Sankhyā : The Indian Journal of Statistics
1988, Volume 50, Series B, Pt. 2, pp. 216–223.

A UNIFIED APPROACH TO ESTIMATION OF LORENZ-RATIO FROM A FINITE POPULATION

By P. MAITI and M. PAL

Indian Statistical Institute

SUMMARY. Most of the results relating to estimation of Lorenz-ratio (LR) are based on the assumption of some distribution on the population. Not much attention has been paid in the literature to provide a design-based unbiased estimator for LR. To the best of authors' knowledge, Taguchi's (1978) is the only one which however, can hardly be used in practice. In case$\overline{Y}$, the population mean is known, an effort, in this paper, has been made to provide some unbiased estimators of LR together with estimates for the variance of the estimators. The relevant expressions have been first found under the general sampling design and then, in particular, the cases of SRSWR and SRSWOR have been discussed.

1. Introduction

Estimation of LR, in the literature, is mainly based on the assumption of a reasonable distribution on income or total expenditure, e.g., Pareto, lognormal or Gamma, which have been found to fit the income distribution very well for many countries. Under the assumption of a continuous distribution and the scheme being that of SRS, estimation of LR is not very difficult.

Taguchi (1978) proposed a sampling scheme using information on an auxiliary variable and provided an unbiased estimator for LR under the assumed model describing the distribution of the auxiliary variable. Though $\overline{Y}$ need not be known to obtain an unbiased estimator for LR under Taguchi's scheme it can hardly be used in practice because of non-availability of such an auxiliary variable. Also in most cases the statistician has to analyse data already available, from surveys not conducted under Taguchi's scheme.

Here, a sample theoretic approach has been adopted to provide an unbiased estimate for LR defined in a finite population under a general sampling design and also the estimate for the variance of the estimator has been found out, following Nanjamma-Murthy-Sethi (1959) approach. In fact, direct application of N-M-S-approach has helped us to find out simple expressions for the estimate of the variance of the estimator, when the scheme is either SRSWR or SRSWOR. It may be pointed out that the usual estimator for LR can be made unbiased under SRSWR or SRSWOR simply by making a correction for the constant multiplication.

2. Definitions, notations and the usual results

Let $U = \{1, 2, ..., N\}$ be a finite population of N given units labelled through 1 to N and y be a variate taking value y_i for the i-th unit of the population.

LR under discrete set up may be defined as

$$LR_{(1)} = \sum_{i=1}^{N} \sum_{j=1}^{N} |y_i - y_j| / 2N^2 \bar{Y} \qquad \dots \quad (2.1)$$

where $\bar{Y}$ is the population mean of Y. Another way of defining LR may be as

$$LR_{(2)} = \sum_{i \neq j} |y_i - y_j| / 2N(N-1) \bar{Y} \qquad \dots \quad (2.2)$$

when $\bar{Y}$ is known, the commonly used esimators for LR are

$$\hat{LR}_{(1)} = \sum_{i,j}^{n} |y_i - y_j| / 2n^2 \bar{Y} \qquad \dots \quad (2.3)$$

and

$$\hat{LR}_{(2)} \sum_{i \neq j = 1}^{n} |y_i - y_j| / 2n(n-1) \bar{Y} \qquad \dots \quad (2.4)$$

It may be observed that under SRSWR and SRSWOR, the estimators defined in (2.3) and (2.4) are biased. However, they can be made unbiased after they are multiplied by the appropriate correction factors depending on n and N only. In particular,

$$\hat{LR}_A = [n/(n-1)] \; \hat{LR}_{(1)} \text{ and } \hat{LR}_B = [n(N-1)/N(n-1)] \; \hat{LR}_{(1)} \qquad \dots \quad (2.5)$$

are unbiased for $LR_{(1)}$ under SRSWR and SRSWOR respectively.

Unbiased estimates for LR_2 may be found in a straight forward manner once we known the unbiased estimates for LR_1. Since LR_1 and LR_2 differ only with respect to the constant multiplication, we shall discuss the estimation problem only for LR_1 in this paper and this will be referred as LR throughout our subsequent discussion.

After some routine calculation, $V(\hat{LR}_A)$ and $V(\hat{LR}_B)$ may be written as,

$$V(\hat{LR}_A) = \Big[\frac{3-2n}{N^2} \sum_{i \neq j \neq k \neq l} |y_i - y_j| \, |y_k - y_l|$$

$$+ \Big(\frac{12-8n}{N^2} - \frac{2n-4}{N} \Big) \sum_{i \neq j \neq k} |y_i - y_j| \, |y_i - y_k|$$

$$+ \Big(\frac{6-4n}{N^2} + \frac{2n-4}{N} + 1 \Big) \sum_{i \neq j} |y_i - y_j|^2 \Big] / (2N^2 n(n-1) \bar{Y}^2) \qquad \dots \quad (2.6)$$

which can equivalently be represented as

$$V(\hat{LR}_A) = \left[(-4n+6)LR^2+\frac{n-2}{N^3}\ \Sigma\ a_{i0}^2/\bar{Y}^2+C_y^2\right] \qquad \dots \quad (2.7)$$

when C_y^2 is the square of the co-efficient of variation of y in the population, and

$$V(\hat{LR}_B) = \Bigg[\left\{\frac{(n-2)(n-3)N(N-1)}{(N-2)(N-3)n(n-1)}-1\right\} \sum_{i\neq j\neq k\neq l} |y_i-y_j|\,|y_k-y_l|$$

$$+4\left\{\frac{(n-2)N(N-1)}{(N-2)n(n-1)}-1\right\} \sum_{i\neq j\neq k} |y_i-y_j|\,|y_i-y_k|$$

$$+2\left\{\frac{N(N-1)}{n(n-1)}-1\right\} \sum_{i\neq j} |y_i-y_j|^2\Bigg] /4N^4\bar{Y}^2 \qquad \dots \quad (2.8)$$

$$= A_1\, LR^2+\frac{A_2-A_1}{N^4\bar{Y}^2}\ \Sigma\ a_{i0}^2+\frac{A_1-2A_2+A_3}{N^2}\ C_y^2 \qquad \dots \quad (2.9)$$

where $$A_1 = \frac{(n-2)\,(n-3)\,N(n-1)}{(N-2)\,(N-3)\,n(n-1)} - 1,$$

$$A_2 = \frac{(n-2)\,N(N-1)}{(N-2)\,n(n-1)} - 1,$$

and $$A_3 = \frac{N(N-1)}{n(n-1)} - 1.$$

Basing on the distinct units $i, j \epsilon s (i \neq j)$, the following Hurvitz-Thompson type estimator for LR can be defined as

$$\hat{LR} = \left[\sum_{i,j\epsilon s} |y_i-y_j|/\pi_{ij}\right]/2N^2\,\bar{Y} \qquad \dots \quad (2.10)$$

where π_{ij} is the second order inclusion probabilities associated with a design $\mathcal{D} = (U, S, P)$. The estimator in (2.10) may be shown to be unbiased for LR and variance of the estimator can be found out without much difficulty. However, an estimator for LR together with the estimate of the variance of the estimator have been obtained as a direct application of Nanjamma-Murthy-Sethi procedure, just by defining the appropriate kernel function for LR.

3. A generalised estimation procedure under N-M-S set up

Let a sample of fixed size n be drawn from a population U, using a sampling design $\mathcal{D} = (U, S, P)$, where P is a probability measure defined on $s \in S$, such that

$$P(s) \geqslant 0 \text{ and } \sum_{s \in S} P(s) = 1.$$

Let π_i, π_{ij} and π_{ijk} be the inclusion probabilities of different order defined as

$$\pi_i = \sum_{s \ni i} P_s,\ \pi_{ij} = \sum_{\substack{s \ni i, j \\ (i \neq j)}} P_s \text{ and } \pi_{ijk} = \sum_{\substack{s \ni (i,j,k) \\ i \neq j \neq k}} P_s.$$

Let A be the class of sets α whose elements belong to U. In such a set, the same unit may or may not occur more than once. Let $\{\alpha\}$ be the set of all samples containing α. In a similar manner, one can define

$$\pi_{\{\alpha\}} = \sum_{s \in \{\alpha\}} P_s.$$

Let the population parameter F be expressible as

$$F = \sum_{\alpha \in A} f(\alpha) \qquad \ldots \quad (3.1)$$

where $f(\alpha)$ is a single valued function defined over the class A. Let S be a sample defined by an ordered sequence of units. It may be noted that the same unit may or may not occur more than once in such an s. For example, a sample of 3 units may be (y_1, y_2, y_1) which is different from (y_1, y_1, y_2) or (y_2, y_1, y_1).

An unbiased estimator of the parameter F may be given by

$$\hat{F} = \sum_{\alpha \in s} f(\alpha). P(s \mid \alpha)/P(s). \qquad \ldots \quad (3.2)$$

The variance and unbiased estimator of variance of $\hat{F}$ would be given by

$$V(\hat{F}) = \sum_{\alpha, \alpha' \subset s} f(\alpha) f(\alpha') \left\{ \frac{P(s \mid \alpha) P(s \mid \alpha')}{P(s)\ P(s \mid \alpha \oplus \alpha')} - 1 \right\} \qquad \ldots \quad (3.3)$$

and

$$\hat{V}(\hat{F}) = (\hat{F})^2 - \sum_{\alpha, \alpha' \subset s} f(\alpha) f(\alpha')\ P(s \mid \alpha \oplus \alpha')/Ps \qquad \ldots \quad (3.4)$$

where

$$P(s \mid \alpha) = P(s)/P(\{\alpha\}) \text{ and } P(s \mid \alpha \oplus \alpha') = P(s)/P(\{\alpha\}\ U \{\alpha'\})$$

are the conditional probabilities of s given the set $\{\alpha\}$ and $\{\alpha\} \bigcup \{\alpha'\}$.

$\hat{F}$ is unbiased only if each s contains at least one α and each set α is contained in at least one s. $\hat{V}(\hat{F})$ is unbiased if every set $(\alpha \oplus \alpha')$ is contained in atlest one s and every s contains atleast one set $(\alpha \oplus \alpha')$. Hence the minimum sample size required is $2r$, where α is defined using r observations. Now, since

$$P(\{\alpha\}) = \sum_{s \in \{\alpha\}} P(s) = \sum_{s \ni \alpha} P(s),$$

where α contains all distinct units, or

$$P(\{\alpha\}) = \pi_\alpha$$

we can in general write

$$P(s \mid \alpha) = P(s)/\pi_{\{\alpha\}}$$

$\pi_{\{\alpha\}} = \pi_\alpha$ in case α contains only distinct units. Hence $\pi_{\{\alpha\}}$ can be regarded as a general version of π_α, where α need not have all distinct units.

4. Estimators based on N-M-S

Case 1. SRSWOR : Under SRSWOR, $\pi_{ij} = n(n-1)/N(N-1)$ and thus

$$\hat{LR}_u = [N(N-1)/n(n-1)] \sum_{\substack{i<j \\ \alpha=(i,j)}} f(\alpha) \qquad \ldots \quad (4.1)$$

where $\qquad f(\alpha) = |y_i - y_j| N^2 \, \overline{Y}$ for $\alpha = (i, j)$.

One may observe that

$$P(s \mid \alpha) = \binom{N-2}{n-2}^{-1}, \quad P(s) = \binom{N}{n}^{-1}$$

and

$$P(s \mid \alpha \oplus \alpha') = \begin{cases} \binom{N-2}{n-2}^{-1} \text{if } \alpha = \alpha' \\ \binom{N-3}{n-3}^{-1} \text{if only one unit is common in } \alpha \text{ and } \alpha'. \\ \binom{N-4}{n-4} \ \text{ if there is no unit in common.} \end{cases}$$

and hence

$$V(\hat{LR}_u) = \Bigg[\sum_{i \neq j \neq k \neq l} |y_i - y_j| \, |y_k - y_l| \left\{ \frac{N(N-1)(n-2)(n-3)}{(N-2)(N-3)n(n-1)} - 1 \right\}$$

$$+ 4 \sum_{i \neq j \neq k} |y_i - y_j| \, |y_i - y_k| \left\{ \frac{N(N-1)(n-2)}{n(n-1)(N-2)} - 1 \right\}$$

$$+ 2 \sum_{i \neq j} |y_i - y_j|^2 \left\{ \frac{N(N-1)}{n(n-1)} - 1 \right\} \Bigg] / 4N^4 \overline{Y}^2 \qquad \ldots \quad (4.2)$$

with

$$\hat{V}(\hat{LR}_u) = a_{00}^2 \frac{N(N-1)}{n(n-1)} \left\{ \frac{N(N-1)}{n(n-1)} - \frac{(N-3)\,(N-2)}{(n-3)\,(n-2)} \right\}$$

$$+ 4\Sigma\, a_{i0}^2 \frac{N(N-1)\,(N-2)\,(N-n)}{n(n-1)\,(n-2)\,(n-3)}$$

$$-2\Sigma a_{ij}^2 \frac{N(N-1)\,(N-n)\,(N-n+1)}{n(n-1)\,(n-2)\,(n-3)} \qquad \ldots \quad (4.3)$$

where

$$a_{ij} = |y_i - y_j|\,,\ a_{i0} = \sum_j |y_i - y_j| \text{ and } a_{00} = \Sigma\,\Sigma\,|y_i - y_j|\,.$$

Case 2. *SRSWR* : Under simple random sampling with replacement, the proposed estimator for LR would be

$$\hat{LR}_{NMS} = \sum_{i<j} |y_i - y_j|\, P(s\,|\,\alpha)/P(s) N^2 \bar{Y} \qquad \ldots \quad (4.4)$$

However, the exact expression can be obtained by substituting the value of $P(s\,|\,\alpha)$.

It follows from the definition that

$$\pi_{\{\alpha\}\cup\{\alpha'\}} = \begin{cases} \pi_{iij} & \text{if } \alpha = (i,\, i) \text{ and } \alpha' = (i,\, j) \\ \pi_{iijk} & \text{if } \alpha = (i,\, i) \text{ and } \alpha' = (j,\, k) \\ \pi_{ii} & \text{if } \alpha = (i,\, i) \text{ and } \alpha' = (i,\, i) \\ \pi_{ijkl} & \text{if } \alpha = (i,\, j) \text{ and } \alpha' = (k,\, l) \\ \pi_{ijk} & \text{if } \alpha = (i,\, j) \text{ and } \alpha' = (i,\, k) \\ \pi_{iijj} & \text{if } \alpha = (i,\, i) \text{ and } \alpha' = (j,\, j) \\ \pi_{ij} & \text{if } \alpha = (i,\, j) \text{ and } \alpha' = (i,\, j) \end{cases} \qquad \ldots \quad (4.5)$$

Hence, with the help of the expressions in (4.5), we have

$$V(\hat{LR}_{N-M-S}) = \Big[\sum_{i \neq j \neq k \neq l} |y_i - y_j|\, |y_k - y_l| \left(\frac{\pi_{ijkl}}{\pi_{ij}\pi_{kl}} - 1 \right)$$

$$+ 4 \sum_{i \neq j \neq k} |y_i - y_j|\, |y_i - y_k| \left(\frac{\pi_{ijk}}{\pi_{ij}\pi_{ik}} - 1 \right)$$

$$+ 2 \sum_{i \neq j} |y_i - y_j|^2 \left(\frac{1}{\pi_{ij}} - 1 \right) \Big] / 4N^4 \bar{Y}^2 \qquad \ldots \quad (4.6)$$

and

$$\hat{V}(\hat{LR}_{N-M-S}) = \Big[\sum_{i \neq j \neq k \neq l} |y_i - y_j| \, |y_k - y_l| \left(\frac{1}{\pi_{ij}^2} - \frac{1}{\pi_{ijkl}} \right)$$

$$+4 \sum_{i \neq j \neq k} |y_i - y_j| \, |y_i - y_k| \left(\frac{1}{\pi_{ij}^2} - \frac{1}{\pi_{ijk}} \right)$$

$$+2 \sum_{i \neq j} |y_i - y_j|^2 \left(\frac{1}{\pi_{ij}^2} - \frac{1}{\pi_{ij}} \right) \Big] / 4N^4 \bar{Y}^2 \qquad \ldots \quad (4.7)$$

Case 3 : *N-M-S estimator based on distinct units on the basis of a sample drawn by SRSWR.* Since each α is defined on two distinct units, the samples of size n like $(y_1, y_1, \ldots, y_1), \ldots, (y_N, y_N, \ldots, y_N)$ can not be used to estimate LR. So assigning zero probabilities to all samples $s = \{y_i, y_i, \ldots, y_i\}$, the probabilities for other samples can be adjusted as

$$P^*(s) = \begin{cases} 1/(N^n - n) \text{ if all the units are not the same} \\ 0, \qquad \text{otherwise} \end{cases} \qquad \ldots \quad (4.8)$$

in order to have $\sum_{s \in S^*} P^*(s) = 1$, S^* being the class of all possible samples, where all the units are not the same. Thus using,

$$\pi_\alpha^* = P^*(s). N_\alpha,$$

where N_α is the number of samples containing α we have

$$P^*(s \mid \alpha) = P^*(s)/\pi_\alpha^*,$$

as before and

$$P^*(s) = P(s) \times N^n/(N^n - n)$$

$$\pi^*(\alpha) = \pi_\alpha \times N^n/(N^n - n)$$

$$P^*(s \mid \alpha) = P(s \mid \alpha)$$

Hence,

$$\hat{F}^* = [(N^n - n)/N^n] \sum_\alpha f(\alpha)/\pi_\alpha^* = F - [n/N^n] \sum_\alpha f(\alpha)/\pi_\alpha^* \qquad \ldots \quad (4.9)$$

and $$V(\hat{F}^*) = \sum_{\alpha, \alpha' \subseteq A} f(\alpha) f(\alpha') \left\{ \frac{P(s \mid \alpha) \; P(s \mid \alpha')}{P(s) \; P(s \mid \alpha \oplus \alpha')} \frac{N^n - n}{N^n} - 1 \right\}$$

$$= V(\hat{F}) - (n/N^n) \sum_{\alpha, \alpha' \subseteq A} f(\alpha) f(\alpha') P(s \mid \alpha) P(s \mid \alpha') / P(s) P(s \mid \alpha \oplus \alpha') \qquad \ldots \quad (4.10)$$

Since (n/N^n) is a negligible term, one can use $\hat{F}$ instead of $\hat{F}^*$ for all practical purpose.

5. Some concluding remarks

One can easily check that $V(L\hat{R}_B) < V(L\hat{R}_A)$ for reasonably large N and it is also expected that $V(L\hat{R}_u)$ (i.e., $V(L\hat{R}_B)) < V(L\hat{R}_{NMS}) < V(L\hat{R}_A)$. Although it was not possible to prove the inequality in a straight forward fashion because of the complicated expressions for π_{ij}, π_{ijk}, but an investigation to specific cases of π_{ij}, π_{ijk} reveal that

(a) $\pi_{ij} \leqslant \pi'_{ij}$

and (b) $\pi'_{ijk}/(\pi'_{ij})^2 \leqslant \pi_{ijk}/(\pi_{ij})^2$ for $f \leqslant 3/13$

π_{ij}, π'_{ij} are being defined under SRSWR and SRSWOR respectively ; similarly π_{ijk} and π'_{ijk} stand for the corresponding expressions under SRSWR and SRSWOR. However taking different values of n and N, it became possible to show that

(c) $\pi'_{ijkl}/(\pi'_{ij})^2 \leqslant \pi_{ijkl}/(\pi_{ij})^2$

holds in most of the situations of $n = (4, 5, 10, 15, 20, 50)$ and $f = (.001, .01, 0.1, 0.2, 0.3, 0.5)$. In the cases, where the inequality is reverse, the differences have been found to be very small, as has been reflected in the following table :

TABLE 1 : VALUES OF $\pi'_{ijkl}/(\pi'_{ij})^2$ AND $\pi_{ijkl}/(\pi_{ij})^2$ FOR DIFFERENT VALUES OF f AND n.

n	$f = 0.001$	
	$\pi'_{ijkl}/(\pi'_{ij})^2$	$\pi_{ijkl}/(\pi_{ij})^2$
5	0.3001801	0.2998379
10	0.6224089	0.6221370
15	0.9602308	0.9590217

So mostly, we can say that without replacement scheme is better than that of with replacement. Also the estimate based on distinct units or pairs of units is better than the corresponding estimate on all units or all pairs of units so long as $V(y)$ is not very small.

N-M-S approach was adopted only to obtain an estimate for the variance of the estimator without much difficulty. Otherwise, it is not very much different from the usual set up.

Comparison between SRSWR and SRSWOR would be more logical, if comparison is made between a SRSWR of size l whose expected effective size would be the size of SRSWOR i.e. if k is the number of distinct units in a smaple drawn by SRSWR, and if $E(k) = n$; then a sample of size n should be drawn by SRSWOR to compare it with SRSWR. However, in that case, we did not arrive at any clear picture. Further studies should be made (say through super population models) in order to establish the superiority of the estimator based on SRSWOR to those under SRSWR.

Acknowledgements. The authors are thankful to the referee for his valuable comments which helped in improving the paper to the present form.

References

Nanjamma, N. S., Murthy M. N. and Sethi V. K. (1959) : Some sampling systems providing unbiased ratio estimators. *Sankhyā*, **21**, 299-314.

Taguchi, Tokio (1978) : On an unbiased, consistent and asymptotically efficient estimation of Gini's concentration coefficient. *Metron*, *XXXVI*, No. 3-4, 57-72.