_{i_v} - Y) + \sum_{i}\sum_{\neq j}^{m}\sum_{u}^{N_i}\sum_{v}^{N_j}\pi_{i_u j_v}(y_{i_u}/p_{i_u} - Y)(y_{j_v}/p_{j_v} - Y) < 0. \tag{4.1}$$

After some lengthy but routine algebra, the inequality (4.1) boils down to

$$-\sum_{i=1}^{n}(n/D_i)\sum_{u=1}^{N_i}(y_{i_u} - Y_i p_{i_u}/P_i)^2/(P_i - 2p_{i_u})$$

$$-n(n-2)\left[\sum_{i=1}^{m}(Y_i/P_i - Y)\right]^2/[(m-2)(m-1)] \tag{4.2}$$

$$-n(m-2)^{-1}\sum_{i=1}^{m}[\{(2n-m-2)P_i - (n-2)(m-1)^{-1}\}(Y_i/P_i - Y)^2] < 0,$$

where $$Y_i = \sum_u y_{i_u}.$$

Obviously, (4.2) holds if

(i) $(2n - m - 2) > 0$, and

(ii) $P_i > (n - 2)/[(m - 1)(2n - m - 2)]$. (4.3)

Also, since we are using Midzuno's procedure at the first stage with revised probabilities $\{P_i'\}$, each P_i must satisfy (2.1), that is, each P_i must satisfy

$$P_i > (n - 2)/[n(m - 1)].$$

Thus, (4.2) holds if

$$m \leq (n - 2). \tag{4.4}$$

It appears, therefore, that for Strategy 1 to be superior to Strategy 2, m should be chosen such that

$$n/2 < m \leq (n - 2). \tag{4.5}$$

However, it is clear that (4.4) is merely a sufficient condition and is not necessary. For $n > 6$, condition (4.5) offers a somewhat wide choice for the value of m, while for $n = 6$, (4.5) implies that $m = 3$. For $n = 4$, (4.5) does not lead to a feasible value of m. Therefore, for $n = 4$, an investigation into the performance of Strategy 1 has been taken up for various values of m, not constrained by (4.5), on certain natural populations. A description of the populations appears in Table 1. Table 2 presents the relative efficiency of Strategy 1 compared to Strategy 2 for the populations in Table 1. The performance of the H-T estimator under Sampford's (1967) scheme (called Strategy 3) is also compared with that of Strategy 2.

It can be observed from Table 2 that the performance of the proposed strategy (Strategy 1) compares favourably with that of Sampford (Strategy 3) for most of the populations. Of course, both strategies are superior to Strategy 2.

To achieve the relative efficiency of Strategy 1, the units were grouped in an ad-hoc manner, ensuring only that requirement (2.1) was satisfied. The procedure of Rao and Lanke (1984) was also attempted in forming the groups. However, the Rao-Lanke procedure did not always result in a high efficiency. Further investigations are necessary to decide the 'best' choice of groups. For certain populations, suitable groups satisfying (2.1) could not be formed for higher values of m, and thus, for these cases, the relative efficiencies are not reported in Table 2.

In conclusion, a brief comment on cases in which the desired sample size, n, is *odd* is in order. An IPPS sample for odd n may be obtained by selecting $(n + 1)$ units by the suggested procedure and then randomly discarding one unit. The expressions for π_i and π_{ij} under this procedure are straghtforward. Obviously, when one of the sample units out of $(n + 1)$ is discarded at random, the resulting sample consists of two units from each of the $(n - 1)/2$ groups and just one unit from one of the groups. An unbiased and positive estimator of $Var(\hat{Y})$ can be obtained, analogous to (3.1), on the basis of the $(n - 1)/2$ groups, each containing two units in the sample.

Table 1

Description of the Populations

Pop. Number	Source	N	y	x
1.	Des Raj (1965)	20	Number of households	Eye-estimated number of households
2.	Rao (1963)	14	Corn acreage in 1960	Corn acreage in 1958
3.	Cochran (1963, p. 204)	10	Weight of peaches	Eye-estimated weight of peaches
4.	Hanurav (1967)	20	Population in 1967	Population in 1957
5.	Hanurav (1967)	19	Population in 1967	Population in 1957
6.	Hanurav (1967)	16	Population in 1967	Population in 1957
7.	Hanurav (1967)	17	Population in 1967	Population in 1957
8.	Cochran (1963, p. 325)	10	Number of persons per block	Number of rooms per block
9.	Cochran (1963, p. 156, cities 1-16)	16	Population in 1930	Population in 1920
10.	Cochran (1963, p. 156, cities 33-49)	17	Population in 1930	Population in 1920
11.	Sampford (1962, p. 61)	35	Oats acreage in 1957	Oats acreage in 1947
12.	Sukhatme and Sukhatme (1970, p. 256, circles 1-20)	20	Wheat acreage	Number of villages
13.	Sukhatme and Sukhatme (1970, p. 256, circles 21-40)	20	Wheat acreage	Number of villages
14.	Yates (1960, p. 163)	20	Volume of timber	Eye-estimated volume of timber

Table 2

Percent Relative Efficiencies of Strategies 1 and 3 over Strategy 2 for the Populations in Table 1 ($n = 4$)

Pop. Number	Strategy 1				Strategy 3
	$m = 3$	4	5	6	
1.	130.1	118.7	120.8	124.5	127.8
2.	132.6	130.2	–	–	127.1
3.	149.1	–	–	–	147.9
4.	120.7	120.6	122.7	129.7	117.8
5.	129.1	138.7	158.7	–	125.1
6.	158.0	173.1	–	–	139.5
7.	151.9	144.8	169.2	–	131.9
8.	168.5	–	–	–	145.5
9.	118.3	116.3	–	–	109.5
10.	126.6	–	–	–	112.2
11.	113.8	116.2	135.6	129.9	113.8
12.	117.4	128.0	119.0	–	119.3
13.	122.2	120.6	–	–	119.7
14.	124.8	123.1	115.4	113.2	116.3

ACKNOWLEDGEMENTS

The authors would like to thank the referee for making many useful suggestions on the first draft.

REFERENCES

BREWER, K.R.W. (1963). A model of systematic sampling with unequal probabilities. *Australian Journal of Statistics*, 5, 5-13.

BREWER, K.R.W. and HANIF, M. (1983). *Sampling with Unequal Probabilities*. Lecture Notes in Statistics, No. 15. New York: Springer-Verlag.

COCHRAN, W.G. (1963). *Sampling Techniques*, (2nd. ed.). New York: John Wiley.

DES RAJ (1965). Variance estimation in randomized systematic sampling with probability proportional to size. *Journal of the American Statistical Association*, 60, 278-284.

DES RAJ (1966). Some remarks on a simple procedure of sampling without replacement. *Journal of the American Statistical Association*, 61, 391-396.

DURBIN, J. (1967). Design of multi-stage surveys for the estimation of sampling errors. *Journal of the Royal Statistical Society*, Ser. C, 16, 152-164.

HANURAV, T. (1967). Optimum utilization of auxiliary information: πps sampling of two units from a stratum. *Journal of the Royal Statistical Society*, Ser. B, 29, 379-391.

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.

MIDZUNO, H. (1951). On the sampling system with probability proportionate to sum of sizes. *Annals of the Institute of Statistical Mathematics*, 2, 99-108.

RAO, J.N.K. (1963). On three procedures of unequal probability sampling without replacement. *Journal of the American Statistical Association*, 58, 202-215.

RAO, J.N.K. and BAYLESS, D.L. (1969). An empirical study of the stabilities of estimators and variance estimators in unequal probability sampling of two units per stratum. *Journal of the American Statistical Association*, 64, 540-559.

RAO, J.N.K. and LANKE, J. (1984). Simplified unbiased variance estimation for multistage designs. *Biometrika*, 71, 387-395.

SAMPFORD, M.R. (1962). *An Introduction to Sampling Theory*. Edinburgh: Oliver and Boyd.

SAMPFORD, M.R. (1967). On sampling without replacement with unequal probabilities of selection. *Biometrika*, 54, 499-513.

SUKHATME, P.V. and SUKHATME, B.V. (1970). *Sampling Theory of Surveys with Applications*, (2nd. ed.). Ames, Iowa: Iowa State University Press.

YATES, F. (1960). *Sampling Methods for Censuses and Surveys*, (3rd. ed.). London: Griffin.

YATES, F. and GRUNDY, P.M. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society*, Ser. B, 15, 253-261.