Chapter 7

Midzuno-Sen Strategy

in

Two-phase Sampling
This chapter deals with the ratio estimation of the finite population total under Midzuno-Sen sampling scheme when complete auxiliary information is not available but another auxiliary variable, closely related to first auxiliary variable but remotely related to the study variable, is available. A chain ratio estimator in two-phase sampling is suggested. An empirical study is carried out to investigate the performance of the suggested estimator relative to the convention two-phase estimator.

7.1 Introduction

Consider a finite population \( U = \{1, ..., i, ..., N\} \). Let \( y \) and \( x \) be the study variable and auxiliary variable taking the values \( y_i \) and \( x_i \) respectively for the population unit \( i \). When the two variables are strongly related but no information is available on the population total \( X = \sum_{U} x_i \) (or mean \( \overline{X} \)), we wish to estimate the population total \( Y = \sum_{U} y_i \) of \( y \) from a sample \( s \), obtained through a two-phase selection.

Scheme 1: The first-phase sample \( s' \) of fixed size \( n' \) is drawn using simple random sampling without replacement (SRSWOR) to observe only \( x \) in order to estimate \( X \). Given \( s' \), the second-phase sample \( s \) of fixed size \( n \) is drawn again using SRSWOR to observe \( y \) only.

The two-phase sampling ratio estimator is defined as

\[
\hat{Y}_{rd} = N \overline{x'} \overline{y}/\overline{x} = N x_s y_s/n' x_s
\]

(7.1.1)

where \( \overline{x'}(x_s') \) is the mean (total) of \( x \) for the sample \( s' \), \( \overline{y}(y_s) \), \( \overline{x}(x_s) \) are means (totals) of \( y \) and \( x \) for the sample \( s \).
Raj (1965) suggests the following two-phase sample selection procedure.

**Scheme II:** The initial sample \( s' \) of size \( n' \) is selected with probability proportional to \( z \) with replacement and the second phase sample \( s \) of size \( n \) is a subsample of \( s' \), selected with equal probabilities without replacement.

An unbiased estimator of \( Y \) is provided by

\[
\hat{\theta}_{DRA} = \sum_{s} \left( \frac{y_i}{n p_i} \right) + k \left[ \sum_{s'} \left( \frac{x_i}{n' p_i} \right) - \sum_{s} \left( \frac{x_i}{n p_i} \right) \right] \tag{7.1.2}
\]

where \( k \) is a suitably chosen scalar. To compute this estimator it is necessary to assess the value of \( k \). Often this is difficult. Motivated by this Srivenkataramana and Tracy (1989) have modified the sampling scheme.

**Scheme III:** Select the first phase sample \( s' \) of size \( n' \) with probability proportional to \( z \) and select a subsample \( s \) of size \( n \) units with probability proportional to \( x/z \), with replacement. Under this scheme an unbiased estimator of \( Y \) is

\[
\hat{\theta}_{STD} = z \sum_{s} \left( \frac{y_i}{n x_i} \right) \sum_{s'} \left( \frac{x_i}{n' z_i} \right) \tag{7.1.3}
\]

Experience shows that of all the unequal probability sampling without replacement schemes available in literature those due to Midzuno (1952) and Sen (1953), and Rao et al. (1962) have wider practical applicability since these procedures are quite simple and involve less computations at estimation stage compared to others. In this chapter to estimate the population total we will use the following two-phase sampling procedure.
**Scheme IV:** A first-phase sample $s'$, of size $n'$, is drawn from $U$ according to a design SRSWOR. Given $s'$, a second-phase sample $s$, of size $n$, is drawn from $s'$ according to the Midzuno-Sen (MS) sampling. The MS scheme of sampling describes the selection of the first unit with probability proportional to a given size measure ($x$) and remaining $(n-1)$ units of the sample by SRSWOR. The size of the measure is an auxiliary variable and is positively correlated with the study variable.

Under this scheme the ordinary ratio estimator is known to be unbiased for the population total. The variance of this estimator is not available in literature in a meaningful form. Rao and Vijayan (1977) investigated the problem of estimating the variance of the ratio estimator, given at (7.1.1), under the MS sampling. The same problem addressed by Patel and Patel (2008, 2009, 2010).

In Section 2 we suggest a chain ratio estimator using two auxiliary variables. To study the performance of the suggested estimator a small scale simulation is presented. Also, comparison of strategy is given in this section. Next, an adjusted Horvitz-Thompson type estimator is suggested in Section 3. Conclusion is given in Section 4.

### 7.2 The Chain-ratio Estimator under Scheme IV

#### 7.2.1 Bias and Variance

In this section a chain ratio estimator

$$\hat{Y}_{rd} = N(\bar{z}/\bar{x})$$

is suggested.
To obtain bias $B(\cdot)$ and variance $Var(\cdot)$ of the chain-ratio estimator under a given design we use the same notations that are given Chapter 6.

**Theorem 7.1** Under Scheme IV, approximate bias and variance, to $O(n^{-1})$, of $\hat{Y}_{Rd}^*$ are given by

$$B(\hat{Y}_{Rd}^*) = \frac{1 - f'}{n} Y (C_y^2 - \rho_{yz} C_y C_z)$$  \hspace{1cm} (7.2.2)$$

$$Var(\hat{Y}_{Rd}^*) = \frac{1 - f'}{n} Y^2 (C_y^2 + C_z^2 - 2\rho_{yz} C_y C_z) + V \left[ 1 + 3 \frac{1-f'}{n} C_y^2 - \frac{2(\Delta - \bar{z}V)}{\bar{z}V} \right]$$  \hspace{1cm} (7.2.3)$$

where

$$\Delta = \frac{1}{N} \left[ \sum_{i \in U} a_{ii} y_i^2 z_i + a \left( \sum_{i \neq j \in U} a_{ij} y_i^2 z_j + 2 \sum_{i \neq j \in U} a_{ij} y_i y_j z_j \right) + b \sum_{i \neq j \neq k \in U} a_{ijk} y_i y_j z_k \right]$$

with $a_{ij}$ and $V$ are defined below and

$$a = \frac{(n' - 1)}{(N - 1)}, \quad b = \frac{(n' - 1)(n' - 2)}{(N - 1)(N - 2)}, \quad f' = n'/N$$

**Proof.** The expectation of $\hat{Y}_{Rd}^*$ is evaluated using the conditional argument $E(\cdot) = E_1 E_2 (\cdot | s')$ as

$$E(\hat{Y}_{Rd}^*) = NE_1 \left[ (\bar{z}/\bar{z}') E_2 (\bar{x}' \bar{y} / \bar{x}) \right] = NE_1 \left[ (\bar{z}/\bar{z}') \bar{y} \right]$$
and consequently using the formula for approximate bias of the ordinary ratio estimator (see, Cochran, 1977) of the population total under SRSWOR we obtain the bias of \( \tilde{Y}_{Rd} \) as given in (7.2.2). Next, the formula for variance of \( \tilde{Y}_{Rd}^* \) is

\[
V \text{ar}(\tilde{Y}_{Rd}^*) = V_1 E_2(\tilde{Y}_{Rd}^*) + E_1 V_2(\tilde{Y}_{Rd}^*) \tag{7.2.4}
\]

The first component is readily seen to be

\[
V_1 E_2(\tilde{Y}_{Rd}^*) = V_1 \left( \frac{N}{n - 1} \frac{\bar{y}}{\tilde{x}} \right) \approx \frac{1}{n} \frac{f'}{y^2} \left( c_y^2 + c_x^2 - 2 \rho_{yx} c_y c_x \right) \tag{7.2.5}
\]

Note that at second phase the ratio estimator \( \tilde{x}/\tilde{y} \) is unbiased for \( \tilde{y}' \) under MS sampling design with variance (see, Rao, 1972)

\[
\nu' = \sum_{s'} a_{ii} y_i^2 + \sum \sum_{s'} a_{ij} y_i y_j
\]

where

\[
a_{ii} = \frac{X}{(N - 1)} \sum_{S \ni i} \frac{1}{x_{si}} - 1 \quad \text{and} \quad a_{ij} = \frac{X}{(N - 1)} \sum_{S \ni i \& j} \frac{1}{x_{sj}} - 1
\]

Also, note that \( \nu' \) is unbiased for

\[
V = \frac{n'}{N} \sum_{U} a_{ii} y_i^2 + \frac{n'(n' - 1)}{N(N - 1)} \sum a_{ij} y_i y_j
\]

at first phase (i.e., under SRSWOR). The second component is written as

\[
E_1 V_2(\tilde{Y}_{Rd}^*) = E_1 \left[ \left( \bar{Z}/\bar{x} \right)^2 V_2(\bar{x}/\bar{x}) \right] = E_1 \left[ \left( \frac{\bar{Z}}{\bar{x}} \right)^2 \cdot \nu' \right]
\]

Inserting \( \nu' = V(1 + \delta_4) \), \( |\delta_4| < 1 \), and using the expansion

\[
(1 + \delta_3)^{-2} = 1 - 2\delta_3 + 3\delta_3^2 + O(n^{-1})
\]

102
after omitting the terms of $\delta$'s having power greater than two, we obtain

$$E_1V_2(\bar{Y}_{\text{rd}}) \equiv V[1 + 3Var(\delta_3) - 2Cov(\delta_3, \delta_4)] \quad (7.2.6)$$

Noting that

$$Var(\delta_3) = \frac{1 - f'}{n'} C_z^2$$

$$Cov(\delta_3, \delta_4) = \frac{E(\bar{Z}'v') - \bar{Z}V}{\bar{Z}V} = \frac{\Delta - \bar{Z}V}{\bar{Z}V}$$

(7.2.6) is written as

$$E_1V_2(\bar{Y}_{\text{rd}}^*) \equiv V \left[ 1 + 3 \frac{1 - f'}{n'} C_z^2 - 2 \frac{(\Delta - \bar{Z}V)}{\bar{Z}V} \right] \quad (7.2.7)$$

Inserting (7.2.5) and (7.2.7) in (7.2.4) we obtain (7.2.2).

### 7.2.2 An Empirical Comparison

The preceding estimators $\bar{Y}_{\text{rd}}$ and $\bar{Y}_{\text{rd}}^*$ were compared empirically on 3 natural populations given below. For comparison of the estimators a two-phase sample was drawn using SRSWOR and MS sampling scheme from each of the populations and these estimators were computed. This procedure was repeated $M = 5000$ times. The relative percentage bias and the relative efficiency were calculated and presented in Table 7.1

Data set I: Murthy (1967)

$y$: No of cultivators  $x$: Area in square miles  $z$: No. of households

Data set II: Murthy (1967)

$y$: workers at household industry  $x$: Cultivated area (in acres)

$z$: No. of households
Data set III: Fisher data (combined all three data sets)

\[ y: \text{Petal width} \quad x: \text{Sepal length} \quad z: \text{Petal length} \]

Table 7.1 Relative bias and MSE

<table>
<thead>
<tr>
<th>Data Set</th>
<th>( n )</th>
<th>( n' )</th>
<th>N</th>
<th>Estimator</th>
<th>RB%</th>
<th>( MSE_{sim} )</th>
<th>RE%</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>10</td>
<td>30</td>
<td>128</td>
<td>( \hat{y}_{rd} )</td>
<td>0.60</td>
<td>3.79E+08</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>( \hat{y}^*_{rd} )</td>
<td>1.02</td>
<td>2.95E+08</td>
<td>128</td>
</tr>
<tr>
<td>II</td>
<td>10</td>
<td>30</td>
<td>128</td>
<td>( \hat{y}_{rd} )</td>
<td>0.89</td>
<td>1.46E+07</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>( \hat{y}^*_{rd} )</td>
<td>0.71</td>
<td>1.28E+07</td>
<td>114</td>
</tr>
<tr>
<td>III</td>
<td>15</td>
<td>40</td>
<td>150</td>
<td>( \hat{y}_{rd} )</td>
<td>-3.16</td>
<td>1.01E+03</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>( \hat{y}^*_{rd} )</td>
<td>-3.32</td>
<td>7.07E+02</td>
<td>143</td>
</tr>
</tbody>
</table>

The above simulation reveals that (1) the absolute RBs % of both the estimators are in reasonable range and (2) the efficiency of the suggested estimator is substantial as compared to the conventional estimator.

7.3 Comparison of Strategies

Here, we compare empirically the suggested strategy (mean a pair of design and estimator) given at (7.2.1) with the conventional strategy given at (1), denoted respectively by \( H_2(\text{srswor} - \text{MS sampling}), \hat{y}^*_{rd} \) and \( H_1(\text{srswor} - \text{srswor}, \hat{y}_{rd}) \). For the comparison of two strategies \( H_1(p_1, \hat{y}_1) \) and \( H_2(p_2, \hat{y}_2) \) we have computed the percentage gain in efficiency of \( H_2 \) over \( H_1 \) as

\[
Gain(H_2, H_1) = \left( \frac{V_{p_1}(\hat{y}_1)}{V_{p_2}(\hat{y}_2)} - 1 \right) \times 100\%
\]

The following table presents gain due to MS sampling scheme over SRSWOR in two-phase sampling.
<table>
<thead>
<tr>
<th>Data Set</th>
<th>I</th>
<th>II</th>
<th>III</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\text{Gain}(H_2, H_1))</td>
<td>33%</td>
<td>-4%</td>
<td>26%</td>
</tr>
</tbody>
</table>

### 7.4 Conclusion

If extra auxiliary information is available at estimation stage then in this case using varying probability sampling scheme at first phase to estimate the unknown mean of auxiliary variable that is correlated with the study variable one can improve the conventional estimator. For varying probability sampling scheme pps without replacement schemes for general \(n\) are not easy to implement but MS scheme is easy to execute.

### 7.5 MS Sampling-PPS Sampling Strategy

In Theorem 7.1 it has been shown that the chain-ratio estimator is biased. In this section we modify the sampling strategy as follows:

**Scheme V:** A first-phase sample \(s'\), of size \(n'\), is drawn from \(U\) according to a MS sampling with \(z\) as size measure, positively correlated with \(y\) variable. Given \(s'\), a second-phase sample \(s\), of size \(n\), is drawn from \(s'\) according to the MS sampling with \(x\) as the size measure which is also positive correlated with \(y\).

Under the Scheme V it is easy to verify that the chain-ratio estimator is unbiased. Its variance can be obtained on the same way that given in Theorem 7.1.

Following Prasad and Venkataramana we suggest another estimator under Scheme V.

Under Scheme V the first and second order inclusion probabilities are
\[ \pi_i = \frac{(N - n')z_i}{(N - 1)Z} + \frac{n' - 1}{N - 1} \]

\[ \pi_{ij} = \frac{(N - n')(n' - 1)(z_i + z_j)}{(N - 1)(N - 2)Z} + \frac{(n' - 1)(n' - 2)}{(N - 1)(N - 2)} \]

Consider the transformation

\[ w_i = x_i + \frac{(n' - 1)b}{N - n'} \quad (i = 1, \ldots, N) \]

where \( b \) is a scalar to be chosen. Then \( X \) can unbiasedly be estimated by

\[ \hat{X} = \sum_{s} w_i \frac{N(n' - 1)b}{N - n'} \]

Consequently the suggested (adjusted Horvitz-Thompson type) estimator under Scheme V would be

\[ \hat{Y}_{adjHT} = \hat{X} \sum_{s} \frac{Y_i}{nX_i} \]

**Comparison of Strategies**

Here we compare the strategies suggested by Tracy and Venkataramana, the chain-ratio estimator with Scheme V and the adjusted HT estimator with Scheme V with convention strategy.

The estimators were compared empirically on natural population given in Appendix A, Table A.2.3 (popl. No.1-8).

The simulated results are presented in the Table 7.2
Table 7.2 The Relative Biased and Simulated Variance

<table>
<thead>
<tr>
<th>Poppl</th>
<th>N</th>
<th>n'</th>
<th>n</th>
<th>$\bar{Y}_{Rd}$ (Scheme I)</th>
<th>$\bar{Y}_{STd}$</th>
<th>$\bar{Y}_{Rd}^*$ (Scheme V)</th>
<th>$\bar{Y}_{adjHT}$ (Scheme V)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>284</td>
<td>70</td>
<td>25</td>
<td>-1.13</td>
<td>0.00</td>
<td>-0.03</td>
<td>-5.35</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>7.98E+07</td>
<td>6.96E+06</td>
<td>1.39E+07</td>
<td>1.65E+07</td>
</tr>
<tr>
<td>2</td>
<td>284</td>
<td>70</td>
<td>25</td>
<td>-0.38</td>
<td>0.73</td>
<td>-0.33</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>8.43E+08</td>
<td>6.25E+08</td>
<td>6.26E+08</td>
<td>7.09E+08</td>
</tr>
<tr>
<td>3</td>
<td>150</td>
<td>50</td>
<td>15</td>
<td>1.97</td>
<td>-0.30</td>
<td>-0.69</td>
<td>-0.66</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>246.334</td>
<td>315.073</td>
<td>251.46</td>
<td>193.04</td>
</tr>
<tr>
<td>4</td>
<td>128</td>
<td>40</td>
<td>10</td>
<td>-3.24</td>
<td>0.24</td>
<td>0.54</td>
<td>0.62</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2.71E+08</td>
<td>3.2E+08</td>
<td>2.38E+08</td>
<td>2.5E+08</td>
</tr>
<tr>
<td>5</td>
<td>128</td>
<td>40</td>
<td>10</td>
<td>2.92</td>
<td>0.11</td>
<td>-0.53</td>
<td>0.43</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>6.49E+09</td>
<td>1.63E+10</td>
<td>1.60E+10</td>
<td>1.58E+10</td>
</tr>
<tr>
<td>6</td>
<td>128</td>
<td>40</td>
<td>10</td>
<td>-0.31</td>
<td>-0.05</td>
<td>-0.04</td>
<td>-0.09</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.31E+09</td>
<td>2.3E+08</td>
<td>1.81E+08</td>
<td>1.24E+08</td>
</tr>
<tr>
<td>7</td>
<td>128</td>
<td>40</td>
<td>10</td>
<td>-0.96</td>
<td>0.09</td>
<td>0.69</td>
<td>-0.01</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3.7E+08</td>
<td>6.28E+08</td>
<td>5.74E+08</td>
<td>3.77E+08</td>
</tr>
<tr>
<td>8</td>
<td>80</td>
<td>20</td>
<td>7</td>
<td>-4.34</td>
<td>-0.08</td>
<td>-2.13</td>
<td>-0.20</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.96E+08</td>
<td>1.14E+08</td>
<td>1.07E+08</td>
<td>8.36E+07</td>
</tr>
</tbody>
</table>

Except Populations 1 and 2 our suggested both the strategies have performed very well as compared to Tracy and Venkataramana strategy.