On the Choice of a Strategy for the Ratio Method of Estimation

By T. J. Rao

Indian Statistical Institute, Calcutta

[Received April 1966. Revised July 1966]

SUMMARY

Three sampling strategies are considered: (i) $H(M, R_1)$ consisting of the Midzuno (1952)–Sen (1952) sampling scheme and the estimator

$$R_1 = X(\sum_{i\epsilon s} Y_i)/(\sum_{i\epsilon s} X_i)$$

of the population total Y, where the symbol $\sum_{i\epsilon s}$ indicates that the summation is over all units U_i contained in the sample s; (ii) $H(M, \hat{Y}_{BT})$ consisting of the Midzuno–Sen sampling scheme and the estimator $\hat{Y}_{BT} = \sum_{i\epsilon s}(Y_i/\pi_i)$, where π_i is the probability of inclusion of the ith unit in the sample; (iii) $H(\pi ps, R_n)$ consisting of the πps sampling scheme (Hanurav, 1965; 1967) and the estimator $R_n = (X/n)\sum_{i\epsilon s}(Y_i/X_i)$ of Y. It is shown that the strategy $H(\pi ps, R_n)$ is suitable for the method of ratio estimation. A direct application to cluster sampling is given at the end.

1. INTRODUCTION

THE method of ratio estimation is used for estimating the population total Y of a characteristic $\mathscr{Y}$, when we have auxiliary information on a characteristic $\mathscr{X}$ related to it. It consists in getting an estimator of the population ratio Y/X and then multiplying this estimator by the known population total X. The usual estimators suggested for estimating the population total are given by

$$R_1 = \frac{\sum_{i\epsilon s} Y_i}{\sum_{i\epsilon s} X_i} X, \tag{1}$$

$$R_n = \frac{X}{n}\sum_{i\epsilon s}\frac{Y_i}{X_i}. \tag{2}$$

Both these estimators are biased, and during the last decade various attempts have been made to construct unbiased ratio estimators. In most cases the bias of one of these estimators is estimated and then the estimator is corrected for its bias. Estimators of this kind were given by Hartley and Ross (1954), Murthy and Nanjamma (1959) and Nieto de Pascual (1961). It was pointed out by Rao (1966a) that these estimators could be obtained as linear combinations of the two above-mentioned estimators R_1 and R_n, and certain relations between these estimators are studied therein.

The other aspect of the problem is to get sampling schemes which provide unbiased ratio estimators. Consider the estimator $(\sum_{i\epsilon s} Y_i)/(\sum_{i\epsilon s} X_i)$. This will be unbiased if the probability of selecting the sample is given by

$$P_s = \sum_{i\epsilon s}(X_i/X)\Big/\binom{N-1}{n-1}.$$

Lahiri (1951) suggested a procedure for selecting a sample with probability proportional to total size, resulting in the required P_r. An easier method was given independently by Midzuno (1952) and Sen (1952). Their scheme consists in taking the first unit with probability proportional to size, and the remaining $n-1$ units by simple random sampling without replacement from the remaining $N-1$ units of the population. Following Hanurav (1967), we denote the Midzuno–Sen sampling scheme together with the estimator R_1 of Y (given by (1)) as the strategy $H(M, R_1)$.

Let π_i' be the probability of inclusion of the ith unit in the sample for the Midzuno–Sen sampling scheme. Then we have, at once, another choice of estimator,

$$\hat{Y}_{HT} = \sum_{i \in s} (Y_i/\pi_i'), \tag{3}$$

introduced by Horvitz and Thompson (1952). The Midzuno–Sen sampling scheme and the Horvitz–Thompson estimator together constitute the strategy $H(M, \hat{Y}_{HT})$.

We consider next the estimator R_n given by (2). This can be written as $\sum_{i \in s}(Y_i/(nX_i/X))$, and when we use a πps sampling scheme (Hanurav, 1967), the denominator of this estimator is π_i, the probability of including the ith unit in the sample ($\mathcal{X}$ being the measure of size). The estimator therefore reduces to

$$\sum_{i \in s}(Y_i/\pi_i),$$

where $\pi_i = np_i$ ($p_i = X_i/X$ assumed less than n^{-1}). We denote the πps design and the estimator R_n of Y by the strategy $H(\pi ps, R_n)$.

2. Comparison of the Strategies

In this section we present a discussion on the choice of a suitable strategy under the assumption of a certain super-population set-up (Cochran, 1946; Hanurav, 1967).

If $V(.)$ is the variance function, we know that

$$V[\sum(Y_i/\pi_i)] = \sum_{i=1}^{N} (Y_i^2/\pi_i) + \sum_{j=1}^{N}\sum_{\substack{i=1 \\ j \neq i}}^{N} \frac{Y_i Y_j}{\pi_i \pi_j}\pi_{ij} - Y^2. \tag{4}$$

We also have (Rao, 1966b)

$$V(R_1) = \sum_{i=1}^{N} T_i Y_i^2 + \sum_{j=1}^{N}\sum_{\substack{i=1 \\ j \neq i}}^{N} T_{ij} Y_i Y_j - Y^2, \tag{5}$$

where

$$T_i = \left\{X \Big/ \binom{N-1}{n-1}\right\} \sum_{\lambda} (X_i + X_\lambda^{(i)})^{-1},$$

$X_\lambda^{(i)}$ being the sum of the λth set of $n-1$ distinct X's not equal to X_i, the summation over λ being taken over the $\binom{N-1}{n-1}$ such sets; and

$$T_{ij} = \left\{X \Big/ \binom{N-1}{n-1}\right\} \sum_{\lambda} (X_i + X_j + X_\lambda^{(ij)})^{-1},$$

$X_\lambda^{(ij)}$ being the sum of the λth set of $n-2$ distinct X's other than X_i and X_j, and the summation over λ being taken over the $\binom{N-2}{n-2}$ such sets.

Let Θ_1 be the class of prior distributions θ satisfying

$$\mathscr{E}_\theta(Y_i|X_i) = aX_i, \tag{6a}$$

$$\mathscr{V}_\theta(Y_i|X_i) = \sigma^2 X_i^g \quad (g>1), \tag{6b}$$

$$\mathscr{C}_\theta(Y_i, Y_j|X_i, X_j) = 0, \tag{6c}$$

where $\mathscr{C}$ denotes covariance. In most of the situations met in practice, the parameter g is found to lie between 1 and 2.

With this model, we have

$$\begin{aligned}\mathscr{E}_\theta\{V(R_1)\} &= \sigma^2 \sum_{i=1}^{N}(T_i-1)X_i^g \\ &\quad + a^2\left\{\sum_{i=1}^{N}(T_i-1)X_i^g + \sum_{\substack{i=1\\ i\neq j}}^{N}\sum_{j=1}^{N}(T_{ij}-1)X_iX_j\right\} \\ &= \sigma^2\sum_{i=1}^{N}(T_i-1)X_i^g,\end{aligned}$$

$$\mathscr{E}_\theta\{V(\hat{Y}_{HT})\} = \sigma^2\sum_{i=1}^{N}\{X_i^g(1-\pi_i')/\pi_i'\} + a^2 V\left\{\sum_{i\in s}(X_i/\pi_i')\right\}$$

and

$$\mathscr{E}_\theta\{V(R_n)\} = \sigma^2\sum_{i=1}^{N}X_i^g(1-\pi_i)/\pi_i.$$

We have the following theorem.

Theorem 1. The sampling strategy $H(\pi ps, R_n)$ is superior to the strategy $H(M, R_1)$ in the Θ_1 sense.

Proof.

$$\begin{aligned}\mathscr{E}_\theta\left\{\frac{V(R_1)-V(R_n)}{\sigma^2}\right\} &= \sum_{i=1}^{N}X_i^{g-1}\{T_iX_i-(X/n)\} \\ &= \sum_{i=1}^{N}X_i^{g-1}\left\{T_iX_i - N^{-1}\sum_{i=1}^{N}(T_iX_i)\right\} \quad \text{since } \sum_{i=1}^{N}T_iX_i = NX/n \\ &= N\operatorname{cov}(T_iX_i, X_i^{g-1})\end{aligned}$$

and this is positive for $g>1$ since $T_iX_i \geq T_jX_j$ if and only if $X_i \geq X_j$. This result appears as a footnote in J. N. K. Rao (1966) who gives a different proof.

We now present a comparison between the strategies $H(\pi ps, R_n)$ and $H(M, \hat{Y}_{HT})$ which are in fact based on a πps sampling scheme and a non-πps sampling scheme respectively. In the latter scheme we have $\pi_i' = \alpha + \beta p_i$, where $p_i = X_i/X$ and $\alpha = (n-1)/(N-1)$, $\beta = (N-n)/(N-1)$. Using a πps scheme, the probability of inclusion of the ith unit in the sample is $\pi_i = np_i$.

Theorem 2. The sampling strategy $H(\pi ps, R_n)$ is superior to the strategy $H(M, \hat{Y}_{HT})$ in the Θ_1 sense if the sample size is greater than

$$n_0 = \frac{2g-3+\{4(N-1)(g-1)(2-g)+1\}^{\frac{1}{2}}}{2(g-1)},$$

when $1<g<2$. When $g=2$, the strategy $H(\pi ps, R_n)$ is always superior to $H(M, \hat{Y}_{HT})$.

Proof. We have

$$D = \mathscr{E}_\theta\left\{\frac{V(\hat{Y}_{HT})-V(R_n)}{\sigma^2}\right\} = \sum_{i=1}^{N}(\pi_i'^{-1}-\pi_i^{-1})X_i^2+(a^2/\sigma^2)V\left\{\sum_{i\in s}(X_i/\pi_i')\right\}$$

$$> \sum_{i=1}^{N}(\pi_i'^{-1}-\pi_i^{-1})X_i^2 = (N^2/n)\{\mathrm{cov}(X_i^2/\pi_i,\pi_i)-\mathrm{cov}(X_i^2/\pi_i',\pi_i')\}$$

$$= (N^2/n)\,\mathrm{cov}\{p_i, X_i^2\{p_i^{-1}-(p_i+\gamma)^{-1}\}\},$$

where $\gamma = \alpha/\beta$. Consequently

$$D > (N^2/n)\,X^2\gamma\,\mathrm{cov}\{p_i, p_i^{g-1}/(p_i+\gamma)\}. \tag{7}$$

We distinguish two cases.

(i) $1<g<2$

Since $n>n_0$, we have $n^2(g-1)-n(2g-3)-N(2-g)>0$, so that

$$n\gamma(g-1)/(2-g)>1>\pi_i = np_i. \tag{8}$$

Consequently, $p_i<\gamma(g-1)/(2-g)$, from which it follows that $p_i^{g-1}/(p_i+\gamma)$ is an increasing function of p_i and hence that $\mathrm{cov}\{p_i, p_i^{g-1}/(p_i+\gamma)\}$ is positive.

(ii) $g=2$

In this case $\mathrm{cov}\{p_i, p_i^{g-1}/(p_i+\gamma)\} = \mathrm{cov}\{p_i, p_i/(p_i+\gamma)\}>0$, and this completes the proof of the theorem.

Remark. The condition $n>n_0$ is therefore sufficient for the covariance to be positive. (This is due to the fact that we are omitting some positive terms in the derivation of (7).) For $g=1{\cdot}5$ we find that $n_0 = N^{\frac{1}{2}}$ and for $1{\cdot}5<g<1{\cdot}9$ we have $n_0<N^{\frac{1}{2}}$. This is not a serious restriction in practice. We present below a table of the minimum sample size, $[n_0]+1$, given by the sufficient condition for the *pps* strategy to be better than the non-*pps* strategy considered, the range of values of g being 1·1 (0·1) 1·9 and of N being 10 (10) 100.

TABLE 1

Minimum sample size required for the pps strategy to be better than the non-pps strategy

g \ N	10	20	30	40	50	60	70	80	90	100
1·1	7	10	13	16	18	20	22	24	25	27
1·2	5	8	10	12	13	15	16	17	18	19
1·3	5	7	8	10	11	12	13	13	14	15
1·4	4	6	7	8	9	10	10	11	12	12
1·5	4	5	6	7	8	8	9	9	10	10
1·6	3	4	5	6	6	7	7	8	8	9
1·7	3	4	4	5	5	6	6	7	7	7
1·8	2	3	4	4	4	5	5	5	6	6
1·9	2	2	3	3	3	4	4	4	4	4

From Table 1 we find that, although the restriction seems to be severe for the smaller values of g, it is a mild one for the larger values of g. In practice, the values of g are found to lie above 1·3 (Fairfield Smith, 1938; Mahalanobis, 1944). Moreover,

we are using 1 as the upper bound for π_i in (8), but in fact $\pi_i \ll 1$, which means that the restriction on the minimum sample size will be still milder.

From Theorems 1 and 2 it follows that the *πps* strategy is better than both the non-*πps* strategy considered here (in most of the practical situations) and the Midzuno–Sen strategy.

3. Application to Cluster Sampling

Suppose we have a population, grouped in N clusters. If M_i is the size of the ith cluster and $\bar{y}_i$ is the mean of the population characteristic $\mathscr{Y}$ in the ith cluster, the population mean of $\mathscr{Y}$ is

$$\bar{Y} = \left(\sum_1^N M_i \bar{y}_i\right) \Big/ \left(\sum_1^N M_i\right).$$

If we want to estimate $\bar{Y}$ from a sample of n clusters, we have at our disposal the biased estimators

$$\bar{Y}_1 = \frac{\sum_{i\in s}(M_i \bar{y}_i)}{\sum_{i\in s}(M_i)}, \qquad \bar{Y}_n = \sum_{i\in s} \bar{y}_i / n.$$

Since we may write $\bar{Y}_n$ in the form $\bar{Y}_n = n^{-1} \sum_{i\in s} M_i \bar{y}_i / M_i$ we may define a sampling scheme (denoted by npM) in which the probability that the ith cluster is included in the sample is proportional to M_i, and this, together with the estimator $\bar{Y}_n$, may be called the strategy $H(npM, \bar{Y}_n)$. Taking cluster size as our auxiliary characteristic, this is clearly a *πps* strategy in Hanurav's sense, and is therefore superior to the Midzuno-Sen scheme (with clusters as the sampling units) and $\bar{Y}_1$ as the estimator.

4. Conclusion

In the above sections we have seen that *πps* sampling schemes can be used for the choice of a suitable strategy for the ratio method of estimation and in cluster sampling. For *πps* sampling schemes and their properties the reader is referred to Hanurav (1967) where he has given a scheme for $n = 2$, and to Vijayan (1967) where a scheme is given for any n.

Acknowledgement

I am thankful to Mr K. Vijayan for discussions and helpful comments.

References

Cochran, W. G. (1946). Relative accuracy of systematic and stratified random samples for a certain class of populations. *Ann. Math. Statist.*, 17, 164–177.

Fairfield Smith, H. (1938). An empirical law governing soil heterogeneity. *J. Agr. Sci.*, 28, 1–23.

Hanurav, T. V. (1965). Optimum sampling strategies and some related problems. (Unpublished Ph.D. thesis; Indian Statistical Institute, Calcutta.)

—— (1967). Optimum utilization of auxiliary information: *πps* sampling of two units from a stratum. *J. R. Statist. Soc.* B, 29, 374–391.

Hartley, H. O. and Ross, A. (1954). Unbiased ratio estimators. *Nature*, 174, 270–271.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Ass.*, 47, 663–685.

Lahiri, D. B. (1951). A method of sample selection providing unbiased ratio estimates. *Bull. Int. Statist. Inst.*, 33(2), 133–140.

Mahalanobis, P. C. (1944). On large-scale sample surveys. *Phil. Trans.* B, 231, 329–451.

MIDZUNO, H. (1952). On the sampling system with probability proportionate to sum of sizes. *Ann. Inst. Statist. Math., Tokyo*, 3, 99–107.

MURTHY, M. N. and NANJAMMA, N. S. (1959). Almost unbiased ratio estimates based on interpenetrating sub-sample estimates. *Sankhyā*, 21, 381–392.

NIETO DE PASCUAL, J. (1961). Unbiased ratio estimators in stratified sampling. *J. Amer. Statist. Ass.*, 56, 70–87.

RAO, J. N. K. (1966). On the relative efficiency of some estimators in *pps* sampling for multiple characteristics. *Sankhyā*, 28, 61–70.

RAO, T. J. (1966a). On certain unbiased ratio estimators. *Ann. Inst. Statist. Math., Tokyo*, 18, 117–121.

—— (1966b). On the variance of the ratio estimator for Midzuno–Sen sampling scheme. *Metrika*, 10, 89–91.

SEN, A. R. (1952). Present status of probability sampling and its use in the estimation of farm characteristics. (Abstract.) *Econometrica*, 20, 103.

VIJAYAN, K. (1967). On an exact *πps* sampling scheme. (Submitted to *J. R. Statist. Soc.* B.)