ON A PRAGMATIC MODIFICATION OF SURVEY SAMPLING IN THREE STAGES

ARIJIT CHAUDHURI
Department of Mathematics and Statistics
University of Nebraska-Lincoln, Lincoln, NE 68588-0323, USA

Key Words and Phrases: Sampling in three stages; unbiased estimation; variance estimator; varying probabilities.

ABSTRACT

Presented are formulae for an unbiased estimator of a finite population total and an unbiased variance estimator for it when samples are taken by usual procedures in the first two stages with varying probabilities but the third stage units are sampled for economy and convenience in a non-standard way from the pool of all sampled second stage units rather than independently from each of the latter separately containing the former.

1. INTRODUCTION

We consider a rural survey in a district containing a number of blocks of villages concerning some aspects of bank transactions of the villagers. The blocks are the first stage units (fsu), the villagers are the second stage units (ssu), and the account holding village customers of the banks within the district are the third stage units (tsu) considered for drawing a suitable sample adopting a three stage design. It is considered convenient to choose the tsu's using the bank ledgers. Also for the sake of economy and quick results it is decided to confine queries among the bank account holders living only within the selected villages. Samples of dwellers of the selected villages may of course be contacted for investigations on other matters of interest and compared with their bank related affairs which may be gathered from the bank ledgers. In section 2 we illustrate a specific sampling procedure with unequal probabilities in the first two stages in a usual way . As an innovation, the third stage sampling scheme we recommend for quick and cheap results in choosing a simple random sample (SRS) without replacement (WOR) from the collection of bank accounts of customers resident only in the selected group of villages. In standard three stage sampling, from each selected village the bank account holding customers are required to be sampled separately and independently across the villages. For the standard three stage sampling schemes, procedures of unbiasedly estimating the population total and unbiasedly estimating the variance of the estimator of the population total are well known from the works of Raj (1968) and Rao (1975) among others. As we are proposing a departure from the traditional way revised estimation procedures are called for and these are presented in section 3 below in a non-trivial way.

2. A SPECIMEN OF A REVISED THREE STAGE SAMPLING PLAN

Let $U = (1, \cdots, i, \cdots, N)$ denote a population of N elements called the first stage units (fsu) with i bearing an unknown value y_i of a variable y of interest and a known normed positive size-measure p_i $(i = 1, \cdots, N)$. Our problem is to estimate the population total $Y = \sum y_i$. By $\sum$ we mean summation over i in U. The fsu i in its turn consists of M_i second stage units (ssu). The jth ssu of ith fsu has the unknown value y_{ij} for y and a known normed positive size-measure p_{ij}. To draw a sample of n fsu's we apply Rao, Hartley and Cochran's (RHC, 1962) scheme using $p_i, i \in U$. We prescribe RHC scheme because it uses known size-measures, is unconditionally applicable, chooses distinct units in the sample, yields higher efficiency than sampling with probability proportional to size with replacement and ensures a non-negative unbiased variance estimator of a total it produces. To implement this scheme, U is divided at random into n groups, taking N_i fsu's in the i th group. This N_i is chosen as the integer part of $\frac{N}{n}$ or 1 added to it subject to the requirement $\sum_n N_i = N$. By $\sum_n$ we mean summation over n groups formed above. From each group one unit is selected with probability proportional to p_i. The selection is done independently across the groups. For simplicity by (y_i, p_i) we denote the value of y and the normed size-measure of the unit chosen from the ith group $(i = 1, \cdots, n)$. We also write s for the sample of n units thus chosen, $\sum'$ for sum over the units in s and $\sum' \sum'$ for sum over the units $i, i'(i \neq i')$ chosen in s. By Q_i we mean the sum of the p_i's of the units falling in the ith group. We also apply the scheme of RHC to draw a sample of m_i ssu's from the ith fsu, if the latter is selected, repeating this independently for every selected fsu. So, we split the M_i ssu's into m_i groups at random taking N_{ij} ssu's in jth group choosing N_{ij} as the integer part of $\frac{M_i}{m_i}$ or 1 added to it subject to $\sum_{m_i} N_{ij} = M_i$. By

$\sum_{m_i}$ we denote summation over the m_i groups formed above. From every group, say the jth, one unit is chosen with a probability proportional to p_{ij}, and this is repeated independently across the groups. For simplicity we write (y_{ij}, p_{ij}) for the values of y and the normed size-measure of the unit selected from the jth group $(j = 1, \cdots, m_i)$. By Q_{ij} we shall denote the sum of the p_{ij}'s falling in the jth group. Let s_i denote the set of ssu's chosen from the ith fsu assuming the latter is selected; $A(s_i)$ the set of bank accounts of customers with dwelling addresses in s_i and $L_i(s_i)$ = cardinality of $A(s_i)$. Let s_i' be an SRSWOR of $l_i(s_i)$ tsu's chosen from $A(s_i)$; this is to be repeated independently across the selected fsu's. This is our proposed specimen of a modified version of three-stage sampling, introduced to achieve economy and speedy execution.

3. UNBIASED ESTIMATOR OF TOTAL AND UNBIASED VARIANCE ESTIMATOR

For Y the unbiased estimator given by RHC is $t = \sum_n y_i \frac{Q_i}{p_i}$ admitting an unbiased variance estimator $v_1(t) = A(\sum_n y_i^2 \frac{Q_i}{p_i^2} - t^2)$, where $A = \frac{\sum_n N_i^2 - N}{N^2 - \sum_n N_i^2}$. However, y_i is supposed non-ascertainable though unbiasedly estimable by $e_i = \sum_{m_i} y_{ij} \frac{Q_{ij}}{p_{ij}}$ which has an unbiased variance estimator $v_2(e_i) = A_i(\sum_{m_i} y_{ij}^2 \frac{Q_{ij}}{p_{ij}^2} - e_i^2)$, where $A_i = \frac{\sum_{m_i} N_{ij}^2 - M_i}{M_i^2 - \sum_{m_i} N_{ij}^2}$. Again y_{ij}'s are non-ascertainable too but are unbiasedly estimable by

$$w_{ij} = \frac{L_i(s_i)}{l_i(s_i)} \sum_{k \in s_i'} y_{ijk} = L_i(s_i) \bar{y}_{ij}, \text{ (say)}.$$

Here, by y_{ijk} we mean the value of y for the kth bank customer living only in the jth village of the ith block. We shall write E_r, V_r to denote operators of expectation, variance over sampling in the rth stage $(r = 1, 2, 3)$ and E, V for over-all expectation, variance. By $C_3(j, j')$ we shall denote the covariance between w_{ij} and $w_{ij'}$ for $j \neq j'$. Further, we shall write

$$E_{12}(.) = E_1 E_2(.), E_{123} = E_1 E_2 E_3(.)$$

$$V_{12} = E_{12}(. - E_{12}(.))^2, V_{123} = E_{123}(. - E_{123}(.))^2 ,$$

the 'dot' here stands for a random variable generated by the sampling design employed.

Let, $\zeta_i = \sum_{m_i} \frac{Q_{ij}}{p_{ij}} w_{ij}, e = \sum_n \frac{Q_i}{p_i} e_i$ and , $u = \sum_n \frac{Q_i}{p_i} \sum_{m_i} \frac{Q_{ij}}{p_{ij}} w_{ij}$. Then,

$$E_2(\zeta_i) = e_i,$$

$$V_2(\zeta_i) = \sum_{m_i} (\frac{Q_{ij}}{p_{ij}})^2 V_3(w_{ij}) + \sum_{m_i} \sum_{\substack{m_i \\ j \neq j'}} \frac{Q_{ij} Q_{ij'}}{p_{ij} p_{ij'}} C_{3i}(j, j').$$

It also follows that

$$v_3(w_{ij}) = \frac{L_i(s_i)(L_i(s_i) - t_i(s_i))}{t_i(s_i)(t_i(s_i) - 1)} \sum_{k \in s_i} (y_{ijk} - \bar{y}_{ij})^2$$

satisfies $E_3 v_3(w_{ij}) = V_3(w_{ij})$ and

$$\text{for } \hat{C}_{3i}(j, j') = \frac{L_i(s_i)(L_i(s_i) - t_i(s_i))}{t_i(s_i)(t_i(s_i) - 1)} \sum_{k \in s_i} (y_{ijk} - \bar{y}_{ij})(y_{ij'k} - \bar{y}_{ij'}) \,,$$

$E_3 \hat{C}_{3i}(j, j') = C_{3i}(j, j')$.

Let $v_{2i} = A_i(\sum_{m_i} \frac{Q_{ij}}{p_{ij}^2} w_{ij}^2 - \zeta_i^2)$. Then,

$$\begin{aligned} E_3(v_{2i}) &= A_i \left[\sum_{m_i} \frac{Q_{ij}}{p_{ij}^2} V_3(w_{ij}) + \left(\sum_{m_i} \frac{Q_{ij}}{p_{ij}^2} y_{ij}^2 - e_i^2 \right) - V_3(\zeta_i) \right] \\ &= v_2(e_i) - A_i \left[\sum_{m_i} \frac{Q_{ij}(1 - Q_{ij})}{p_{ij}^2} V_3(w_{ij}) - \sum_{m_i} \sum_{\substack{m_i \\ j \neq j'}} C_{3i}(j, j') \right] \end{aligned} \qquad (1)$$

Let us now express $v_1(t)$ in the form

$$v_1(t) = \sum{}' b_{si} y_i^2 + \sum{}' \sum{}' b_{sij} y_i y_j$$

with b_{si}, b_{sij} as constants independent of $\underline{Y} = (y_1, \cdots, y_N)$. In particular, $b_{si} = \frac{Q_i}{p_i}$ for $i \in s$. Also let,

$$\begin{aligned} v_2(t) &= \sum{}' b_{si} e_i^2 + \sum{}' \sum{}' b_{sij} e_i e_j \text{ and} \\ v_3(t) &= \sum{}' b_{si} \zeta_i^2 + \sum{}' \sum{}' b_{sij} \zeta_i \zeta_j. \end{aligned}$$

Then, we have the following theorems with easy proofs.

Theorem 1.

$$u = \sum_n \frac{Q_i}{p_i} \sum_{m_i} \frac{Q_{ij}}{p_{ij}} w_{ij} \text{ satisfies } E(u) = Y.$$

Proof: $E_2E_3(\sum_{m_i} \frac{Q_{ij}}{p_{ij}} w_{ij}) = E_2(\sum_{m_i} \frac{Q_{ij}}{p_{ij}} y_{ij}) = \sum_{j=1}^{M_i} y_{ij} = y_i$ and
$E(u) = E_1(\sum_n \frac{Q_i}{p_i} y_i) = Y$.
Theorem 2.

$$v_{2i}^{*} = v_{2i} - A_i \left[\sum_{m_i} \frac{Q_{ij}(1 - Q_{ij})}{p_{ij}^2} v_3(w_{ij}) - \sum_{\substack{m_i \\ j \neq j'}} \sum_{m_i} \hat{C}_{3i}(j, j')\right]$$

satisfies $E_3(v_{2i}^{*}) = v_2(e_i)$, using (1).
Theorem 3.

$$v_{12}(e) = v_2(t) - \sum_n \frac{Q_i}{p_i} v_3(e_i) \text{ satisfies } E_{12}v_{12}(e) = V_{12}(e).$$

Proof:

$$\begin{aligned}
E_2 v_{12}(e) &= (\sum{}' b_{si} y_i^2 - \sum{}' \sum{}' b_{sij} y_i y_j) - \sum{}' b_{si} V_2(e_i) \\
&= v_1(t) + \sum{}' b_{si} V_2(e_i) \\
E_{12} v_{12}(e) &= E_1 v_1(t) + E_1 \sum{}' b_{si} V_2(e_i) \\
&= V_1(t) + E_1 \sum_n (\frac{Q_i}{p_i})^2 V_2(e_i) \\
&= V_1 E_2(e) + E_1 V_2(e) \\
&= V_{12}(e).
\end{aligned}$$

Theorem 4.

$$\begin{aligned}
E_3 v_2(t) &= \sum{}' b_{si} \left[e_i^2 + \sum_{m_i} (\frac{Q_{ij}}{p_{ij}})^2 V_3(w_{ij}) + \sum_{\substack{m_i \\ j \neq j'}} \sum_{m_i} \frac{Q_{ij}Q_{ij'}}{p_{ij}p_{ij'}} C_{3i}(j, j')\right] \\
&\quad - \sum{}' \sum{}' b_{sij} \left\{e_i e_j + \sum_{\substack{m_i \\ j \neq j'}} \sum_{m_i} \frac{Q_{ij}Q_{ij'}}{p_{ij}p_{ij'}} C_{3i}(j, j')\right\} \\
&= v_2(t) + \sum{}' b_{si} V_3(\zeta_i) + \sum{}' \sum{}' b_{sij} \sum_{\substack{m_i \\ j \neq j'}} \sum_{m_i} \frac{Q_{ij}Q_{ij'}}{p_{ij}p_{ij'}} C_{3i}(j, j')
\end{aligned}$$

Theorem 5.

$$\begin{aligned}
v &= v_3(t) - \sum{}' b_{si} v_3(\zeta_i) + \sum{}' \sum{}' b_{sij} \sum_{\substack{m_i \\ j \neq j'}} \sum_{m_i} \frac{Q_{ij}Q_{ij'}}{p_{ij}p_{ij'}} \hat{C}_{3i}(j, j') \\
&\quad + \sum_n \frac{Q_i}{p_i} v_{2i}^{*} + \sum_n (\frac{Q_i}{p_i})^2 \left[\sum_{m_i} (\frac{Q_i}{p_i})^2 v_3(\bar{y}_{ij})\right.
\end{aligned}$$

$$+ \sum_{m_i} \sum_{\substack{m_i \\ j \neq j'}} \frac{Q_{ij} Q_{ij'}}{p_{ij} p_{ij'}} \hat{C}_{3i}(j, j') \bigg]$$

satisfies $E(v) = V(u)$.

Proof: Follows, using theorems 2-4 and observing that

$$\begin{aligned} V(u) &= V_{12} E_3(u) + E_{12} V_3(u) \\ &= V_{12}(e) + E_{12} \Big[\sum_n \left(\frac{Q_i}{p_i}\right)^2 \sum_{m_i} \left(\frac{Q_{ij}}{p_{ij}}\right)^2 V_3(y_{ij}) \\ &\quad + \sum_n \left(\frac{Q_i}{p_i}\right)^2 \Big\{ \sum_{m_i} \sum_{\substack{m_i \\ j \neq j'}} \frac{Q_{ij} Q_{ij'}}{p_{ij} p_{ij'}} C_{3i}(j, j') \Big\} \Big] \end{aligned}$$

Remark I. Our proposed estimator for Y is u and the variance estimator for u is v if one adopts our recommended specimen of a version of three stage sampling scheme.

Remark II. The work originates from the requirement of an actual survey carried out in Indian Statistical Institute, Calcutta.

BIBLIOGRAPHY

Raj, D. (1968). *Sampling Theory.* Mcgraw-Hill, N.Y.

Rao, J.N.K. (1975). "Unbiased variance estimation for multi stage designs", *Sankhyā C*, 37, 133-139.

Rao, J.N.K., Hartley, H.O. and Cochran, W.G. (1962). "On a simple procedure of unequal probability sampling without replacement", *Jour. Roy. Stat. Soc., B*, 24, 482-491.

Received November, 1995; Revised March, 1997.