Journal of the Korean Statistical Society (2000), **29**: 4, pp 455–471

Estimation of Mean Using Balanced Systematic Sampling and Interpolation for Population with Linear Trend[†]

Hyuk Joo Kim[1]

ABSTRACT

A new method is developed for estimating the mean of a population which has a linear trend. The proposed estimator is based on the balanced systematic sampling method and the concept of interpolation. The efficiency of the proposed method is compared with that of existing methods.

Keywords: Population with a linear trend ; Balanced systematic sampling ; Interpolation ; Infinite superpopulation model.

1. Introduction

When conducting statistical surveys, we sometimes meet with a population which has a linear trend. For example, suppose we wish to estimate the average sales of the supermarkets in a certain city. If the supermarkets in that city are arranged in increasing or decreasing order of the number of employees, we can expect that there will be a linear trend in this population.

In estimating the mean of a population which has a linear trend, ordinary systematic sampling (OSS) is known to be much better than simple random sampling (SRS). Several researchers have suggested sampling methods which are versions of systematic sampling. Among them, end corrections (EC) proposed by Yates (1948), centered systematic sampling (CSS) proposed by Madow (1953), balanced systematic sampling (BSS) proposed by Sethi (1965) and Murthy (1967), and modified systematic sampling (MSS) proposed by Singh et al. (1968) are well-known methods.

Recently, Kim (1998, 1999) proposed two methods for the case when n (the sample size) is an odd number ($n \geq 3$) and k (the reciprocal of the sampling fraction) is an even number. These methods, one using MSS and interpolation

[†]This paper was supported by Wonkwang University in 2000.

[1]Division of Mathematical Science, Wonkwang University, Iksan, Chonbuk, 570-749, Korea.

and the other using BSS and interpolation and extrapolation, turned out to be relatively efficient as compared with conventional methods in many cases.

In this paper, we propose another method for efficiently estimating the mean of a population which has a linear trend. The linear trend will be specified by using a mathematical expression in Section 3. The proposed method, based on BSS and the concept of interpolation, will be developed for use in the case when n is an odd number ($n \geq 5$) and k is an even number, and will be compared with several existing methods under the expected mean square error criterion based on the infinite superpopulation model introduced by Cochran (1946).

2. Proposition of the method

Suppose we have a population of size $N = kn$, the units of which are denoted by $U_1, U_2, \cdots, U_N$. We wish to select a sample of size n from this population.

First, let us briefly review the balanced systematic sampling (BSS) method. This sampling method, proposed by Sethi (1965) and Murthy (1967) as stated in the previous section, was developed for populations having linear trends.

BSS selects one of the k clusters $C_1', C_2', \cdots, C_k'$ with respective probability $1/k$, and then estimates the population mean by the sample mean, $\overline{y}_{bal}$, which is the mean of the selected cluster. Here the cluster C_i' is defined by

$$C_i' = \{U_{i+2(j-1)k} : j = 1, 2, \cdots, n/2\} \cup \{U_{2jk+1-i} : j = 1, 2, \cdots, n/2\} \quad (i = 1, 2, \cdots, k)$$

for n even, and

$$C_i' = \{U_{i+2(j-1)k} : j = 1, 2, \cdots, (n+1)/2\} \cup \{U_{2jk+1-i} : j = 1, 2, \cdots, (n-1)/2\} \quad (i = 1, 2, \cdots, k)$$

for n odd. For example, if $N = 28$, $n = 7$ and $k = 4$, then the four clusters are as follows :

$$\begin{aligned} C_1' &= \{U_1, U_8, U_9, U_{16}, U_{17}, U_{24}, U_{25}\} \\ C_2' &= \{U_2, U_7, U_{10}, U_{15}, U_{18}, U_{23}, U_{26}\} \\ C_3' &= \{U_3, U_6, U_{11}, U_{14}, U_{19}, U_{22}, U_{27}\} \\ C_4' &= \{U_4, U_5, U_{12}, U_{13}, U_{20}, U_{21}, U_{28}\}. \end{aligned}$$

The sample mean $\overline{y}_{bal}$ obtained by BSS is easily seen to be an unbiased estimator of $\overline{Y}$, the population mean, with variance

$$V(\overline{y}_{bal}) = \frac{1}{k}\sum_{i=1}^{k}(\overline{y}_i' - \overline{Y})^2,$$

where $\overline{y}_i'$ is the mean value for the units in C_i' $(i = 1, 2, \cdots, k)$.

Throughout this paper the following notation will be used :

y_i : value for the ith unit in the population $(i = 1, 2, \cdots, N)$,

$\overline{Y} = \frac{1}{N}\sum_{i=1}^{N} y_i$: population mean to be estimated,

y_{ij}' : value for the jth unit in C_i' $(i = 1, 2, \cdots, k;\ j = 1, 2, \cdots, n)$, that is,

$$\begin{aligned} y_{ij}' &= y_{i+(j-1)k} \ (j = 1, 3, 5, \cdots, n-1) \\ y_{ij}' &= y_{1-i+jk} \ (j = 2, 4, 6, \cdots, n) \end{aligned}$$

for n even, and

$$\begin{aligned} y_{ij}' &= y_{i+(j-1)k} \ (j = 1, 3, 5, \cdots, n) \\ y_{ij}' &= y_{1-i+jk} \ (j = 2, 4, 6, \cdots, n-1) \end{aligned}$$

for n odd,

$\overline{y}_i' = \frac{1}{n}\sum_{j=1}^{n} y_{ij}'$: mean for the units in C_i' $(i = 1, 2, \cdots, k)$.

Now we are in a position to introduce a new method for estimating the population mean $\overline{Y}$. This method involves the same sampling method as BSS, but it estimates $\overline{Y}$ by an adjusted estimator, not by the sample mean itself. We only consider the case when n is an odd number $(n \geq 5)$ and k is an even number, because the method is defined and has a practical meaning in this case.

Consider again the case of $N = 28$, $n = 7$ and $k = 4$. One of C_1', C_2', C_3', C_4' is selected with respective probability 1/4. We notice that the sums of the numbers assigned to the units in C_1', C_2', C_3' and C_4' are, respectively, 100, 101, 102 and 103, showing differences ranging from 1 to 3. When the population has a linear trend, it would be desirable to remove such differences. Our idea is to replace y_9, y_{10}, y_{11}, or y_{12} by "$y_{10.5}$" (or, alternatively, to replace y_{17}, y_{18}, y_{19}, or y_{20} by "$y_{18.5}$") according as C_1', C_2', C_3', or C_4' is selected. Here $y_{10.5}$ and $y_{18.5}$ are imaginary values which do not actually exist.

If C_1' is selected, then we can "estimate" $y_{10.5}$ by use of y_9 and y_{16}. By the interpolation method, $y_{10.5}$ is estimated by $(1/14)(11y_9 + 3y_{16})$. Therefore, by using this value in place of y_9, we can estimate $\overline{Y}$ by

$$\begin{aligned}\overline{y}_1'^*(3) &= \frac{1}{7}\{y_1 + y_8 + \frac{1}{14}(11y_9 + 3y_{16}) + y_{16} + y_{17} + y_{24} + y_{25}\} \\ &= \overline{y}_1' + \frac{3}{98}(y_{16} - y_9) \\ &= \overline{y}_1' + \frac{3}{98}(y_{14}' - y_{13}').\end{aligned}$$

Here the number '3' in the parentheses means that this estimate, $\overline{y}_1'^*(3)$, is obtained by using an adjusted value instead of the third value in C_1'. Instead of estimating $y_{10.5}$, we may estimate $y_{18.5}$ by using y_{17} and y_{24}, and use the resultant value in place of y_{17}.

Suppose now that the selected cluster is C_3'. Then by applying the interpolation method to y_6 and y_{11} to estimate $y_{10.5}$ and using the resultant value in place of y_{11}, we can estimate $\overline{Y}$ by

$$\begin{aligned}\overline{y}_3'^*(3) &= \frac{1}{7}\{y_3 + y_6 + \frac{1}{10}(y_6 + 9y_{11}) + y_{14} + y_{19} + y_{22} + y_{27}\} \\ &= \overline{y}_3' - \frac{1}{70}(y_{11} - y_6) \\ &= \overline{y}_3' - \frac{1}{70}(y_{33}' - y_{32}').\end{aligned}$$

Alternatively, we may estimate $y_{18.5}$ by using y_{14} and y_{19}, and use the resultant value in place of y_{19}.

Similar arguments enable us to estimate $\overline{Y}$ when C_2' or C_4' is selected.

The above method can be generalized as follows. One of the k clusters $C_1', C_2', \cdots, C_k'$ is selected with respective probability $1/k$. If the selected cluster is C_i', then the population mean $\overline{Y}$ is estimated by one of $\overline{y}_i'^*(3), \overline{y}_i'^*(5), \cdots, \overline{y}_i'^*(n-2)$(with respective probability $2/(n-3)$), where

$$\overline{y}_i'^*(m) = \overline{y}_i' + \frac{k+1-2i}{2n(2k+1-2i)}(y_{i,m+1}' - y_{im}')(m = 3, 5, \cdots, n-2)$$

for $i = 1, 2, \cdots, k/2$, and

$$\overline{y}_i'^*(m) = \overline{y}_i' - \frac{2i-k-1}{2n(2i-1)}(y_{im}' - y_{i,m-1}')(m = 3, 5, \cdots, n-2)$$

for $i = k/2+1, k/2+2, \cdots, k$.

Let us denote this method and the resultant estimator of $\overline{Y}$ as BI and $\overline{y}_{bi}$, respectively. Then $\overline{y}_{bi}$ is biased for $\overline{Y}$ and it is easy to show that $\overline{y}_{bi}$ has bias

$$B(\overline{y}_{bi}) = \frac{2}{k(n-3)} \sum_{i=1}^{k} \sum_{m} \overline{y}_i^{'*}(m) - \overline{Y}$$

and mean square error

$$MSE(\overline{y}_{bi}) = \frac{2}{k(n-3)} \sum_{i=1}^{k} \sum_{m} \{\overline{y}_i^{'*}(m) - \overline{Y}\}^2.$$

Here and hereafter $\sum_m$ means summing over $m = 3, 5, \cdots, n-2$.

3. Expected mean square error of $\overline{y}_{bi}$

In this section, the expected mean square error of $\overline{y}_{bi}$ is obtained by using Cochran's (1946) infinite superpopulation model.

We regard the finite population as a sample from an infinite superpopulation. First, as a general case, we set up the model as

$$y_i = \mu_i + e_i \ (i = 1, 2, \cdots, N), \tag{3.1}$$

where μ_i is a function of i and the random error e has properties $\mathcal{E}(e_i) = 0$, $\mathcal{E}(e_i^2) = \sigma^2$, $\mathcal{E}(e_i e_j) = 0$ $(i \neq j)$. The operator $\mathcal{E}$ denotes the expectation over the infinite superpopulation.

From now on, with regard to μ and e also we will use the same style of notation as adopted for y. That is,

$$\begin{aligned}
\overline{\mu} &= \frac{1}{N} \sum_{i=1}^{N} \mu_i, \\
\mu_{ij}^{'} &= \mu_{i+(j-1)k} \ (j = 1, 3, 5, \cdots, n), \\
\overline{\mu}_i^{'} &= \frac{1}{n} \sum_{j=1}^{n} \mu_{ij}^{'}, \\
\overline{\mu}_i^{'*}(m) &= \overline{\mu}_i^{'} + \frac{k+1-2i}{2n(2k+1-2i)} (\mu_{i,m+1}^{'} - \mu_{im}^{'}) (i = 1, 2, \cdots, k/2; m = 3, 5, \cdots, n-2),
\end{aligned}$$

and so on.

The following theorem is very important in evaluating the efficiency of $\overline{y}_{bi}$. The proof of this theorem is given in Appendix.

Theorem 1. *Assuming the model* (3.1), *the expected mean square error of* $\overline{y}_{bi}$ *for* k *even and* n *odd* ($n \geq 5$) *is*

$$\begin{aligned} \mathcal{E}MSE(\overline{y}_{bi}) &= \frac{2}{k(n-3)} \sum_{i=1}^{k} \sum_{m} \{\overline{\mu}_i'^{*}(m) - \overline{\mu}\}^2 + \frac{\sigma^2}{n} \frac{N-n}{N} \\ &\quad + \frac{\sigma^2}{2n^2}(1 - 4A_k + 2kB_k) \end{aligned} \tag{3.2}$$

where

$$\begin{aligned} A_k &= \sum_{i=1}^{k/2} \frac{1}{2k+1-2i} = \frac{1}{2}\left\{\psi\left(k+\frac{1}{2}\right) - \psi\left(\frac{k+1}{2}\right)\right\} \\ B_k &= \sum_{i=1}^{k/2} \frac{1}{(2k+1-2i)^2} = -\frac{1}{4}\left\{\psi^{(1)}\left(k+\frac{1}{2}\right) - \psi^{(1)}\left(\frac{k+1}{2}\right)\right\} \\ \psi(x) &= \frac{d}{dx}\ln\Gamma(x)\ (x>0) : \textit{the polygamma function} \\ \Gamma(x) &= \int_0^{\infty} t^{x-1}e^{-t}dt\ (x>0) : \textit{the gamma function} \\ \psi^{(1)}(x) &= \frac{d}{dx}\psi(x) \end{aligned}$$

Now, let us consider the case of $\mu_i = a + bi$, where a and b are constants with $b \neq 0$. In other words, the assumed model is

$$y_i = a + bi + e_i\ (i = 1, 2, \cdots, N). \tag{3.3}$$

This is the case of a population which has a linear trend.

In this case, as a preparatory stage for obtaining $\mathcal{E}MSE(\overline{y}_{bi})$ we get the following formulas :

$$\overline{\mu} = a + \left(\frac{b}{2}\right)(N+1) \tag{3.4}$$

$$\overline{\mu}_i' = a + \left(\frac{b}{2}\right)(N+1) + \left(\frac{b}{n}\right)\left(i - \frac{k+1}{2}\right) \tag{3.5}$$

$$\mu_{i,m+1}' = \mu_{1-i+(m+1)k} = a + b\{1 - i + (m+1)k\} \tag{3.6}$$

$$\mu'_{im} = \mu_{i+(m-1)k} = a + b\{i + (m-1)k\} \tag{3.7}$$

$$\mu'_{i,m-1} = \mu_{1-i+(m-1)k} = a + b\{1 - i + (m-1)k\} \tag{3.8}$$

Thus we have, for $i = 1, 2, \cdots, k/2$,

$$\begin{aligned}\overline{\mu}'^{*}_{i}(m) &= \overline{\mu}'_i + \frac{k+1-2i}{2n(2k+1-2i)}(\mu'_{i,m+1} - \mu'_{im}) \\ &= a + \left(\frac{b}{2}\right)(N+1),\end{aligned} \tag{3.9}$$

and for $i = k/2+1, k/2+2, \cdots, k$,

$$\begin{aligned}\overline{\mu}'^{*}_{i}(m) &= \overline{\mu}'_i - \frac{2i-k-1}{2n(2i-1)}(\mu'_{im} - \mu'_{i,m-1}) \\ &= a + \left(\frac{b}{2}\right)(N+1).\end{aligned} \tag{3.10}$$

Now using (3.4), (3.9), (3.10) and the result of Theorem 1, we obtain the following theorem :

Theorem 2. *For a population characterized by* (3.3), *the expected mean square error of* $\overline{y}_{bi}$ *is*

$$\mathcal{E}MSE(\overline{y}_{bi}) = \frac{\sigma^2}{n}\frac{N-n}{N} + \frac{\sigma^2}{2n^2}(1 - 4A_k + 2kB_k)(k : even,\ n : odd,\ n \geq 5), \tag{3.11}$$

where A_k *and* B_k *are as defined in Theorem 1.*

4. Comparison of efficiency with other methods

In this section, the efficiency of $\overline{y}_{bi}$ is compared with that of estimators resulting from other methods. First, let us consider SRS, OSS, MSS, BSS and CSS. Discussions on comparisons of the performances of OSS, MSS, BSS and CSS were also given in Bellhouse and Rao (1975).

For a population characterized by the model (3.3), the following were obtained in Singh et al. (1968) and Kim (1985):

$$\mathcal{E}MSE(\overline{y}_{ran}) = \left(\frac{b^2}{12}\right)(N+1)(k-1) + \frac{\sigma^2}{n}\frac{N-n}{N} \tag{4.1}$$

$$\mathcal{E}MSE(\overline{y}_{sys}) = \left(\frac{b^2}{12}\right)(k+1)(k-1) + \frac{\sigma^2}{n}\frac{N-n}{N} \tag{4.2}$$

$$\mathcal{E}MSE(\overline{y}_{mod}) = \mathcal{E}MSE(\overline{y}_{bal}) = \left(\frac{b^2}{12n^2}\right)(k+1)(k-1) + \frac{\sigma^2}{n}\frac{N-n}{N} \ (n: \text{ odd}) \tag{4.3}$$

$$\mathcal{E}MSE(\overline{y}_{cen}) = \frac{b^2}{4} + \frac{\sigma^2}{n}\frac{N-n}{N} \ (k: \text{ even}) \tag{4.4}$$

Here $\overline{y}_{ran}, \overline{y}_{sys}, \overline{y}_{mod}, \overline{y}_{bal}$ and $\overline{y}_{cen}$ denote the sample mean, which is used as the estimator of $\overline{Y}$, obtained from SRS, OSS, MSS, BSS and CSS, respectively.

On the basis of formulas (3.11) and (4.1) through (4.4), we can arrange the methods under consideration according to the magnitude of the expected mean square error as the following theorem. For simplicity's sake, $\mathcal{E}MSE(\overline{y}_{bi})$ is abbreviated as "bi", $\mathcal{E}MSE(\overline{y}_{sys})$ as "sys", and so on. Thus, for example, "$bi < sys$" means that BI is more efficient than OSS.

Theorem 3. *For a population having a linear trend represented by (3.3), the following hold (Here $T_k = 1 - 4A_k + 2kB_k$):*

(1) The case of $k = 2$ and $n = 3, 5, 7, \cdots$

(i) If $\sigma^2 < 9b^2/2$, then $bi < mod = bal < cen = sys < ran$.

(ii) If $9b^2/2 \le \sigma^2 < 9b^2n^2/2$, then $mod = bal \le bi < cen = sys < ran$.

(iii) If $9b^2n^2/2 \le \sigma^2 < 3b^2n^2(N+1)/2$, then $mod = bal < cen = sys \le bi < ran$.

(iv) If $3b^2n^2(N+1)/2 \le \sigma^2$, then $mod = bal < cen = sys < ran \le bi$.

(2) The case of $k = 4, 6, 8, \cdots$, $n = 3, 5, 7, \cdots$ and $n < \sqrt{(k^2-1)/3}$

(i) If $\sigma^2 < b^2n^2/2T_k$, then $bi < cen < mod = bal < sys < ran$.

(ii) If $b^2n^2/2T_k \le \sigma^2 < b^2(k^2-1)/6T_k$, then $cen \le bi < mod = bal < sys < ran$.

(iii) If $b^2(k^2-1)/6T_k \le \sigma^2 < b^2n^2(k^2-1)/6T_k$, then $cen < mod = bal \le bi < sys < ran$.

(iv) If $b^2n^2(k^2-1)/6T_k \le \sigma^2 < b^2n^2(N+1)(k-1)/6T_k$, then $cen < mod = bal < sys \le bi < ran$.

(v) If $b^2n^2(N+1)(k-1)/6T_k \le \sigma^2$, then $cen < mod = bal < sys < ran \le bi$.

(3) The case of $k = 4, 6, 8, \cdots$, $n = 3, 5, 7, \cdots$ and $n = \sqrt{(k^2-1)/3}$ (for example, $k = 26$ and $n = 15$)

(i) If $\sigma^2 < b^2n^2/6T_k$, then $bi < cen = mod = bal < sys < ran$.

(ii) If $b^2n^2/6T_k \le \sigma^2 < b^2n^2(k^2-1)/6T_k$, then $cen = mod = bal \le bi < sys < ran$.

(iii) If $b^2n^2(k^2-1)/6T_k \leq \sigma^2 < b^2n^2(N+1)(k-1)/6T_k$, *then* $cen = mod = bal < sys \leq bi < ran$.

(iv) If $b^2n^2(N+1)(k-1)/6T_k \leq \sigma^2$, *then* $cen = mod = bal < sys < ran \leq bi$.

(4) The case of $k = 4, 6, 8, \cdots, n = 3, 5, 7, \cdots$ *and* $n > \sqrt{(k^2-1)/3}$

(i) If $\sigma^2 < b^2(k^2-1)/6T_k$, *then* $bi < mod = bal < cen < sys < ran$.

(ii) If $b^2(k^2-1)/6T_k \leq \sigma^2 < b^2n^2/2T_k$, *then* $mod = bal \leq bi < cen < sys < ran$.

(iii) If $b^2n^2/2T_k \leq \sigma^2 < b^2n^2(k^2-1)/6T_k$, *then* $mod = bal < cen \leq bi < sys < ran$.

(iv) If $b^2n^2(k^2-1)/6T_k \leq \sigma^2 < b^2n^2(N+1)(k-1)/6T_k$, *then* $mod = bal < cen < sys \leq bi < ran$.

(v) If $b^2n^2(N+1)(k-1)/6T_k \leq \sigma^2$, *then* $mod = bal < cen < sys < ran \leq bi$.

Example 1. Suppose that we wish to draw a sample of size $n = 25$ from a population consisting of $N = 500$ units. We have $k = 500/25 = 20$. Assume that the slope of the linear trend is $b = 0.4$. Then by use of Mathematica we get $\psi(20.5) = 2.995836$, $\psi(10.5) = 2.303001$, $\psi^{(1)}(20.5) = 0.0499896$, $\psi^{(1)}(10.5) = 0.0999170$, and hence

$$\begin{aligned} A_{20} &= \frac{1}{2}\{\psi(20.5) - \psi(10.5)\} = 0.346418 \\ B_{20} &= -\frac{1}{4}\{\psi^{(1)}(20.5) - \psi^{(1)}(10.5)\} = 0.0124818 \\ T_{20} &= 1 - 4A_{20} + (2)(20)B_{20} = 0.113600. \end{aligned}$$

Therefore, by (4) of Theorem 3, the efficiency of the estimation methods can be compared as follows :

(i) If $\sigma^2 < 93.662$, then $bi < mod = bal < cen < sys < ran$.

(ii) If $93.662 \leq \sigma^2 < 440.141$, then $mod = bal \leq bi < cen < sys < ran$.

(iii) If $440.141 \leq \sigma^2 < 58538.732$, then $mod = bal < cen \leq bi < sys < ran$.

(iv) If $58538.732 \leq \sigma^2 < 1396566.901$, then $mod = bal < cen < sys \leq bi < ran$.

(v) If $1396566.901 \leq \sigma^2$, then $mod = bal < cen < sys < ran \leq bi$.

We can see from this example that BI is relatively efficient as compared with other methods unless σ^2 is preposterously large.

Now let us compare BI with methods which estimate $\overline{Y}$ by a weighted mean, not by the simple mean, of the sample values. The methods and the expected mean square errors of the resultant estimators are as follows :

(1) End corrections (EC) (See Yates (1948).)

$$\mathcal{E}MSE(\overline{y}_{ec}) = \frac{\sigma^2}{n}\frac{N-n}{N} + \frac{\sigma^2(k^2-1)}{6k^2(n-1)^2}.$$

(2) Modified systematic sampling with interpolation (MI) (See Kim (1998).)

$$\mathcal{E}MSE(\overline{y}_{mi}) = \frac{\sigma^2}{n}\frac{N-n}{N} + \frac{\sigma^2}{12n^2}(4 - 12A_k + 6kB_k - \frac{1}{k^2}) \quad (k : \text{even}, n : \text{odd}, n \geq 3),$$

where A_k and B_k are as defined in Theorem 1.

(3) Balanced systematic sampling with interpolation and extrapolation (BIE) (See Kim (1999).)

$$\mathcal{E}MSE(\overline{y}_{bie}) = \frac{\sigma^2}{n}\frac{N-n}{N} + \frac{\sigma^2}{2n^2}(1 - \gamma - 2ln2 + C_k) \quad (k : \text{even}, n : \text{odd}, n \geq 3),$$

where $\gamma = 0.577215\cdots$ is the Euler constant, and $C_k = \frac{k}{8}\{\pi^2 - 2\psi^{(1)}(k+\frac{1}{2})\} - \psi(k+\frac{1}{2})$.

Theorem 4. *Consider the four methods : EC, MI, BIE and BI. For k even and $n(\geq 5)$ odd, the following holds :*

$$\mathcal{E}MSE(\overline{y}_{bi}) < \mathcal{E}MSE(\overline{y}_{mi}) < \mathcal{E}MSE(\overline{y}_{ec}) < \mathcal{E}MSE(\overline{y}_{bie}).$$

This means that BI is the most efficient of these four methods.

The proof of Theorem 4 is also given in Appendix. For various k, the values of the second terms of $\mathcal{E}MSE(\cdot)$'s for EC, MI, BIE and BI are given in Table 4.1. Note that the first terms are all the same for the four methods. We can see again that BI is the most efficient of the four methods.

Table 4.1 : The values of the second terms of $\mathcal{E}MSE(\cdot)$'s for EC, MI, BIE and BI

k	EC	MI	BIE	BI
4	$0.1563\sigma^2/(n-1)^2$	$0.1061\sigma^2/n^2$	$1.1668\sigma^2/n^2$	$0.0559\sigma^2/n^2$
8	$0.1641\sigma^2/(n-1)^2$	$0.1103\sigma^2/n^2$	$3.2882\sigma^2/n^2$	$0.0566\sigma^2/n^2$
12	$0.1655\sigma^2/(n-1)^2$	$0.1111\sigma^2/n^2$	$5.5529\sigma^2/n^2$	$0.0567\sigma^2/n^2$
16	$0.1660\sigma^2/(n-1)^2$	$0.1114\sigma^2/n^2$	$7.8765\sigma^2/n^2$	$0.0568\sigma^2/n^2$
20	$0.1663\sigma^2/(n-1)^2$	$0.1115\sigma^2/n^2$	$10.2324\sigma^2/n^2$	$0.0568\sigma^2/n^2$
∞	$0.1667\sigma^2/(n-1)^2$	$0.1117\sigma^2/n^2$	∞	$0.0569\sigma^2/n^2$

Example 2. The following data for a small artificial population were adopted from Cochran (1977, p.211). Some modification has been made to the original data in order to make k even and n odd. We draw a sample of size $n = 9$ from this population of size $N = 36$.

1	2	5	4	7	7	8	6	6	8
9	10	13	12	15	16	16	17	18	19
20	20	24	23	25	28	29	27	26	30
31	33	32	35	37	38				

The mean of this population is $\overline{Y} = 18.2500$. As we can see from Figure 4.1, this population is exhibiting a linear increasing trend. The MSEs of the estimators of $\overline{Y}$ by the existing methods are

$$MSE(\overline{y}_{ran}) = 9.8351, \quad MSE(\overline{y}_{sys}) = 1.9653, \quad MSE(\overline{y}_{mod}) = 0.1875,$$
$$MSE(\overline{y}_{bal}) = 0.1875, \quad MSE(\overline{y}_{cen}) = 1.0224, \quad MSE(\overline{y}_{ec}) = 0.2667,$$
$$MSE(\overline{y}_{mi}) = 0.2446, \quad MSE(\overline{y}_{bie}) = 0.2831.$$

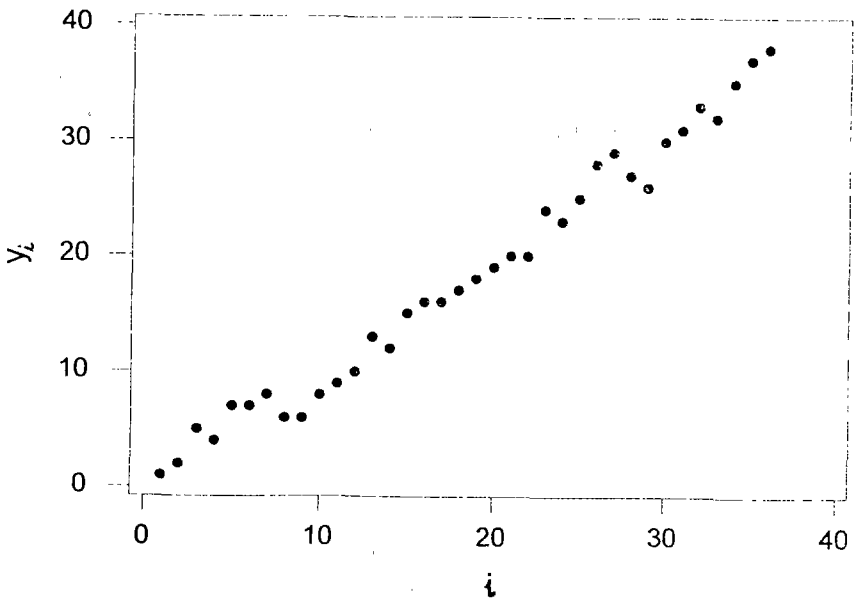

Figure 4.1 : The population in Example 2

On the other hand, using BI proposed in this paper, $\overline{Y}$ is estimated by one of the following twelve values :

$$\begin{array}{lll} \overline{y}_1'^*(3) = 17.7937, & \overline{y}_1'^*(5) = 17.7222, & \overline{y}_1'^*(7) = 17.7460, \\ \overline{y}_2'^*(3) = 18.7444, & \overline{y}_2'^*(5) = 18.7444, & \overline{y}_2'^*(7) = 18.7000, \\ \overline{y}_3'^*(3) = 18.5333, & \overline{y}_3'^*(5) = 18.4889, & \overline{y}_3'^*(7) = 18.4556, \\ \overline{y}_4'^*(3) = 18.1508, & \overline{y}_4'^*(5) = 18.0794, & \overline{y}_4'^*(7) = 18.0556. \end{array}$$

Therefore, the MSE of our estimator $\overline{y}_{bi}$ is computed as

$$MSE(\overline{y}_{bi}) = 0.1407,$$

which shows that BI is the most efficient of the methods considered.

5. Concluding remarks

In this paper, we proposed a new method for estimating the mean of a population of size $N = kn$ which has a linear trend, for the case of k even and n odd ($n \geq 5$). The proposed method, BI, consists of selecting a sample of size n by BSS, and then estimating the population mean by using the concept of interpolation.

BI turned out to be relatively efficient as compared with the conventional methods if σ^2, the variance of the random error term in the infinite superpopulation model, is not preposterously large. It was found to be especially efficient as σ^2 becomes smaller. Moreover, BI was found to be more efficient than EC, MI and BIE.

APPENDIX

(1) Proof of Theorem 1

We know that

$$MSE(\overline{y}_{bi}) = \frac{2}{k(n-3)} \sum_{i=1}^{k} \sum_{m} \{\overline{y}_i'^*(m) - \overline{Y}\}^2, \tag{A.1}$$

and by (3.1) we obtain

$$\overline{Y} = \overline{\mu} + \overline{e}. \tag{A.2}$$

On the other hand, from (3.1) it can be written that

$$y'_{ij} = \mu'_{ij} + e'_{ij} \ (i = 1, 2, \cdots, k;\ j = 1, 2, \cdots, n), \tag{A.3}$$

from which we obtain

$$\overline{y}'^{*}_{i}(m) = \overline{\mu}'^{*}_{i}(m) + \overline{e}'^{*}_{i}(m) \ (i = 1, 2, \cdots, k; m = 3, 5, \cdots, n-2). \tag{A.4}$$

Substituting (A.2) and (A.4) into (A.1) and taking expectation, we have

$$\begin{aligned} \mathcal{E}MSE(\overline{y}_{bi}) &= \frac{2}{k(n-3)} \sum_{i=1}^{k} \sum_{m} \mathcal{E}[\{(\overline{\mu}'^{*}_{i}(m) - \overline{\mu}) + (\overline{e}'^{*}_{i}(m) - \overline{e})\}^2] \\ &= \frac{2}{k(n-3)} \sum_{i=1}^{k} \sum_{m} [\{\overline{\mu}'^{*}_{i}(m) - \overline{\mu}\}^2 + \mathcal{E}\{(\overline{e}'^{*}_{i}(m) - \overline{e})^2\}]. \end{aligned} \tag{A.5}$$

We also have, for $i = 1, 2, \cdots, k/2$,

$$\begin{aligned} \mathcal{E}[\{\overline{e}'^{*}_{i}(m) - \overline{e}\}^2] &= \mathcal{E}[\{\overline{e}'_{i} - \overline{e} + P_i(m)\}^2] \\ &= \mathcal{E}\{(\overline{e}'_{i} - \overline{e})^2\} + 2\mathcal{E}\{(\overline{e}'_{i} - \overline{e})P_i(m)\} + \mathcal{E}\{P_i^2(m)\}, \end{aligned} \tag{A.6}$$

where

$$P_i(m) = \frac{k + 1 - 2i}{2n(2k + 1 - 2i)} (e'_{i,m+1} - e'_{im}). \tag{A.7}$$

We further have, for $i = 1, 2, \cdots, k/2$,

$$\mathcal{E}\{(\overline{e}'_{i} - \overline{e})^2\} = \mathcal{E}\{(\overline{e}'_{i})^2\} - 2\mathcal{E}\{(\overline{e}'_{i})(\overline{e})\} + \mathcal{E}\{(\overline{e})^2\} \tag{A.8}$$

and

$$\mathcal{E}\{(\overline{e}'_{i})^2\} = \mathcal{E}\left\{\left(\frac{1}{n}\sum_{j=1}^{n} e'_{ij}\right)^2\right\}$$

$$= \frac{1}{n^2}\mathcal{E}\left\{\sum_{j=1}^{n}(e'_{ij})^2 + 2\sum_{j<j'}(e'_{ij})(e'_{ij'})\right\}$$
$$= \frac{1}{n^2}\left[\sum_{j=1}^{n}\mathcal{E}\{(e'_{ij})^2\} + 2\sum_{j<j'}\mathcal{E}\{(e'_{ij})(e'_{ij'})\}\right]$$
$$= \frac{1}{n^2}(n\sigma^2 + 0) \text{ (by the assumptions on the model)}$$
$$= \frac{\sigma^2}{n}, \tag{A.9}$$

and similarly

$$\mathcal{E}\{(\overline{e}'_i)(\overline{e})\} = \mathcal{E}\{(\overline{e})^2\} = \frac{\sigma^2}{N}. \tag{A.10}$$

The second term in the rightmost side of (A.6) is easily shown to be zero, and the third term is also easily obtained as

$$\mathcal{E}\{P_i^2(m)\} = \frac{(k+1-2i)^2\sigma^2}{2n^2(2k+1-2i)^2}. \tag{A.11}$$

Substitution of these results into (A.6) gives

$$\mathcal{E}[\{\overline{e}'^*_i(m) - \overline{e}\}^2] = \frac{\sigma^2}{n}\frac{N-n}{N} + \frac{(k+1-2i)^2\sigma^2}{2n^2(2k+1-2i)^2} \ (i = 1, 2, \cdots, k/2). \tag{A.12}$$

For $i = k/2+1, k/2+2, \cdots, k$, we obtain

$$\mathcal{E}[\{\overline{e}'^*_i(m) - \overline{e}\}^2] = \frac{\sigma^2}{n}\frac{N-n}{N} + \frac{(2i-k-1)^2\sigma^2}{2n^2(2i-1)^2} \tag{A.13}$$

by quite a similar method to that used in the above.

Substituting (A.12) and (A.13) into (A.5), we have

$$\begin{aligned}\mathcal{E}MSE(\overline{y}_{bi}) &= \frac{2}{k(n-3)}\Big[\sum_{i=1}^{k}\sum_{m}\{\overline{\mu}'^*_i(m) - \overline{\mu}\}^2 \\ &\quad + \sum_{i=1}^{k/2}\sum_{m}\Big\{\frac{\sigma^2}{n}\frac{N-n}{N} + \frac{(k+1-2i)^2\sigma^2}{2n^2(2k+1-2i)^2}\Big\} \\ &\quad + \sum_{i=k/2+1}^{k}\sum_{m}\Big\{\frac{\sigma^2}{n}\frac{N-n}{N} + \frac{(2i-k-1)^2\sigma^2}{2n^2(2i-1)^2}\Big\}\Big]\end{aligned}$$

$$= \frac{2}{k(n-3)} \sum_{i=1}^{k} \sum_{m} \{\overline{\mu}_i'^{*}(m) - \overline{\mu}\}^2 + \frac{\sigma^2}{n} \frac{N-n}{N}$$
$$+ \frac{\sigma^2}{2kn^2} \{ \sum_{i=1}^{k/2} (1 - \frac{k}{2k+1-2i})^2 + \sum_{i=k/2+1}^{k} (1 - \frac{k}{2i-1})^2 \}, \quad \text{(A.14)}$$

and by using the following facts concerning summation (See, for reference, Abramowitz and Stegun (1982, p.258)), we obtain (3.2) after straightforward calculation.

$$\sum_{i=1}^{k/2} \frac{1}{2k+1-2i} = \sum_{i=k/2+1}^{k} \frac{1}{2i-1} = \frac{1}{2} \left\{ \psi\left(k + \frac{1}{2}\right) - \psi\left(\frac{k+1}{2}\right) \right\}$$
$$\sum_{i=1}^{k/2} \frac{1}{(2k+1-2i)^2} = \sum_{i=k/2+1}^{k} \frac{1}{(2i-1)^2} = -\frac{1}{4} \left\{ \psi^{(1)}\left(k + \frac{1}{2}\right) - \psi^{(1)}\left(\frac{k+1}{2}\right) \right\}$$

□

(2) Proof of Theorem 4

We prove here the first inequality : $\mathcal{E}MSE(\overline{y}_{bi}) < \mathcal{E}MSE(\overline{y}_{mi})$. The remaining inequalities can be proved by quite similar methods.

We have

$$\mathcal{E}MSE(\overline{y}_{mi}) - \mathcal{E}MSE(\overline{y}_{bi}) = \frac{\sigma^2}{12n^2}(12A_k - 6kB_k - 2 - \frac{1}{k^2}). \quad \text{(A.15)}$$

Using the fact that

$$A_k = \sum_{i=k/2+1}^{k} \frac{1}{2i-1} \text{ and } B_k = \sum_{i=k/2+1}^{k} \frac{1}{(2i-1)^2},$$

the quantity in the parentheses of the righthand side of (A.15) can be written as

$$12A_k - 6kB_k - 2 - \frac{1}{k^2}$$
$$= \sum_{i=k/2+1}^{k} \{\frac{12}{2i-1} - \frac{6k}{(2i-1)^2} - \frac{4}{k} - \frac{2}{k^3}\}$$
$$= \sum_{i=k/2+1}^{k} \frac{12k^3(2i-1) - 6k^4 - 4k^2(2i-1)^2 - 2(2i-1)^2}{k^3(2i-1)^2}$$
$$= \sum_{i=k/2+1}^{k} \frac{(-16k^2-8)i^2 + (24k^3+16k^2+8)i - 6k^4 - 12k^3 - 4k^2 - 2}{k^3(2i-1)^2}.$$

Let us consider the numerator in the last expression. Since the coefficient of i^2 is negative, the numerator takes its minimum at $i = k/2+1$ or at $i = k$; and the values taken by the numerator at these two values of i are $2k^4+4k^3-6k^2-4k-2$ (at $i = k/2+1$) and $2k^4+4k^3-12k^2+8k-2$ (at $i = k$). Since both of these two values are positive for $k \geq 2$, we can easily see that the numerator is positive for each i, from which it follows that $\mathcal{E}MSE(\overline{y}_{bi}) < \mathcal{E}MSE(\overline{y}_{mi})$.

□

REFERENCES

Abramowitz, M. and Stegun, I. A.(1982). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables.* National Bureau of Standards, U.S. Department of Commerce.

Bellhouse, D.R. and J.N.K. Rao (1975). Systematic sampling in the presence of a trend. *Biometrika*, 62, 694-697.

Cochran, W.G.(1946). Relative accuracy of systematic and stratified random samples for a certain class of populations. *Annals of Mathematical Statistics*, 17, 164-177.

Cochran, W.G.(1977). *Sampling Techniques* (3rd ed.), John Wiley and Sons, Inc., New York.

Kim, H.J.(1985). New systematic sampling methods for populations with linear or parabolic trends. Unpublished Master Thesis, Department of Computer Science and Statistics, Seoul National University.

Kim, H.J.(1998). Estimation of population mean using interpolation in modified systematic sampling. *Korean Annals of Mathematics*, 15, 217-231.

Kim, H.J.(1999). A study on estimating population mean by use of interpolation and extrapolation with balanced systematic sampling. *Journal of the Korean Data and Information Science Society*, 10, 91-102.

Madow, W.G.(1953). On the theory of systematic sampling, III. Comparison of centered and random start systematic sampling. *Annals of Mathematical Statistics*, 24, 101-106.

Murthy, M.N.(1967). *Sampling Theory and Methods.* Statistical Publishing Society, Calcutta, India.

Sethi, V.K.(1965). On optimum pairing of units. *Sankhya*, B, 27, 315-320.

Singh,D., K.K. Jindal and J.N. Garg (1968). On modified systematic sampling. *Biometrika*, 55, 541-546.

Yates, F.(1948). Systematic sampling. *Philosophical Transactions of the Royal Society of London*, A, 241, 345-377.