Survey Methodology, December 1998
Vol. 24, No. 2, pp. 185-192
Statistics Canada

Sampling on Two Occasions: Estimation of Population Total

RAGHUNATH ARNAB[1]

ABSTRACT

Two sampling strategies have been proposed for estimating the finite population total for the most recent occasion, based on the samples selected over two occasions involving varying probability sampling schemes. Attempts have been made to utilize the data collected on a study variable, in the first occasion, as a measure of size and a stratification variable for selection of the matched-sample on the second occasion. Relative efficiencies of the proposed strategies have been compared with suitable alternatives.

KEY WORDS: Composite estimator; Matched-sample; Sampling schemes; Sampling strategies; Varying probability sampling schemes.

1. INTRODUCTION

We very often survey the same population at regular time intervals to estimate the same population characteristics which change over time. For example, many countries collect data to estimate total number of unemployed persons, HIV infected people, immigrants *etc.*, on an annual or quarterly basis. In this article, we consider a finite population $U = (U_1, ..., U_i, ..., U_N)$ of N identifiable units, which is supposed to be sampled over two occasions, to estimate the population total of a variable under study for the current (second) occasion. In successive sampling, one utilizes data collected on the previous (first) occasion effectively, to get an efficient strategy in consideration of cost, and providing an efficient estimator of the population total for the current occasion. Extensive literature is now available for this purpose. Singh (1967), and Avadhani and Sukhatme (1970) utilized information, collected on the first occasion as a measure of size, for the selection of the matched sample on the second occasion; while Arnab (1991) utilized such information as a stratification variable, as well as the measure of size, for selection of the sample on the second occasion. Recently, Prasad and Graham (1994) modified Raj's (1965) and Chotai's (1974) sampling strategies, by using information of the first occasion as a measure of size, for the selection of the matched sample in the second occasion. They found empirically, that one of their proposed strategies fares better than that given by Chotai (1974). In this article, two alternative strategies are proposed. One of them utilizes information in the first occasion as a measure of size, and the other utilizes information as a measure of size and also as a stratification variable for selection of the matched sample in the second occasion. In this paper, it is shown that one of the proposed strategies is better than that given by Prasad and Graham (1994) and for the other, we do not have any definite theoretical conclusion. However, empirical evidence shows that the latter is more efficient than that described by Prasad and Graham (1994), as well as the former proposed strategy. This is possible because it utilizes first occasion values in all possible stages *viz.*, stratification, estimation and selection of the matched sample in the second occasion.

The general methods of selection of samples and estimation over two occasions are described below.

1.1 Sampling Schemes

On the first occasion, a sample s_1, of size n, is selected by some suitable sampling design, say $\boldsymbol{P}_1$, and the data y_{1i} $i\varepsilon s_1$, is obtained where y_{1i} (y_{2i}) is the value of the variate y under study, for the i-th unit on the first (second) occasion. On the second occasion, a matched sample (sub-sample) s_m of size m ($= n\lambda$, assumed to be an integer, $0 \le \lambda \le 1$) is selected from s_1 by some suitable sampling scheme P_m, and it is supplemented by an un-matched sample s_u of size u ($= n\mu = n - m, \mu = 1 - \lambda$) either from the entire population U or from U/s_1, the set of units not selected in the first occasion, by some suitable sampling design $\boldsymbol{P}_u$, and information y_{2i} ($i\varepsilon s_m, i\varepsilon s_u$) on the second occasion is obtained. It is obvious that the cost of survey for the matched sampled units is expected to be much lower than that of the un-matched units, but for the sake of simplicity, we assume that the cost of the survey remains the same for all the units in the second occasion.

1.2 Method of Estimation

From the data y_{1i}, $i\varepsilon s_1$, and y_{2i}, $i\varepsilon s_m$ collected through the initial sample s_1, and the matched sample s_m, an unbiased estimator $\hat{Y}_{2m}$ for Y_2, the population total for the second occasion, is formed by treating the y_{1i}'s, $i\varepsilon s_1$, as auxiliary information. Thus $\hat{Y}_{2m}$ is normally a difference, ratio or regression estimator. From the un-matched sample s_u, an unbiased estimator $\hat{Y}_{2u}$ is also constructed for Y_2. Finally, a composite estimator, a combination of $\hat{Y}_{2m}$ and

[1] Raghunath Arnab, Department of Statistics, University of Durban-Westville, Private Bag-X54001, Durban - 4000, South Africa.

$\hat{Y}_{2u}$, is obtained by using a suitable weight of φ $(0 \le \varphi \le 1)$, as

$$\hat{Y}_2 = \varphi\, \hat{Y}_{2m} + (1 - \varphi)\, \hat{Y}_{2u}. \qquad (1)$$

The optimum value of $\varphi = \varphi(\lambda)$ is obtained by minimizing $V(\hat{Y}_2)$, the variance of $\hat{Y}_2$ with respect to φ, for a given value of m (*i.e.*, λ). The expressions for $\varphi(\lambda)$ and $V(\hat{Y}_2 I\lambda)$, the variance of $\hat{Y}_2$ with $\varphi = \varphi(\lambda)$ are obtained as follows, when $\hat{Y}_{2m}$ and $\hat{Y}_{2u}$ are independent:

$$\varphi(\lambda) = (1/V_m)\,[1/V_m + 1/V_u]^{-1},$$

$$V(\hat{Y}_2 I\lambda) = [1/V_m + 1/V_u]^{-1},$$

where V_m and V_u are variances of $\hat{Y}_{2m}$ and $\hat{Y}_{2u}$ respectively. The optimum proportion of matched sample $\lambda = \lambda_0$, is obtained by minimizing $V(\hat{Y}_2 I\lambda)$ with respect to λ. Finally, putting $\lambda = \lambda_0$ in the expression for $V(\hat{Y}_2 I\lambda)$, the minimum variance of $\hat{Y}_2$ is obtained, and it will be denoted by $V_{\min}(\hat{Y}_2) = V(\hat{Y}_2 I\lambda_0)$. Our object is to find a suitable strategy, which is a combination of $P = (P_1, P_m, P_u)$ and $\hat{Y}_2$, to control the magnitude of $V_{\min}(\hat{Y}_2)$ to a minimum.

1.3 A Few Sampling Strategies

1.3.1 Avadhani and Sukhatme (1970)

On the first occasion, the initial sample s_1 of size n was selected by simple random sampling without replacement (SRSWOR) method, assuming that no auxiliary information is available prior to this survey. On the second occasion, the matched sample s_m of size m was selected from s_1 by the Rao, Hartley and Cochran (RHC, in brief, 1962) sampling scheme using y_{1i} as a measure of size for the i-th unit $i\varepsilon s_1$, assuming y_{1i}'s are positive. Under the RHC sampling scheme, the selected n units of s_1, are divided at random into m groups, each of size n/m, which is assumed to be an integer. From each of the selected groups, one unit is selected independently with probability proportional to the measure of size. Thus if the i-th unit, U_i, belongs to the j-th group G_j $(j = 1, ..., m)$ then U_i will be selected with the probability $q_i^*(i\varepsilon s_1) = y_{1i}/\sum_{i\varepsilon s_1} y_{1i}$. The un-matched sample s_u was selected from U/s_1 by SRSWOR.

1.3.2 Chotai (1

On the first occasion, the initial sample s_1 of size n was selected by the RHC scheme of sampling (assuming N/n is an integer), as described above with probability proportional to z_i, the size measure for the i-th unit which is, assumed to be positive and known for every $i\varepsilon U$. Let $\Delta_j = \sum_{k\varepsilon G_j} p_k$, the sum of p_k $(= z_k/Z, Z = \sum_{i\varepsilon U} z_i)$ values that belong to the random group G_j $(j = 1, ..., n)$, which is formed in selecting the sample s_1 by the RHC method. The matched sample s_m was selected from s_1 by the RHC scheme, with normed size measure Δ_i, for the i-th unit $i\varepsilon s_1$ $(\sum_{i\varepsilon s_1} \Delta_i = 1)$ assuming n/m is an integer. The un-matched sample, s_u was selected by the RHC sampling scheme with normed size measure p_i for the i-th unit assuming N/u is an integer. Let P_i^+ (P_i') = total of the Δ_i (p_i) values associated with those units that belong to the random group from which the i-th unit was selected in s_m (s_u) by the RHC sampling scheme with $\sum_{i\varepsilon s_m} P_i^+ = 1$ $(\sum_{i\varepsilon s_u} P_i' = 1)$.

The composite estimator for Y_2 is given by

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

where

$$\hat{Y}_{2m} = \sum_{i\varepsilon s_m} (y_{2i}/p_i)\, P_i^+ - \gamma\left[\sum_{i\varepsilon s_m} (y_{1i}/p_i)\, P_i^+ - \sum_{i\varepsilon s_1} (y_{1i}/p_i)\, \Delta_i\right];$$

$$\hat{Y}_{2u} = \sum_{i\varepsilon s_u} (y_{2i}/p_i) P_i' \qquad (2)$$

where γ is a suitably chosen constant to minimize variance of $\hat{Y}_{2m}$. Chotai (1974) derived the expression for the minimum variance of $\hat{Y}_2$ as

$$V_{\min}(\hat{Y}_2) = k\,[1 - f + \sqrt{(1 - \delta^{*2})}]\, \sigma_2^2/2 = V_c \text{ (say)} \qquad (3)$$

where

$$k = N/\{n(N-1)\},\ f = n/N,$$

$$\sigma_t^2 = \sum_{i\varepsilon U} p_i (y_{ti}/p_i - Y_t)^2,\ t = 1, 2$$

$$Y_t = \sum_{i\varepsilon U} y_{ti},\ t = 1, 2$$

$$\delta^* = \sum_{i\varepsilon U} p_i (y_{2i}/p_i - Y_2)\ (y_{1i}/p_i - Y_1)/(\sigma_1 \sigma_2). \qquad (4)$$

1.3.3 Arnab (1991)

Arnab (1991) presented several strategies where the initial sample s_1 was selected by probability proportional to size with replacement (PPSWR) using normed size measure $p_i = z_i/Z$ for the i-th unit. Utilizing the ascertain values y_{1i}'s $(i\varepsilon s_1)$ on the basis of certain criteria, the n sample units are assigned to a suitable number of L strata. Let s_{1h} be the sample of size n_h, belonging to the h-th stratum $(s_1 = U_h s_{1h}$ and $\sum_h n_h = n)$. Here, it is assumed that n is large enough to ensure that n_h is positive for every h in practice. On the second occasion, sub-samples s_{mh}'s

of size m_h's ($= v_h n_h$, v_h is a predetermined fraction and m_h is assumed to be an integer) are selected from s_{1h}'s independently, by suitable sampling schemes involving y_{1i}'s, $i\varepsilon s_1$ in the selection of matched samples s_{mh}'s. The unmatched sample s_u is selected by PPSWR method from the entire population U using z_i' as measure of size.

1.3.4 Prasad and Graham (1994)

Here the initial sample s_1 is selected by the RHC scheme of sampling similar to Chotai (1974) with normed size measure $p_i = z_i / Z$ for the i-th unit. The matched sample s_m is selected from s_1 by the RHC scheme with $p_i^* = (y_{1i}\Delta_i/p_i)/\sum_{i\varepsilon s_1}(y_{1i}\Delta_i/p_i)$ for the i-th unit, $i\varepsilon s_1$; where Δ_i is the sum of the p_j values for the group containing the i-th unit, formed in selecting s_1 by the RHC sampling scheme of sampling. The un-matched sample, s_u was selected from the entire population U by the RHC scheme similar to that presented by Chotai (1974). Here also N/n, n/m and N/u are assumed to be integers. Prasad and Graham (1994) proposed the following composite estimator for Y_2:

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

where $\hat{Y}_{2m} = \sum_{i\varepsilon s_m}(y_{2i}^*/p_i^*)\,\tilde{P}_i$; $\hat{Y}_{2u} = \sum_{i\varepsilon s_u}(y_{2i}/p_i)P_i'$; $y_{2i}^* = y_{2i}\Delta_i/p_i$; $\tilde{P}_i\,(P_i') =$ total of the $p_i^*\,(p_i)$ values associated with those units that belong to the random group from which the i-th unit was selected in $s_m\,(s_u)$. The expression for minimum variance of $\hat{Y}_2$, is obtained as:

$$V_{\min}(\hat{Y}_2) = k(1 - f + \sqrt{\zeta})\,\sigma_2^2/2 = V_{PG}\ (\text{say}) \qquad (5)$$

where

$$\zeta = \sigma_3^2/\sigma_2^2,\ \sigma_3^2 = \sum_{i\varepsilon U} q_i(y_{2i}/q_i - Y_2)^2,\ q_i = y_{1i}/Y_1; \qquad (6)$$

k, f, σ_2^2 and Y_1 are defined in (4).

In Prasad and Graham's (1994) expression for $V_{\min}(\hat{Y}_2)$, the divisor 2 was omitted and is obviously a typographical error.

Remark 1.1

From the strategies described in section 1.3, we note that the Avadhani and Sukhatme (1970) scheme does not require information on size measures in the whole frame, and hence is less demanding than the others. Chotai (1974) used the original size measures p_i in selection, but the first survey values y_{1i}'s, $i\varepsilon s_1$ were used additionally in estimation only. The use of additional information, p_i's, for the selection of the initial sample s_1 will make Chotai's (1974) strategy more efficient than that of Avadhani and Suhkatme (1970). But to use the optimal estimator $\hat{Y}_2$ for the Avadhani and Sukhatme (1970) strategy, one needs to estimate φ, the only unknown parameter. However, in Chotai's (1974) strategy, both the parameters φ and γ have to be estimated in order to use the optimum $\hat{Y}_2$. Prasad and Graham (1994) used both these variables in the selection of the matched sample (hence automatically in the estimation) and showed empirically that their strategy fares better than that of Chotai (1974). In addition, to gain in efficiency, Prasad and Graham's (1994) strategy can be used in practice, because $\hat{Y}_2$ involves only one unknown parameter, φ. It should be noted that Arnab (1991) first introduced the principle of stratification using y_{1i}'s, $i\varepsilon s_1$ as a stratification variable. This should always be done in practice whenever the necessary information is available, particularly in the selection of large units with marked size differences of the type considered in the numerical examples in section 3. Arnab's (1991) strategy is expected to be more efficient than the preceding strategies, since it utilizes first occasion values for stratification in addition to estimation. However, the optimal estimator $\hat{Y}_2$ contains the several unknown parameters (for details see Arnab 1991) which may hinder the application of the strategy especially when the sample size is not large enough.

2. PROPOSED STRATEGIES

Here two sampling strategies have been proposed which are modifications of strategies proposed by Prasad and Graham (1994) and Arnab (1991), respectively.

2.1 Strategy 1

The sampling scheme for this strategy is the same as was considered by Prasad and Graham (1994), and described in section 1.3.4. Here, only the estimator based on the matched sample s_m, has been modified by introducing the original size measure into the estimation. The proposed modified estimator $\hat{Y}_{2m}^*$ and the composite estimators for Y_2 are as follows:

$$\hat{Y}_{2m}^* = \sum_{i\varepsilon s_m}(y_{2i}^*/p_i^*)\,\tilde{P}_i - \beta\left[\sum_{i\varepsilon s_m}(z_i^*/p_i^*)\,\tilde{P}_i - Z\right] =$$

$$\sum_{i\varepsilon s_m}(r_i^*/p_i^*)\,\tilde{P}_i + \beta Z$$

where $z_i^* = z_i\Delta_i/p_i$, $y_{2i}^* = y_{2i}\Delta_i/p_i$, $r_i^* = r_i\Delta_i/p_i$, $r_i = y_{2i} - \beta z_i$ and β is a suitably chosen constant to minimize variance of $\hat{Y}_{2m}^*$; $p_i^*, \tilde{P}_i$ and Δ_i are as described in the section 1.3.4;

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

where $\hat{Y}_{2u}$ is given in (2).

Denoting $E_1(V_1)$ as unconditional expectation (variance) over selection of the sample s_1, and $E_2(V_2)$ the conditional expectation (variance) over s_m when s is fixed, one gets the variance of $\hat{Y}_{2m}^*$ for a given value of β, as

$$V(\hat{Y}_{2m}^* I\beta) = E_1 V_2(\hat{Y}_{2m}^* I\beta) + V_1 E_2(\hat{Y}_{2m}^* I\beta).$$

Following Prasad and Graham (1994), we obtain

$$E_1 V_2(\hat{Y}^*_{2m} I\beta) = k_1 \sigma_3^{*2}(\beta)$$

and

$$V_1 E_2(\hat{Y}^*_{2m}) = k(1-f)\sigma_2^2$$

where

$$k_1 = N(n-m)/\{nm(N-1)\};$$

$$\sigma_3^{*2}(\beta) = \sum_{i\varepsilon U} q_i (r_i/q_i - R)^2$$

$$= \sigma_3^2 + \beta^2 \sigma_0^2 - 2\beta\sigma_0\sigma_3\delta;$$

$$R = \sum_{i\varepsilon U} R_i = Y_2 - \beta Z, \delta = \sigma_{03}/(\sigma_0\sigma_3),$$

$$\sigma_0^2 = \sum_{i\varepsilon U} q_i (z_i/q_i - Z)^2,$$

$$\sigma_{03} = \sum_{i\varepsilon U} q_i (y_{2i}/q_i - Y_2)(z_i/q_i - Z) \qquad (7)$$

σ_2^2, k and σ_3^2, q_i are as in (4) and (6), respectively. The optimum value of β that minimizes $V(\hat{Y}^*_{2m} I\beta)$ comes out as, opt $\beta = \beta_0 = \delta\sigma_3/\sigma_0$.

Putting the optimum value of $\beta = \beta_0$ in the expression of $V(\hat{Y}^*_{2m} I\beta)$, we get the optimum value of

$$V(\hat{Y}^*_{2m} I\beta) = V(\hat{Y}^*_{2m} I\beta_0) = k[(1-f) + (1-\lambda)\zeta^*/\lambda]\sigma_2^2$$

where $\zeta^* = (1-\delta^2)\zeta$; k, f and ζ are defined in (4) and (6) respectively.

The optimum variance of $\hat{Y}_2$ for a given value of λ is obtained by minimizing the variance of $\hat{Y}_2$ with respect to φ when $\beta = \beta_0$, and is given by

$$V_{\text{opt}}(\hat{Y}_2 I\lambda) = [1/V(\hat{Y}^*_{2m} I\beta_0) + 1/(\hat{Y}_{2u})]^{-1}$$

$$= [1/\{k(1-f) + (1-\lambda)\zeta^*/\lambda\} + \mu/\{k(1-f\mu)\}]^{-1}\sigma_2^2.$$

Finally, minimizing $V_{\text{opt}}(\hat{Y}_2 I\lambda)$ with respect to λ, the optimum proportion of the matched sample and minimum variance of $\hat{Y}_2$ are obtained respectively as

$$\text{opt } \lambda = \lambda_0 = \sqrt{\zeta^*}/(1+\sqrt{\zeta^*})$$

and

$$V_{\min}(\hat{Y}_2) = k(1-f+\sqrt{\zeta^*})\sigma_2^2/2 = M_1 \text{ (say)} \qquad (8)$$

Remark 2.1

The estimator $\hat{Y}^*_{2m}$, described in (1) is usable in practice when the optimum value of $\beta = \beta_0$ is known, or a good guess value of β_0 is available from some previous surveys. If instead of the regression estimator $\hat{Y}^*_{2m}$ described above, one uses the difference estimator $\hat{Y}^{**}_{2m} = \sum_{i\varepsilon s_m}(y^*_{2i}/p_i^*)\tilde{P}_i - [\sum_{i\varepsilon s_m}(z_i^*/p_i^*)\tilde{P}_i - Z]$ based on the matched sample, the expression for the minimum variance of $\hat{Y}_2$ would be as follows:

$$V_{\min}(\hat{Y}_2) = k(1-f+\sqrt{\tilde{\zeta}})\sigma_2^2/2 = \tilde{M}_1 \text{ (say)}$$

with

$$\tilde{\zeta} = (1+\tau^2 - 2\tau\delta)\zeta, \ \tau = \sigma_0/\sigma_3.$$

2.1.1 Variance Estimation

To get approximate unbiased estimators for $V_{\text{opt}}(\hat{Y}_2)$, we first present the following theorems without proof:

Theorem 1

$$\hat{V}(\hat{Y}^*_{2m}) = \{k/(1-k)\}\left[\left\{\sum_{i\varepsilon s_m}(y_{2i}^2\Delta_i/p_i^2)\tilde{P}_i/p_i^* - \hat{Y}_{2m}^{*2}\right\}\right.$$

$$\left. + \{k_2/k\}\sum_{i\varepsilon s_m}\tilde{P}_i\left(r_i^*/p_i^* - \sum_{i\varepsilon s_m}\tilde{P}_i r_i^*/p_i^*\right)^2\right]$$

is an unbiased estimator of $V(\hat{Y}^*_{2m})$, when β_0 is known, $k = (N-n)/\{n(N-1)\}$ and $k_2 = (n-m)/\{m(n-1)\}$.

Theorem 2

$E_1 V_2[\sum_{i\varepsilon s_m}\tilde{r}_i^*/p_1^*] = N(n-m)/\{nm(N-1)\}[\sigma_3^2 + \sigma_0^2 - 2\sigma_{03}]$ can be estimated unbiasedly by

$$\{(n-m)/n(m-1)\}\sum_{i\varepsilon s_m}(\tilde{r}_i^*/p_i^* - \sum_{i\varepsilon s_m}\tilde{r}_i^*/p_i^*)^2\tilde{P}_i$$

where $\tilde{r}_i^* = \tilde{r}_i\Delta_i/p_i$, $\tilde{r}_i = y_{2i} - z_i$; σ_3^2, σ_0^2 and σ_{03} are given in (4) and (7) respectively.

From the Theorem 2 we note that

$$\hat{\sigma}_0^2 = d\sum_{i\varepsilon s_m}\left(z_i/p_i^* - \sum_{i\varepsilon s_m} z_i\tilde{P}_i/p_i^*\right)^2\tilde{P}_i,$$

$$\hat{\sigma}_3^2 = d\sum_{i\varepsilon s_m}\left(y_{2i}/p_i^* - \sum_{i\varepsilon s_m} y_{2i}\tilde{P}_i/p_i^*\right)^2\tilde{p}_i$$

and

$$\hat{\sigma}_{30}^2 = d\sum_{i\varepsilon s_m}\left(z_i/p_i^* - \sum_{i\varepsilon s_m} z_i\tilde{P}_i/p_i^*\right)\left(y_{2i}/p_i^* - \sum_{i\varepsilon s_m} y_{2i}\tilde{P}_i/p_i^*\right)\tilde{P}_i$$

are unbiased estimators of σ_0^2, σ_3^2 and σ_{30}, respectively where $d = m(N-1)/\{N(m-1)\}$.

Estimator for $V_{opt}(\hat{Y}_2 I\lambda)$

Thus for a given value of m (*i.e.*, λ), we can suggest an approximate unbiased estimator of $V_{opt}(\hat{Y}I\lambda)$ as,

$$V_{opt}(\hat{Y}I\lambda) = (1/\hat{V}_m + 1/\hat{V}_u)^{-1},$$

where $\hat{V}_m = \hat{V}(\hat{Y}_{2m}^* I\beta_0)$ and $\hat{V}_u$ = an unbiased estimator of $V(\hat{Y}_{2u}) = \{(N-u)/N(u-1)\}\sum_{i\varepsilon s_u} P_i'(y_{2i}/p_i - \hat{Y}_{2u})^2$.

Estimator for $V_{min}(\hat{Y}_2)$

Putting suitable estimators for λ, ζ^* and σ_2^2 in the expression for $V_{min}(\hat{Y}_2)$, we get an approximate unbiased estimator for $V_{min}(\hat{Y}_2)$ as,

$$\hat{V}_{min}(\hat{Y}_2) = k[1 - f + (1-\hat{\lambda})\hat{\zeta}^*/\hat{\lambda}]/\hat{\sigma}_2^2,$$

where

$$\hat{\zeta}^* = (1-\hat{\delta}^2)\hat{\zeta}, \hat{\lambda} = \sqrt{\hat{\zeta}^*}/(1+\sqrt{\hat{\zeta}^*}),$$

$$\hat{\delta} = \hat{\sigma}_{03}/(\hat{\sigma}_0^2\hat{\sigma}_3^2)^{1/2}, \hat{\zeta}^* = \hat{\sigma}_3^2/\hat{\sigma}_2^2,$$

$$\hat{\sigma}_2^2 = \hat{\lambda}\hat{\sigma}_2^2(m) + (1-\hat{\lambda})\hat{\sigma}_2^2(u)$$

$\hat{\sigma}_2^2(m)$ = an approximate unbiased estimator of σ_2^2 based on the matched sample $s_m = \sum_{i\varepsilon s_m}(y_{2i}^2\Delta_i/p_i^2)\tilde{P}_i/p_i^* - \{\hat{Y}_{2m_2}^{*2} - \hat{V}_m\}$, $\hat{\sigma}_2^2(u)$ = an approximate unbiased estimator of σ_2^2 based on the un-matched sample $s_u = u(N-1)/\{N(u-1)\}\sum_{i\varepsilon s_u} P_i'(y_{2i}P_i'/p_i - \hat{Y}_{2u})^2$; k and f are as in (4).

Remark 2.2

Ideally one should estimate σ_2^2 through the optimum combination of $\hat{\sigma}^2(m)$ and $\hat{\sigma}^2(u)$ and in this case, the optimum combination will involve unknown parameters. To avoid this complexity, the simpler estimator $(\hat{\sigma}^2)$ of σ^2 has been suggested above.

2.2. Strategy 2

The population is supposed to consist of L strata with N_h as the unknown size of the h-th stratum ($h = 1, ..., L$; $\sum_h N_h = N$) stipulating that one can identify the stratum to which a unit belongs, as soon as its value is observed on the first occasion. On the first occasion, the initial sample s_1 of size n was selected by PPSWR method with normed size p_i attached to the i-th unit. Let n_h units of s_1, falling in the h-th stratum, be denoted as s_{1h}. Let $y_{1i}(h), y_{2i}(h)$ be respectively the value of the variate under study, of the i-th unit of the h-th stratum for the first and second occasions, and $z_i(h)$ be the corresponding size measure. On the second occasion, independent samples s_{mh}'s of sizes $m_h = m n_h/n$ (assumed an integer for every h), keeping $\sum_h m_h = m$ as fixed, are selected by the RHC sampling scheme with normed size $q_{hi}^* = [y_{1i}(h)/z_i(h)]/\sum_{i\varepsilon s_1}[y_{1i}(h)/z_i(h)]$ for the i-th unit of h-th stratum. The unmatched sample s_u was selected from the entire population by the RHC method with normed size measure p_i for the i-th unit as in strategy 1. The proposed estimators for Y_2, based on the matched-sample s_m, and the un-matched sample s_u are respectively as follows:

$$\hat{Y}_{2m} = \sum_h w_h \hat{Y}_{2m}(h);\ \hat{Y}_{2u} = \sum_{s_u}(y_{2i}/p_i)P_i' \qquad (9)$$

where

$$\hat{Y}_{2m}(h) = \sum_{s_{mh}} r_i(h)Q_{hi}/(n_{1h}p_{hi}q_{hi}^* + c_h\sum_{s_{1h}} z_j(h)/(n_{1h}p_{hj}), w_h = n_{1h}/n, p_{hj} = z_j(h)/Z,$$

$$r_i(h) = y_{2i}(h) - c_h y_{1i}(h),$$

Q_{hi} = sum of q_{hj}^* for the group containing i-th unit of the h-th stratum, that was formed for selection of the matched sample s_{mh} by RHC method. c_h's are constants chosen to minimize variance of $\hat{Y}_{2m}(h)$. Following Arnab (1991), the expression for variance of $\hat{Y}_{2m}$ is obtained as:

$$V(\hat{Y}_{2m}) = k_2\sum_h\sum_{j=1}^{N_h} q_{hj}(r_{hj}/q_{hj} - R_h)^2/P(h) + \sigma_2^2/n$$

where $k_2 = (n-m)/n, q_{hj} = y_{1j}(h)/y_1(h), Y_1(h) = \sum_{j=1}^{N_h} y_{1j}(h), N_h$ = population size of the h-th stratum, $P(h) = Z_h/Z, Z = \sum_{j=1}^{N_h} z_j(h)$.

The optimum value of c_h that minimizes $V(\hat{Y}_{2m})$ and the corresponding value of $V(\hat{Y}_{2m})$ comes out respectively as

$$\text{opt } c_h = c_h(0) = \delta_{h3} = \sum_{j=1}^{N_h} q_{hj}\alpha_{hj}\beta_{hj}/(\sigma_{h0}\sigma_{h3})$$

and $[1 + (n-m)\theta/m]\sigma_2^2/n$, where

$$\alpha_{hj} = y_{2j}(h)/q_{hj} - Y_2(h),\ \beta_{hj} = z_{hj}/q_{hj} - Z_h,$$

$$\sigma_{h3}^2 = \sum_{j=1}^{N_h} q_{hj}\alpha_{hj}^2, \sigma_{h0}^2 = \sum_{j=1}^{N_h} q_{hj}\beta_{hj}^2, Y_2(h) = \sum_{j=1}^{N_h} y_{2j}(h)$$

and $\theta = \sum_h(1-\delta_h^2)\sigma_{h3}^2/\{P_h\sigma_2^2\}$.

The proposed composite estimator for Y_2, the optimum proportion of matched sample and the expression for the minimum variance of the composite estimator $\hat{Y}_2$ are given respectively by

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

$$\text{opt}\,\lambda = \lambda_0 = [\theta - (1-f)\sqrt{\theta}\sqrt{f^*}]/[\theta + f\sqrt{\theta}\sqrt{f^*} - 1]$$

$$V_{\min}(\hat{Y}_2) = k(1/\mu_0 - f)\sigma_2^2/[1 + (\lambda_0/\mu_0)\sqrt{f^*}/\sqrt{\theta}]$$

$$= M_2 \text{ (say)}$$

where $\hat{Y}_{2m}$ and $\hat{Y}_{2u}$ are given in (9), $f^* = N/(N-1)$, $\mu_0 = 1 - \lambda_0$; k, f and σ_2^2 are given in (4).

3. EFFICIENCIES OF THE PROPOSED STRATEGIES

The proposed Strategy 1 is more efficient than the strategy proposed by Prasad and Graham (1994) in the sense of yielding smaller minimum variance, as $\delta^2 \le 1$. Efficiency of the Strategy 1 increases as δ, the correlation between y_{2i}/q_i and z_i/q_i increases. The efficiency of the Strategy 1 and Prasad and Graham's (1994) strategy increases as ζ decreases. The value of $\zeta = \sigma_3^2/\sigma_2^2$ depends on the magnitudes of σ_3^2 and σ_2^2. σ_3^2 will be smaller (greater) than σ_2^2 if the proportionality of y_{2i} on y_{1i} is higher (lower) than that of y_{2i} on z_i. Obviously, Strategy 1 can be used in practice when a good guess value of β is available from the past surveys. If the difference estimator is used in Strategy 1 instead of the regression estimator mentioned in Remark 2.1, then the proposed Strategy 1 fares better than that of Prasad and Graham (1994) whenever $\delta > \frac{1}{2}\sigma_0/\sigma_3$. Strategy 1 fares better or worse than Chotai's (1974) strategy according to $\zeta^* = (1 - \delta^2)\zeta <$ or $> (1 - \delta^{*2})$. Here, δ^* may be regarded as a correlation coefficient between y_{2i}/p_i and y_{1i}/p_i. In particular, if z_i's, are constant, then δ^* becomes the simple correlation coefficient between y_{1i}'s and y_{2i}'s. The expression for the minimum variance M_2 for Strategy 2 is complex and does not yield any simple comparison with the other strategies described here. However, we note that the efficiency of the Strategy 2 increases as the stratum correlation δ_{h3} increases. Following numerical examples based on the live data reveals that the proposed Strategy 2 fares better than Strategy 1 and also the alternatives proposed by Prasad and Graham (1994) and Chotai (1974).

For numerical comparisons, three data sets are considered. One of them (will be called Population 1) was considered by Prasad and Graham (1994) which relates to the area under wheat in 1937 (y_2) and 1936 (y_1) and cultivated area (z) for a set of 34 villages in India, compiled by Sukhatme and Sukhatme (1970). The population 1 is stratified in two strata in accordance with area under wheat in 1936 less than or more than 200 acres. Parameters for this population are: $N = 34$, $N_1 = 20$, $N_2 = 14$, $\delta^* = .7635$, $\delta = .3638$, $\zeta = .3811$, $\theta = .2436$. The Population 2 comprises of production of cereals in South America for the years 1980 (z), 1988 (y_1) and 1989 (y_2), compiled from The Statistical year book, United Nations (1988/89). The population is stratified in two strata considering 1988 production of more or less than 570 (thousand metric tons). The parameters for this population 2 are: $N = 19$, $N_1 = 7$, $N_2 = 12$, $\delta^* = -.6939$, $\delta = .7666$, $\zeta = 1.1478$, $\theta = .3681$. The population 3 compiled by Singh and Chaudhuri (1986) relates to the area under wheat in hector during 1979-80 (y_2) and 1978-79 (y_1) and total cultivated area in 1978-79 (z) of 16 villages of Meerut District. The parameters for the population 3 are: $N = 16$, $N_1 = 9$, $N_2 = 7$, $\delta^* = .7729$, $\delta = .1057$, $\zeta = .3965$, $\theta = .2827$.

The following table shows relative efficiencies of the proposed Strategies 1, 2 and the one proposed by Prasad and Graham (1994) with respect to Chotai (1974) which are respectively denoted by $E_1 = V_c/M_1$, $E_2 = V_c/M_2$ and $E_3 = V_c/V_{PG}$.

Table 1
Efficiencies of the Strategies

f	Population 1			Population 2			Population 3		
	E_1	E_2	E_3	E_1	E_2	E_3	E_1	E_2	E_3
.05	1.0463	1.1033	1.0181	1.0196	1.0850	.8262	1.0053	1.0864	1.0030
.10	1.0479	1.0895	1.0187	1.0202	1.0711	.8212	1.0055	1.0711	1.0031
.15	1.0496	1.0776	1.0194	1.0209	1.0579	.8172	1.0057	1.0577	.0033
.20	1.0514	1.0683	1.0200	1.0216	1.0519	.8123	1.0058	1.0469	1.0034
.25	1.0533	1.0622	1.0208	1.0224	1.0490	.8071	1.0061	1.0396	1.0035
.30	1.0554	1.0604	1.0216	1.0232	1.0530	.8017	1.0063	1.0368	1.0036

From the above table, we note that in all the three populations, Strategy 2 fares better than the others. It is also worth noting that both the proposed strategies fare better than those of Chotai (1974) and Prasad and Graham (1994). For the population 1, $\zeta = .3811$ which is quite favourable for Prasad and Graham's (1994) strategy, hence for the proposed Strategy 1. Both Prasad and Graham's strategy and Strategy 1, performed better than Chotai's (1974) strategy. For the population 2, $\zeta = 1.1478$ which is high and unfavourable for Prasad and Graham's (1994) strategy, but $\delta = .7666$ is quite favourable to Strategy 1. Hence, for the population 2, Prasad and Graham's strategy becomes less efficient than that of Chotai (1974), but the proposed Strategy 1 remains better. For the population 3, $\zeta = .3965$ which is quite favourable for Prasad and Graham (1994) but at the same time $\delta^* = .7729$ and this (δ^*) favours Chotai (1974). In fact Chotai's (1974) strategy is marginally inferior to Prasad and Graham's (1994) strategy but the proposed Strategy 2 remains better than both. It should be noted that the examples shown here are quite unusual in the

sense that they present low correlation between y_2 and z (in example 1, $\delta = .3638$ and in example 3, $\delta = .1057$) and there is a negative correlation between y_2 and y_1 ($\delta^* = -.6939$) in example 2. The correlations δ and δ^* are expected to be high and positive. Hence, further investigation is needed to compare the performances of the present strategies with suitable data.

Table 2
Sensitivity of Efficiency $E^* = V_{PG}/M_{\tilde{\beta}}$

$\lvert v \rvert$	.05	.10	f .15	.20	.25	.30
			Population 1			
0	1.028	1.029	1.030	1.031	1.032	1.033
.2	1.027	1.027	1.028	1.029	1.031	1.032
.4	1.023	1.024	1.027	1.026	1.027	1.028
.6	1.017	1.108	1.019	1.019	1.020	1.021
.8	1.010	1.010	1.010	1.011	1.011	1.011
1.0	1.000	1.000	1.000	1.000	1.000	1.000
1.2	.989	.988	.988	.988	.988	.987
1.4	.976	.976	.975	.974	.973	.972
			Population 2			
0	1.234	1.241	1.249	1.257	1.266	1.278
.2	1.219	1.227	1.233	1.241	1.249	1.258
.4	1.180	1.186	1.191	1.197	1.204	1.211
.6	1.125	1.128	1.133	1.137	1.141	1.146
.8	1.063	1.065	1.067	1.068	1.070	1.073
1.0	1.000	1.000	1.000	1.000	1.000	1.000
1.2	.939	.938	.936	.935	.933	.931
1.4	.883	.880	.877	.875	.871	.869
			Population 3			
0	1.002	1.002	1.004	1.003	1.003	1.003
.2	1.002	1.002	1.002	1.002	1.003	1.002
.4	1.002	1.002	1.002	1.002	1.002	1.002
.6	1.001	1.002	1.002	1.002	1.002	1.001
.8	1.001	1.001	1.001	1.001	1.001	1.001
1.0	1.000	1.000	1.000	1.000	1.000	1.000
1.2	.999	.999	.999	.999	.999	.999
1.4	.998	.997	.998	.998	.998	.998

To study the effect of departure of the optimum value of $\beta = \beta_0$ when some guess value of β is used in Strategy 1, one may consider sensitivity of efficiency of $\hat{Y}_2$ for the Strategy 1 for different choices of β, following Prasad and Srivenkataramana (1980). The minimum variance of $\hat{Y}_2$ for the Strategy 1 when some guess value of $\beta_0 = \tilde{\beta}$ is used, produces

$$V_{\min}(\hat{Y}_2 \mid \tilde{\beta}) = k(1 - f + \sqrt{\zeta^{**}})\,\sigma_2^2/2 = M_{\tilde{\beta}} \qquad (9)$$

where $\zeta^{**} = [1 - (1 - v^2)\,\delta^2]\,\zeta$ and $v = 1 - \tilde{\beta}/\beta_0$.

From (9), we note that the proposed Strategy 1 with the guess value $\tilde{\beta}$ fares better or worse than Prasad and Graham's (1994) strategy according to $|v| < 1$ or $|v| > 1$. Similarly, the proposed Strategy 1 with $\beta = \tilde{\beta}$ performs better or worse than Chotai's (1974) strategy according to $v^2 > \text{or} < (1 - 1/\delta^2)(1 - 1/\zeta)$. Table 2 proceeds sensitivity E^* of the estimator $\hat{Y}_2$ compared to Prasad and Graham's (1994) strategy where $E^* = V_{PG}/M_{\tilde{\beta}}$. From the Table 2, the loss with $v > 1$ is likely to be more than the gain with $v < 1$ for population 1 and population 3 but the situation is reverse for population 2.

CONCLUSION

In sampling over two occasions, one should utilize data collected on the first occasion to get an efficient estimator for the population total on the second occasion. Chotai (1974) used data collected on the first occasion at the stage of estimation, while Prasad and Graham did so at the stage of selection (and hence estimation) of the matched sample. In this article, two strategies have been proposed. The first one utilizes data collected at the first occasion for the selection of the matched sample similar to Prasad and Graham and formation of a regression estimator as determined by Chotai (1974). These make Strategy 1 more efficient than that of Prasad and Graham. The proposed Strategy 2 utilized first occasion values as a stratification variable, measure of size for the selection of the matched sample for the second occasion, and formation of a regression type estimator involving auxiliary variable (z), available on the first occasion. Intuitively one should expect the proposed Strategy 2 to perform better than the others mentioned here, but no theoretical result was established due to the complexity of the expression for the minimum variance of the proposed estimator. However, superiority of the Strategy 2 was established through numerical data.

ACKNOWLEDGEMENTS

The author is grateful to the referee, Associate Editor and the Editor for their valuable comments that substantially improved the earlier version of this paper. This work was supported by the FRD, South Africa.

REFERENCES

ARNAB, R. (1991). On sampling over two occasions using varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, 43(3), 282-290.

AVADHANI, M.S., and SUKHATME, B.V. (1970). A comparison of two sampling procedures with an application to successive sampling. *Applied Statistics*, 19, 251-259.

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

PRASAD, N.G.N., and SRIVENKATARAMANA, T. (1980). A modification to the Horvitz-Thompson estimator under the Midzuno sampling scheme. *Biometrika*, 67, 709-711.

PRASAD, N.G.N., and GRAHAM, J.E. (1994). PPS sampling over two occasions. *Survey Methodology*, 20, 59-64.

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

SINGH, D., and CHAUDHURI, F.S. (1986): *Theory and Analysis of Sample Survey Designs*. India: Wiley Eastern Limited, 166.

SINGH, M.P. (1967). The relative efficiency of some two-phase sampling schemes. *Annals of Mathematical Statistics*, 38, 937-940.

SUKHATME, P.V., and SUKHATME, B.V. (1970). *Sampling Theory of Surveys with Applications*. Ames, Iowa: Iowa State University Press, 185.

UNITED NATIONS (1992). *Statistical Year Book*, (1988/89). New York: United Nations, 356.