Survey Methodology, June 1994
Vol. 20, No. 1, pp. 59–64
Statistics Canada

PPS Sampling over Two Occasions

N.G.N. PRASAD and J.E. GRAHAM[1]

ABSTRACT

The Random Group Method for sampling with probability proportional to size(PPS) is extended to sampling over two occasions. Information on a study variate observed on the first occasion is used to select the matched portion of the sample on the second occasion. Two real data sets are considered for numerical illustration and for comparsion with other existing methods.

KEY WORDS: Composite estimator; Efficiency comparisons; Random group method; Probability proportional to size.

1. INTRODUCTION

The practice of using a partial replacement sampling scheme in repeated surveys is quite common due, in part, to an anticipated increase in the efficiency of estimation as well as a reduction in the burden of response. Essentially, after each sampling occasion a fraction of the units observed on that occasion is rotated out of the sample and replaced by a fresh sub-sample from the population.This set of unmatched units is then observed on the next sampling occasion along with the remaining set of matched units. The literature abounds with discussions of sampling and estimation procedures for sampling with equal selection probabilities on two occasions. A particularly important case is the situation where the units are chosen on a given occasion with unequal selection probabilities. In the literature to date,information collected on the previous occasion is used to improve upon the customary estimator of the total or mean for the current occasion by using a difference method of estimation. In this article we present a sampling and estimation procedure for sampling on two occasions which incorporates information collected on the first (previous) occasion in selecting the sub-sample for observation on the second (current) occasion. For the sake of completeness and parsimony, we review only unequal probability selection procedures for two occasions in this section.

Consider a finite population of N units, labelled 1, 2, ..., N, and two sampling occasions: 1 (previous occasion) and 2 (current occasion). Let y_{1i} and y_{2i} denote the values of a characteristic y for the i-th unit observed on the first and second occasions respectively. Let Y_1 and Y_2 denote the respective population totals. Suppose a size measure x is known for each of the population units.

1.1 The Des Raj Scheme

Raj (1965) considered the following PPS (probability proportional to size) sampling scheme: On the first occasion a sample s of size n is selected with probabilities p_i proportional to the x_i values, $i = 1, 2, \ldots, N$, and with replacement (wr). On the second occasion a simple random sample s_1 of m units is selected from s without replacement (wor) and an independent PPS sample s_2 of $u = n - m$ units is selected wr from the entire population. Then Y_1 and Y_2 are respectively unbiasedly estimated by:

$$\hat{Y}_1 = \sum_{i \in s} y_{1i}/(np_i) \tag{1.1}$$

and

$$\hat{Y}_2 = Q\hat{Y}_{2u} + (1 - Q)\hat{Y}_{2m}, \tag{1.2}$$

where

$$\hat{Y}_{2u} = \sum_{i \in s_2} y_{2i}/(up_i), \tag{1.3}$$

$$\hat{Y}_{2m} = \sum_{i \in s} y_{1i}/(np_i) + \sum_{i \in s_1} (y_{2i} - y_{1i})/(mp_i), \tag{1.4}$$

and Q is a weight, $0 \leq Q \leq 1$. Assuming that

$$V_1 = \sum_{i=1}^{N} (y_{1i}/p_i - Y_1)^2 p_i = V_2$$

$$= \sum_{i=1}^{N} (y_{2i}/p_i - Y_2)^2 p_i = V, \tag{1.5}$$

[1] N.G.N. Prasad, Associate Professor, Department of Statistics and Applied Probability, University of Alberta, Edmonton, Alberta, Canada T6G 2G1; J.E. Graham, Professor, Department of Mathematics and Statistics, Carleton University, Ottawa, Ontario, Canada K1S 5B6.

the minimum variance of $\hat{Y}_2$ was found to be

$$V_{min}(\hat{Y}_2) = V[1 + \sqrt{2(1-\delta)}/(2n)], \quad (1.6)$$

where δ is given by

$$V\delta = \sum_{i=1}^{N} (y_{1i}/p_i - Y_1)(y_{2i}/p_i - Y_2)p_i. \quad (1.7)$$

1.2 The Ghangurde-Rao (G-R) Scheme

Under the PPSWOR framework, Ghangurde and Rao (1969) extended the Rao-Hartley-Cochran (RHC) Method, also known as the Random Group Method (See: Rao, Hartley and Cochran 1962) to sampling on two occasions. Under the RHC Method, the population of N units is split at random into n groups of sizes $N_1, N_2, \ldots, N_n$ such that $\sum_{h=1}^{n} N_h = N$, and a sample of one unit is drawn independently from each of the n groups with probabilities proportional to the initial selection probabilities, p_i. Under the G-R Method, the population is first divided at random into n groups, each of size N/n (assumed to be an integer). On the first occasion, one unit is drawn from each random group (as described above), giving a sample s of n units. On the second occasion, a simple random wor sample s_1 of $m = \lambda n (o < \lambda < 1)$ matched units is selected from s and an independent sample s_2 of $u = n - m$ units is drawn from the whole population of N units by the same method that was used in obtaining s. Then, a composite estimator of Y_2 is given by

$$\hat{Y}_2' = Q'\hat{Y}_{2u}' + (1 - Q')\hat{Y}_{2m}', \quad (1.8)$$

where $0 \leq Q' \leq 1$,

$$\hat{Y}_{2u}' = \sum_{i \epsilon s_2} \frac{y_{2i}P_i^*}{p_i}, \quad (1.9)$$

and

$$\hat{Y}_{2m}' = \sum_{i \epsilon s} \frac{y_{1i}P_i}{p_i} + nm^{-1} \sum_{i \epsilon s_1} \frac{(y_{2i} - y_{1i})P_i}{p_i}, \quad (1.10)$$

with P_i and P_i^* denoting the totals of the p_i values for the groups containing the i-th unit ($i = 1, 2, \ldots, N$) in the selection of s and s_2 respectively. Under assumption (1.5), the variance of $\hat{Y}_2'$ (with optimum values of Q' and λ) is given by

$$V_{min}(\hat{Y}_2') = \frac{NV}{2n(N-1)} \times [1 - n/N + \sqrt{2(1-\delta)}(1+\gamma)n/N], \quad (1.11)$$

where

$$\gamma = \frac{(1-\rho')V'}{(1-\delta)V} - 1,$$

$$V' = N^{-1}\sum_{i=1}^{N}(y_{1i} - \bar{Y}_1)^2 = N^{-1}\sum_{i=1}^{N}(y_{2i} - \bar{Y}_2)^2$$

and

$$\rho' = N^{-1}\sum_{i=1}^{N}(y_{1i} - \bar{Y}_1)(y_{2i} - \bar{Y}_2)/V'.$$

1.3 The Chotai Scheme

Chotai (1974), under the additional assumption that n/m is an integer, modified the G-R sampling design on the second occasion. A sample s is selected as in the G-R scheme on the first occasion. On the second occasion, the n units in the sample s are split at random into $m (= \lambda n)$ groups each of size n/m. One unit is selected from each of the m groups independently with probabilities proportional to the P_i's as defined in the G-R scheme. This selection yields the sample s_1. The selection of s_2 is the same as in the G-R scheme. Then a composite estimator of Y_2 is given by

$$\hat{Y}_2^C = Q^C\hat{Y}_{2u}^C + (1 - Q^C)\hat{Y}_{2m}^C, \quad (1.12)$$

where $0 \leq Q^C \leq 1$,

$$\hat{Y}_{2u}^C = \sum_{i \epsilon s_2} \frac{y_{2i}P_i^*}{p_i}, \quad (1.13)$$

and

$$\hat{Y}_{2m}^C = \sum_{i \epsilon s_1} \frac{(y_{2i} - y_{1i})P_i^+}{p_i} + \sum_{i \epsilon s} \frac{y_{1i}P_i}{p_i}. \quad (1.14)$$

Here, P_i and P_i^* are as defined in the G-R scheme, and P_i^+ denotes the total of the P_i-values for those random groups of s containing the i-th unit ($i = 1, 2, \ldots, N$) in the selection of s_1. The minimum variance of $\hat{Y}_2^C$ under assumption (1.5), obtained by using the optimum values of Q^C and λ, is given by

$$V_{min}(\hat{Y}_2^C) = \frac{NV}{2n(N-1)}[1 - n/N + \sqrt{2(1-\delta)}]. \quad (1.15)$$

Under this scheme, but without assumption (1.5), Chotai also considered an estimator of Y_2 (similar to Kulldorff's estimator for simple random sampling: See Kulldorff 1963), given by

$$\hat{Y}_2^{CM} = Q^{CM}\hat{Y}_{2u}^{C} + (1 - Q^{CM})\hat{Y}_{2m}^{CM}, \qquad (1.16)$$

where $\hat{Y}_{2u}^{C}$ is as defined in (1.13), Q^{CM} $(0 \leq Q^{CM} \leq 1)$ is an assigned weight to be determined and

$$\hat{Y}_{2m}^{CM} = \sum_{i \in s_1} \frac{(y_{2i} - \beta y_{1i})P_i^{+}}{p_i} + \beta \sum_{i \in s} \frac{y_{1i}P_i}{p_i}, \qquad (1.17)$$

with

$$\beta = \delta \left[\frac{\sum_{i=1}^{N} p_i(y_{2i}/p_i - Y_2)^2}{\sum_{i=1}^{N} p_i(y_{1i}/p_i - Y_1)^2} \right] = \delta \frac{V_2}{V_1}, \qquad (1.18)$$

and δ as defined in (1.7). The minimum variance of $\hat{Y}_2^{CM}$, using optimum values of Q^{CM} and λ, is given by

$$V_{min}(\hat{Y}_2^{CM}) = \frac{N}{2n(N-1)}(1 + \sqrt{1 - \delta^2} - n/N)V_2. \qquad (1.19)$$

To actually use $\hat{Y}_2^{CM}$ it is evidently necessary to first assess the value of β, which is usually not possible in practice. An estimate of β based on the available sample can be used but this will induce a bias in the estimator $\hat{Y}_2^{CM}$.

2. ALTERNATIVE SCHEMES FOR SAMPLING PPS OVER TWO OCCASIONS

We now present an alternative sampling and estimation procedure which does not require a known value of β as defined in (1.18). In this scheme information collected on the first occasion is used in selecting the sample on the second occasion. The approach is based upon a procedure developed by Prasad and Srivenkataramana (1980) and was used there in the context of double sampling where a second phase sub-sample is selected using information obtained from an initial sample. For simplicity, we first consider its implementation in Raj's (1965) scheme (described earlier).

2.1 A Modification of Des Raj's Scheme

On the first occasion a sample s of size n is selected with probabilities p_i proportional to the x_i values and with replacement. On the second occasion, instead of choosing a SRSWOR sub-sample, a sub-sample s_1 of m units is selected from s using a PPSWR scheme with size measure $z_i = y_{1i}/x_i$, where y_{1i} is the observed value for the y characteristic for unit i on the first occasion. A sample s_2 of size $u = n - m$ is drawn, independent of s, as in Raj (1965). A composite estimator of Y_2 is given by

$$\tilde{Y}_2 = Q\,\hat{Y}_{2u} + (1 - Q)\,\tilde{Y}_{2m},$$

where $\hat{Y}_{2u}$ is as defined in (1.3) and

$$\tilde{Y}_{2m} = \frac{1}{nm} \sum_{i \in s_1} \frac{(y_{2i}/p_i)}{(y_{1i}/p_i)} \sum_{i \in s} (y_{1i}/p_i),$$

with Q being a weight, $0 \leq Q \leq 1$. The minimum variance of $\tilde{Y}_2$, obtained by minimizing the variance of $\tilde{Y}_2$ with respect to Q, is given by

$$V_{min}(\tilde{Y}_2) = V_1C_1(n + C_1m)^{-1},$$

where $C_1 = \sum_{i=1}^{N} (y_{2i}/p_{1i} - Y_2)^2 p_{1i} V_1^{-1}$, with $p_{1i} = y_{1i}/Y_1$ and V_1 as defined in (1.5).

2.2 A Modification to Chotai's Scheme

As in Chotai (1974), assume that N, n, and $m(< n)$ are all positive integers such that N/n, N/u and n/m are also all integers. Then:

1. For the first occasion select a sample s of size n in the same manner as that adopted in the G-R procedure. For this set of units, observations y_{1i}, $i = 1, \ldots, n$, are made on a characteristic y.
2. For the second occasion, (a) split the n units in s at random into m groups, each of size n/m and draw one unit with PPS, $p_i^{*} = (y_{1i}P_i)/p_i$, independently from each of the m groups, yielding a sub-sample s_1, where P_i is as defined in the G-R scheme; (b) select s_2, a fresh sample of $u = n - m$ units from the entire population, and observe the second occasion y values, y_{2i}, for these u units in the same manner as in the G-R scheme.

Note that the difference between the proposed procedure and that of Chotai (1974) lies in the selection of s_1: in the former, information collected on the first occasion is used in selecting s_1 on the second occasion.

We now consider an estimator of the second occasion total Y_2 that exploits the proposed procedure. Let

$$y_{2i}^{*} = \frac{y_{2i}P_i}{p_i}.$$

A composite estimator of Y_2 is given by

$$\hat{Y}_2^{*} = Q^{**}\hat{Y}_{2u}^{C} + (1 - Q^{**})\hat{Y}_{2m}^{*}, \qquad (2.1)$$

where $\hat{Y}_{2u}^{C}$ is defined as in (1.13), $0 \le Q^{**} \le 1$ and

$$\hat{Y}_{2m}^{*} = \sum_{i \in s_1} \frac{y_{2i}^{*} \tilde{P}_i}{p_i^{*}}.$$

Here $\tilde{P}_i$ denotes the total of the p_i^{*} values associated with those units that belong to the random group from which the i-th unit was selected in s_1. Let E_1 and E_2 denote expectation and V_1 and V_2 denote variance over all s and for a given s, respectively. The unbiasedness of $\hat{Y}_{2m}^{*}$ and hence of $\hat{Y}_2^{*}$ for Y_2 follows by noting that the expected value of $\hat{Y}_{2m}^{*}$ is

$$E(\hat{Y}_{2m}^{*}) = E_1 E_2(\hat{Y}_{2m}^{*}) = E_1\left(\sum_{i \in s} \frac{y_{2i} P_i}{p_i}\right) = Y_2. \quad (2.2)$$

To obtain the variance of $\hat{Y}_{2m}^{*}$, consider

$$V_2(\hat{Y}_{2m}^{*}) = \frac{n-m}{m(n-1)} \sum_{i \in s} \left(\frac{y_{2i}^{*}}{p_i^{*}} - \sum_{i \in s} y_{2i}^{*}\right)^2 p_i^{*}$$

$$= \frac{n-m}{m(n-1)}\left[\sum_{i \in s} \frac{(y_{2i}^2/y_{1i})}{p_i} P_i \sum_{i \in s} \frac{y_{1i} P_i}{p_i} - \left(\sum_{i \in s} \frac{y_{2i} P_i}{p_i}\right)^2\right],$$

which leads, after considerable algebraic simplication, to

$$E_1 V_2(\hat{Y}_{2m}^{*}) = \frac{N(n-m)}{mn(N-1)} \sigma_3^2,$$

where

$$\sigma_3^2 = \sum_{i=1}^{N} \left(\frac{y_{2i}}{y_{1i}} Y_1 - Y_2\right)^2 \frac{y_{1i}}{Y_1}.$$

Noting that

$$V_1 E_2(\hat{Y}_{2m}^{*}) = \frac{N-n}{n(N-1)} \sigma_2^2,$$

it follows that

$$V(\hat{Y}_{2m}^{*}) = \frac{N}{n(N-1)}\left[(1 - n/N) + \frac{1-\lambda}{\lambda} h\right] \sigma_2^2, \quad (2.3)$$

where

$$h = \frac{\sigma_3^2}{\sigma_2^2}, \quad \sigma_2^2 = V_2 = \sum_{i=1}^{N} (y_{2i}/p_i - Y_2)^2 p_i \quad \text{and} \quad \lambda = \frac{m}{n}.$$

Because $\hat{Y}_{2u}^{C}$ and $\hat{Y}_{2m}^{*}$ are independent, the variance of $\hat{Y}_2^{*}$ is given by

$$V(\hat{Y}_2^{*}) = Q^{**2} V(\hat{Y}_{2u}^{C}) + (1 - Q^{**})^2 V(\hat{Y}_{2m}^{*}),$$

where

$$V(\hat{Y}_{2u}^{C}) = \frac{N-u}{u(N-1)} \sigma_2^2,$$

and $V(\hat{Y}_{2m}^{*})$ is given by (2.3).

The minimum variance of $V(\hat{Y}_2^{*})$ is obtained by using optimum values of Q^{**} and λ, respectively given by

$$Q^{**} = \frac{(1 - n/N) + \frac{(1-\lambda)}{\lambda} h}{(1 - n/N) + \frac{(1-\lambda)}{\lambda} h + \frac{(1 - (1-\lambda)n/N)}{(1-\lambda)}},$$

and

$$\lambda = \frac{\sqrt{h}}{1 + \sqrt{h}}.$$

Hence, the minimum variance of $V(\hat{Y}_2^{*})$ is given by

$$V_{min}(\hat{Y}_2^{*}) = \frac{N\sigma_2^2}{n(N-1)}[1 - n/N + \sqrt{h}]. \quad (2.4)$$

Note that the quantity h reflects the efficiency of the estimator using the p_i's as initial selection probabilities over the estimator with initial selection probabilities y_{1i}/Y_1. A "small" value of h leads to an increase in the efficiency of the proposed method over Chotai's.

3. NUMERICAL EFFICIENCY COMPARISONS

The composite estimators $\hat{Y}_2^{C}$ defined in (1.12), $\hat{Y}_2^{CM}$ defined in (1.16) and $\hat{Y}_2^{*}$ defined in (2.1) are now compared at their respective optimum Q and λ values. The efficiency of the scheme proposed in 2.2 relative to Chotai's (1974) procedure is examined through a comparison of the following two relative efficiencies:

$$\text{RE1} = \frac{V_{min}(\hat{Y}_2^{C})}{V_{min}(\hat{Y}_2^{*})} = \frac{(1 - n/N) + \sqrt{2(1-\delta)}}{(1 - n/N) + \sqrt{h}}$$

and

$$\text{RE2} = \frac{V_{min}(\hat{Y}_2^{CM})}{V_{min}(\hat{Y}_2^{*})} = \frac{(1 - n/N) + \sqrt{1 - \delta^2}}{(1 - n/N) + \sqrt{h}},$$

evaluated respectively obtained using (1.15) and (2.4), and (1.19) and (2.4). It follows that the proposed scheme is superior to that of Chotai using Kulldorff's estimator (which depends on the unknown constant β) for those populations having $h < (1 - \delta^2)$. In order to permit meaningful numerical comparisons, two data sets that have appeared elsewhere in the literature are used here.

Data Set A: This data set relates to the area under wheat in 1964 (y_2), in 1963 (y_1) and cultivated area in 1961 (x) for 34 villages in India (See Murthy 1967). The parameter values for this data set are $\delta = 0.6404$ and $h = 0.1868$.

Data Set B: This data set relates to the area under wheat in 1937 (y_2) and in 1936 (y_1) and cultivated area in 1930 (x) for a sample of 34 villages in India (see: Sukhatme, P.V. and Sukhatme, B.V. 1970). The corresponding parameter values for this data set are $\delta = 0.7635$ and $h = 0.3811$.

Using these values for δ and h the two relative efficiencies values RE1 and RE2 (expressed as percentages) were computed for selected values of n/N and are given in Tables 1 and 2.

Table 1

RE1% - Values for Data Sets A and B

n/N	Data Set A	Data Set B
0.05	130.09	124.30
0.10	131.22	125.21
0.15	132.43	126.19
0.20	133.75	127.25
0.25	135.18	128.41
0.30	136.73	129.66

Table 2

RE2% - Values for Data Sets A and B

n/N	Data Set A	Data Set B
0.05	104.49	101.82
0.10	104.64	101.88
0.15	104.80	101.94
0.20	104.97	102.01
0.25	105.15	102.08
0.30	105.34	102.16

An examination of Table 1 leads to the conclusion that the proposed scheme out performs that of Chotai (1974). The gain in the efficiency ranges from 30% to 37% for Data Set A and from 24% to 30% for Data Set B as the sampling fraction varies from 0.05 to 0.30. Note that the increase in efficiency is greater for Data Set A than for Data Set B because of the difference in the value of the parameters h (0.1868 *vs.* 0.3811) and of δ (0.6404 *vs.* 0.7635). Recall that h measures the efficiency of p_i as a size measure for unit i compared to the use of y_{1i} as a size measure in estimating the total Y_2 for the current occasion and δ is the correlation between y_{1i}/p_i and y_{2i}/p_i as defined in (1.7). When h is relatively small, greater gains in efficiency are realized with the proposed scheme than when h is not small. In both cases, however, the efficiency gains using the proposed procedure are worthwhile.

The efficiency gains using the proposed method compared to the use of Chotai's scheme with Kulldorff's estimator (as reported in Table 2) are minimal, varying from 4.5% to 5.3% for Data Set A and from 1.8% to 2.2% from Data Set B. But in order to use Kulldorff's estimator, the value of β must be available. In practice this is not the case. It follows that the proposed strategy performs well from the point of view of actual implementation and of efficiency gain.

There are situations where the auxiliary information needed to compute the initial selection probabilities is not available. A simple random sampling scheme may then be used in place of the RHC procedure in selecting the sample for the first occasion enumeration; the RHC procedure can then be adopted in selecting s_1 by using the SRS information on the study variable collected on the first occasion. The theory for such a procedure follows directly as a special case of that presented by taking $p_i = 1/N$, $i = 1, \ldots, N$. One would anticipate that substantial gains in efficiency would then result in this situation.

ACKNOWLEDGEMENTS

The authors thank Professor J.N.K. Rao for suggesting this problem, and the referee for constructive suggestions. This research was supported by grants from the Natural Sciences and Engineering Research Council of Canada.

REFERENCES

CHOTAI, J. (1974). A note on the Rao-Hartley-Cochran method for PPS sampling over two occasions. *Sankhyā*, Series C, 36, 173-180.

GHANGURDE, P., and RAO, J.N.K. (1969). Some results on sampling over two occasions. *Sankhyā*, Series A, 31, 463-472.

KULLDORFF, G. (1963). Some problems of optimum allocation for sampling on two occasions. *Review of the International Statistical Institute*, 31, 24-57.

MURTHY, N.N. (1967). *Sampling Theory and Methods*. Calcutta, India: Statistical Publishing Society.

PRASAD, N.G.N., and SRIVENKATARAMANA, T. (1980). Double sampling with PPS selection. *Vignana Bharathi*, 6, 52-58.

RAJ, D. (1965). On sampling over two occasions with probabilities proportional to size. *Annals of Mathematical Statistics*, 36, 327-330.

RAO, J.N.K., HARTLEY, H.O., and COCHRAN, W.G. (1962). On a simple procedure of unequal probability sampling without replacement. *Journal of the Royal Statistical Society*, 24, 482-491.

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