ELSEVIER

Journal of Statistical Planning and
Inference 102 (2002) 179–187

journal of
statistical planning
and inference

www.elsevier.com/locate/jspi

Location-shifts for improved estimation under Midzuno–Sen sampling scheme

T. Srivenkataramana

Department of Statistics, Bangalore University, JnanaBharathi, Bangalore 560 056, India

Received 1 August 1999; received in revised form 1 November 1999

Abstract

Simple location-shifts for the study or auxiliary character are proposed under Midzuno–Sen sampling from a finite population. These aim at improving the efficiency of the classical Horvitz–Thompson estimator or the unbiased ratio estimator of a population total. It is demonstrated that the choice of the translation parameters is flexible. A few methods for assessing these parameters are outlined. The gain in efficiency of estimation is illustrated. © 2002 Elsevier Science B.V. All rights reserved.

MSC: primary 62 D 05

Keywords: Horvitz–Thompson estimator; Location-shift; Midzuno–Sen sampling scheme; Unbiased ratio estimator

1. Introduction

The well-known Midzuno–Sen (MS) scheme (Midzuno, 1952; Sen, 1952) is an elegant method for unequal probability sampling (UPS). Here the first sample unit is drawn with probability proportional to size (PPS) and the other units by simple random sampling without replacement (SRSWOR) from those units remaining in the population after the first draw. An added advantage with the scheme is that it accomplishes a probability proportional to aggregate size (PPAS) sample which makes the conventional ratio estimator exactly design unbiased. However, the use of the classical Horvitz–Thompson estimator (HTE) with the MS scheme is not quite efficient. To remedy this, efforts have been made to revise the initial probabilities of selection such that the first order inclusion probabilities Π_i are proportional to size (ΠPS) (e.g. see Rao, 1963; Asok and Sukhatme, 1978). But these generally impose stringent restrictions on the initial probabilities which are difficult to meet in survey practice (Brewer and Hanif, 1983, p. 25). As a consequence, the use of HTE after MS sampling is generally not preferred.

Suppose that $U=\{1,2,\ldots,N\}$ denotes a labelled finite population of N units. The study variate y and a related auxiliary variate x take real values (Y_i,X_i) on unit $i\in U$.

0378-3758/02/$ - see front matter © 2002 Elsevier Science B.V. All rights reserved.
PII: S0378-3758(01)00188-4

The X_i values are assumed to be positive and known for all the units. In this set-up UPS is often used for improved estimation and the character x provides the basis for P_i, the initial selection probabilities of units. With x as size measure, $P_i = X_i/X$, where $X = \sum_{i \in U} X_i$. We consider below, the estimation of the population total $Y = \sum_{i \in U} Y_i$ from a random sample $\mathscr{s}$ $(\subset U)$ of n units drawn without replacement, with particular focus on the MS scheme.

Horvitz–Thompson estimator

The classical HTE of Y is

$$\hat{Y} = \sum_{i \in \mathscr{s}} Y_i/\Pi_i \tag{1.1}$$

with variance

$$V(\hat{Y}) = \sum_{i,j \in U}{}_1 \Delta_{ij}(Y_i/\Pi_i - Y_j/\Pi_j)^2, \tag{1.2}$$

where $\sum_1$ denotes the sum over all the different pairs of units in the population, $\Delta_{ij} = (\Pi_i \Pi_j - \Pi_{ij})$ and Π_{ij} are second-order inclusion probabilities for the units i and j. It is immediately noted from (1.1) or (1.2) that the success of HTE hinges on near proportionality between Y_i and Π_i. In order to achieve this, we first consider a transformation of $\mathscr{y}$ by confining our attention to the class of schemes where Π_i is a linear function of X_i. That is, for $i = 1, 2, \ldots, N$

$$\Pi_i = v + \delta X_i, \tag{1.3}$$

where v and δ are constants which depend on the sampling scheme. For instance, the MS scheme belongs to this class and has

$$v = (n-1)/(N-1), \quad \delta = (N-n)/(N-1)X. \tag{1.4}$$

PPSWOR schemes which ensure $\Pi_i = nP_i$ also belong to this class and have $v = 0$, $\delta = n/X$. Examples for this are the commonly used PPx systematic sampling (Madow, 1949) and Sampford's (1967) rejective scheme. In order to motivate a suitable transformation of $\mathscr{y}$, consider the simple situation where, for $i = 1, 2, \ldots, N$

$$Y_i = \alpha + \beta X_i. \tag{1.5}$$

Then

$$Y_i/\Pi_i = (\alpha + \beta X_i)/(v + \delta X_i) \text{ is not a constant in general,}$$

whereas

$$\frac{Y_i - (\alpha - \beta v/\delta)}{\Pi_i} = \frac{\beta X_i + \beta v/\delta}{v + \delta X_i} = \beta/\delta \tag{1.6}$$

is a constant for all i. Motivated by this fact, consider the location-shift for $\mathscr{y}$ defined by

$$Y_i^* = (Y_i - \theta); \quad i = 1, 2, \ldots, N, \tag{1.7}$$

where $\theta = (\alpha - \beta v/\delta)$. For the MS scheme, v and δ are as in (1.4) and

$$\theta = \alpha - \beta(n-1)X/(N-1). \tag{1.8}$$

Thus, it follows that if a good assessment of θ can be made, even as late as the estimation stage in a sample survey, the sampling variance of $\hat{Y}$ can be controlled by a location shift as in (1.7). The methods for assessing θ are discussed in the next section. The proposed alternative to $\hat{Y}$ is

$$\hat{Y}_1 = \sum_{i\in s} (Y_i^*/\Pi_i) + N\theta, \tag{1.9}$$

which is readily noted to be unbiased for Y. Further, an expression for the sampling variance is given by (1.2), with Y_i^* now replacing Y_i.

2. Choice of θ

The choice of θ for which the variance is minimised is easily obtained by writing

$$V(\hat{Y}_1) = \underset{i,j\in U}{\sum_1} \Delta_{ij} \left[\left(\frac{Y_i}{\Pi_i} - \frac{Y_j}{\Pi_j}\right) - \theta\left(\frac{1}{\Pi_i} - \frac{1}{\Pi_j}\right)\right]^2$$

$$= V(\hat{Y}) + \theta^2\Delta_1 - 2\theta\Delta_2, \tag{2.1}$$

where

$$\Delta_1 = \underset{i,j\in U}{\sum_1} \Delta_{ij} \left(\frac{1}{\Pi_i} - \frac{1}{\Pi_j}\right)^2, \tag{2.2}$$

$$\Delta_2 = \underset{i,j\in U}{\sum_1} \Delta_{ij} \left(\frac{Y_i}{\Pi_i} - \frac{Y_j}{\Pi_j}\right)\left(\frac{1}{\Pi_i} - \frac{1}{\Pi_j}\right). \tag{2.3}$$

It is noted from (2.1) that $V(\hat{Y}_1)$ is minimised with respect to θ by the choice

$$\theta_{\text{opt}} = \Delta_2/\Delta_1 \tag{2.4}$$

and the corresponding

$$V_{\min}(\hat{Y}_1) = V(\hat{Y}) - \Delta_2^2/\Delta_1. \tag{2.5}$$

Thus, the reduction in sampling variance associated with optimum location shift for y is Δ_2^2/Δ_1. In general, $V(\hat{Y}_1) < V(\hat{Y})$ as long as θ lies between 0 and $2\theta_{\text{opt}}$. From (2.3), we notice that when $\{Y_i/\Pi_i\}$ differ considerably among themselves Δ_2 can be expected to be large and in this situation there will be sufficient flexibility in the choice of θ. This is precisely when the classical HTE by itself is not very efficient.

Interestingly, Δ_1 is free from the study character and it depends only on Π_i. Thus, for a given procedure like the MS scheme, Δ_1 can be computed. But Δ_2 cannot be computed in general as it depends on all the y-values. Stuart (1986) has shown that for any general UPS design, small fixed errors in determining θ_{opt} are not serious. If, instead of θ_{opt}, we use fixed $\theta = \theta_{\text{opt}}(1+\varepsilon)$, then the variance reduction is $(1-\varepsilon^2)\ \Delta_2^2/\Delta_1$. A glance at (1.8) and (2.4) shows that θ can be chosen in the following two different ways.

(1) By assessing α and β separately.
(2) By estimating θ directly.

A few guidelines for this purpose are given below.

Method 1: *Assessing* α *and* β: Since α and β may be considered as the $\mathscr{y}$-intercept and the slope in the population regression line of $\mathscr{y}$ on x, we may assess their values as follows.

(1a). Based on experience, past data or a pilot study make a careful assessment of α and β.

(1b). In a scatter plot of Y_i/Π_i versus X_i/Π_i following MS sampling, gauge the intercept on the vertical axis and also the slope of the best-fitting line. Use these, respectively, as α and β.

Method 2: *Estimating* θ: Obtain sample analogues of Δ_1 and Δ_2 after attaching weights Π_{ij}^{-1}. Thus estimate θ by

$$\hat{\theta} = \frac{\sum_{1 i,j \in \mathscr{s}} \Delta_{ij}(Y_i/\Pi_i - Y_j/\Pi_j)(1/\Pi_i - 1/\Pi_j)\Pi_{ij}^{-1}}{\sum_{1\ i,j \in \mathscr{s}} \Delta_{ij}(1/\Pi_i - 1/\Pi_j)^2 \Pi_{ij}^{-1}}. \tag{2.6}$$

A fixed θ in (1.9) does not affect sampling variance, while it is not so when a sample based estimate of θ, as in (2.6), is used. This impact can be illustrated by writing the estimator $\hat{Y}_1$ in (1.9) with θ estimated as

$$t = \hat{Y} - \hat{\theta} \sum_{i \in \mathscr{s}} \left(\frac{1}{\Pi_i} - \frac{N}{n} \right), \tag{2.7}$$

which has the form $(t_1 - t_2 \cdot t_3)$ involving three non-independent statistics. An interesting discussion in this context, pointing out that the use of $\hat{\theta}$ may not necessarily inflate variance is in Stuart (1986). In any case, the use of nonrandom α, β, as implied by *Method* 1, is simpler to implement though it has an element of subjectivity.

Particular case of $\alpha = 0$: In model (1.5) if we overlook the intercept term and take $\alpha = 0$, we have further simplified premises which will need the assessment of only β. In this set-up Narasimha Prasad and Srivenkataramana (1980) discuss a location shift for $\mathscr{y}$ after MS sampling while Rao (1988) proposes a location shift for x of the type

$$X_i^* = X_i + d\bar{X}, \quad i = 1, 2, \ldots, N, \tag{2.8}$$

where $\bar{X} = X/N$ and suggests the choice $d = -\{(n-1)N\}/\{n(N-1)\}$ which renders the scheme IPPS. However, for (2.8) to be operative we need $X_i^* > 0$ for all i. This in turn imposes the restriction that $P_i > (n-1)/n(N-1)$ for all i. A translation of x has the advantage that its values are readily available while translation of $\mathscr{y}$ allows the flexibility of using a separate shift for each study variate when multiple characteristics are being estimated from the same sample. It may be remarked that the approach of the present paper is more general as it does not implicitly force the intercept term α to be zero. For other transformations on x, we refer to Bedi (1996).

3. Unbiased ratio estimator (URE)

A specific advantage with the MS scheme is that it provides a simple procedure for PPAS sampling which renders the ratio estimator

$$\hat{Y}_2 = X\left(\sum_{i\in s} Y_i \Big/ \sum_{i\in s} X_i\right) \tag{3.1}$$

exactly design unbiased for Y. A formal expression for variance of $\hat{Y}_2$ can be written down from first principles (Des Raj, 1968, p. 94) as

$$V(\hat{Y}_2) = \left(\frac{X}{N^*}\right)\sum\nolimits_2 \left[\left(\sum_{i\in s} Y_i\right)^2 \Big/ \sum_{i\in s} X_i\right] - Y^2, \tag{3.2}$$

where $\sum_2$ denotes the sum over all possible samples s and $N^* = \binom{N-1}{n-1}$ is the number of possible SRSWOR samples after the first draw. It follows from (3.2) that $V(\hat{Y}_2)$ vanishes when Y_i is proportionate to X_i. Also see Rao (1983). The premises in (1.5) thus motivate a location shift

$$Y_i^{**} = Y_i - \alpha, \quad i = 1, 2, \ldots, N \tag{3.3}$$

for y and prompt the unbiased ratio estimator

$$\hat{Y}_3 = X\left(\sum_{i\in s} Y_i^{**} \Big/ \sum_{i\in s} X_i\right) + N\alpha \tag{3.4}$$

with variance given by (3.2), where Y_i^{**} now replaces Y_i and $(Y - N\alpha)$ replaces Y.

A bit of algebra shows that we can write

$$V(\hat{Y}_3) = V(\hat{Y}_2) + \alpha^2\tau_1 - 2\alpha\tau_2, \tag{3.5}$$

where

$$\tau_1 = (nX/N^*)\sum\nolimits_2 (1/\bar{x} - 1/\bar{X}), \tag{3.6}$$

$$\tau_2 = (nX/N^*)\sum\nolimits_2 (\bar{y}/\bar{x} - \bar{Y}/\bar{X}) \tag{3.7}$$

and $\bar{y} = \sum_{i\in s} Y_i/n$, $\bar{X} = \sum_{i\in s} X_i/n$, $\bar{Y} = Y/N$. It follows from (3.5) that the variance minimizing location shift is given by

$$\alpha_{\text{opt}} = \tau_2/\tau_1 = \frac{\sum_2(\bar{y}/\bar{x} - \bar{Y}/\bar{X})}{\sum_2(1/\bar{x} - 1/\bar{X})} \tag{3.8}$$

with the corresponding

$$V_{\min}\,(\hat{Y}_3)=V(\hat{Y}_2)-\tau_2^2/\tau_1 \tag{3.9}$$

and τ_2^2/τ_1 is the reduction in variance provided by the optimum location shift. However, this cannot be determined, being dependent on all possible samples. In survey practice, α may be assessed as in *Method* 1 outlined in the previous section and as long as it is between 0 and $2\alpha_{\text{opt}}$, there will be a variance reduction.

Location-shift for x: We may promote a translation of the size measure x by rewriting (1.5) as

$$Y_i=\beta\,(X_i+\phi), \tag{3.10}$$

where $\phi=\alpha/\beta$. Thus one may consider a location shift

$$X_i^{**}=X_i+\phi \tag{3.11}$$

and use X_i^{**} as the size measure for MS sampling and construct an unbiased ratio estimator

$$\hat{Y}_4=X^{\|}\left(\sum_{i\in s} Y_i \Big/ \sum_{i\in s} X_i^{**}\right) \tag{3.12}$$

for Y, where $X^{**}=(X+N\phi)$ is the population total of the location shifted measure of size. Also see Bedi (1996). The sampling variance is now given by (3.2) with x_i replaced by $x_i^{\|}$. However, a simple representation of variance similar to that in (3.5) is not possible here in view of the parameter occuring in the denominator of the expression. In any case the variance minimising choice for ϕ is obtained as a solution of

$$\sum\nolimits_2\left[\left(\frac{\bar{y}}{(\bar{x}+\phi)}\right)^2(\bar{x}-\bar{X})\right]=0. \tag{3.13}$$

Again, for practical use $\phi=\alpha/\beta$ may be chosen by assessing α and β separately as outlined in the preceding section. However, the use of X_i^{**} as size measure imposes a restriction, viz.

$$X_i+\phi>0 \quad \text{for all } i=1,2,\ldots,N. \tag{3.14}$$

Since the study variate is directly related to the auxiliary variate, β is always positive. But this may not be the case with the intercept term α. If α is also positive then $\phi>0$ and the requirement (3.14) is readily met. On the other hand, a negative α renders $\phi<0$ and (3.14) needs the smallest X_i to exceed $-\phi$. If this is not true, one may either reset ϕ so that (3.14) is met or alternatively translate y as in (3.3).

4. Illustration

In order to compare the efficiency of the suggested estimators with that of the usual estimators under MS sampling, four populations are considered. Population I consists of the number of cattle (y) and the number of farms (x) in 13 clusters as in Table 1.

Table 1
Population I

x	:	19	28	28	30	31	46	51	53	55	56	61	64	83
y	:	168	326	396	360	331	697	586	739	914	930	619	784	906

Table 2
Populations A, B and C

Unit	X_i	y values for population		
		A	B	C
1	0.1	0.5	0.8	0.2
2	0.2	1.2	1.4	0.6
3	0.3	2.1	1.8	0.9
4	0.4	3.2	2.0	0.8
Total	1.0	7.0	6.0	2.5

Table 3
Efficiency E_1 of HTE

Population	θ_{opt}	Departure percentage $100 \cdot \lvert 1 - \theta/\theta_{\text{opt}} \rvert$			
		0	20	40	60
I	4217.44	239	226	200	159
A	−4.8390	5655	1755	644	269
B	−1.6662	724	586	389	224
C	−1.1479	294	273	231	173

The other three are small hypothetical populations considered by Yates and Grundy (1953) with details as shown in Table 2.

We examine the efficiency under MS sampling for $n = 2$ units and with estimation strategies as follows:

I. HTE without and with a translation of y as in (1.7).
II.URE without and with a translation of y as in (3.3).

The efficiencies $E_1 = 100V(\hat{Y})/V(\hat{Y}_1)$ under I and $E_2 = 100V(\hat{Y}_2)/V(\hat{Y}_3)$ under II for optimum choice of translation parameter as well as for specified departures from it are displayed in Tables 3 and 4. They show that the efficiency gain is often substantial and it is not too sensitive to departures from the optimum. For population C, α_{opt} is quite small and hence the translation $Y_i^{**} = Y_i - \alpha$ is unable to provide much improvement over the URE $\hat{Y}_2$. Similar is the case with Population I, relative to values of y.

5. Discussion

The usefulness of location shifts for study and auxiliary variates, y and x, respectively, is demonstrated for improving the efficiency of estimators in UPS from finite

Table 4
Efficiency E_2 of URE

Population	α_{opt}	Departure percentage $100 \cdot \|1 - \alpha/\alpha_{\text{opt}}\|$			
		0	20	40	60
I	−13.19	100.22	100.21	100.18	100.07
A	−0.5	681	551	353	220
B	0.5	681	551	353	220
C	−0.096	113	112	110	107

populations. The main focus is on the MS sampling scheme which is simple to operate and guarantees a non-negative Yates–Grundy variance estimator for HTE. The scheme also ensures a PPAS sample which renders the usual ratio estimator exactly design unbiased. However, the sampling variability of the HTE or URE is often unacceptably high so that MS sampling followed by HTE or URE is not generally preferred. In this scenario variate translations are proposed in order to reduce sampling variance. A linear relation between y and x is assumed to motivate suitable location shifts. The classical HTE and URE strategies are examined in this context. It is illustrated that an attractive gain in efficiency can be achieved with reasonable flexibility in the choice of the translation parameters. The methods for assessment of these parameters are also discussed.

A translation of y has the inbuilt advantages that (a) it can be effected after the sample arrives, and (b) separate location shifts can be made for different study variates as dictated by the relation with x. As a consequence of (a), the information in the current sample may be used for deciding the translation parameter. Advantage (b) particularly suits multivariate surveys with several important study characters.

A location shift for x is associated with the convenience that its values are known beforehand for the entire population. But these translations are subject to non-negativity conditions on the values of the location shifted x, to be used as a measure of size of the units, which may sometimes be very restrictive. The generalisation of the results of this paper to a stratified population is straightforward. Separate location shifts may be adopted in the different strata. It may be pointed out that translations of both y and x may be used in conjunction in a given survey. This is particularly handy in multivariate surveys where a translation of x may be followed by suitable individual translations for the y-characters. The modalities of such several translations are worth being looked into.

Acknowledgements

The author wishes to thank the referee for several useful suggestions and Dr. K. Harish Chandra for computational help.

References

Asok, C., Sukhatme, B.V., 1978. A note on Midzuno scheme of sampling—an abstract. J. Ind. Soc. Agric. Statist. 30 (2), 131.

Bedi, P.K., 1996. Efficient utilization of auxiliary information at estimation stage. Biometrical J. 8, 973–976.

Brewer, K.R.W., Hanif, M., 1983. Sampling with Unequal Probabilities. Lecture Notes in Statistics, Vol. 15. Springer, New York.

Madow, W.G., 1949. On the theory of systematic sampling II. Ann. Math. Statist. 20, 333–354.

Midzuno, H., 1952. On the sampling system with probability proportional to sum of sizes. Ann. Inst. Statist. Math. 3, 99–107.

Narasimha Prasad, N.G., Srivenkataramana, T., 1980. A modification to the Horvitz–Thompson estimator under the Midzuno–Sen sampling scheme. Biometrika 67, 709–711.

Rao, J.N.K., 1963. On two systems of unequal probability sampling without replacement. Ann. Inst. Statist. Math. 15, 67–72.

Rao, T.J., 1983. Transformation on the auxiliary variate for Midzuno–Sen sampling scheme. Technical Report No. 8/83, Stat-Math Division, Indian Statistical Institute, Calcutta.

Rao, T.J., 1988. Transformation on the auxiliary variate for Midzuno–Sen sampling scheme. J. Ind. Soc. Agric. Statist. 40 (3), 173–177.

Sampford, M.R., 1967. On sampling without replacement with unequal probabilities of selection. Biometrika 54, 449–513.

Sen, A.R., 1952. Present status of probability sampling and its use in estimation of farm characteristics, An abstract. Econometrica 27, 130.

Stuart, A., 1986. Location-shifts in sampling with unequal probabilities. J. Roy. Statist. Soc. Ser. A 149, 349–365.

Yates, F., Grundy, P.M., 1953. Selection with replacement from within strata with probability proportional to size. J. Roy. Statist. Soc. Ser. B 15, 253–262.