The Korean Communications
in Statistics Vol. 4, No. 3, 1997
pp. 743 - 752

A Second Type of Centered Balanced Systematic Sampling Method[1]

Hyuk Joo Kim[2]

Abstract

Kim (1985) proposed the so-called "centered balanced systematic sampling" for estimating the mean of a population with a linear trend. In this paper, a version of this sampling method is proposed. It is shown that this version is as efficient as the original method from the viewpoint of the expected mean square error criterion. It is also shown to be quite an efficient method as compared with other existing methods.

1. Introduction

When conducting statistical surveys, we sometimes encounter 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, there is expected to be a linear trend in this population.

Several researchers have proposed sampling methods for estimating the mean of such a population. In particular, *centered balanced systematic sampling* (CBSS) was proposed by Kim (1985). This sampling method turned out to be quite efficient as compared with existing methods.

In this paper, some modification will be made with CBSS when n (the sample size) is an odd number and k (the reciprocal of the sampling fraction) is an even number. The resulting method will be compared with the original method and other methods under the expected mean square error criterion based on the infinite superpopulation model introduced by Cochran (1946).

2. CBSS1 and CBSS2

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.

1) This paper was supported by Wonkwang University in 1996.
2) Associate Professor, Department of Statistics, Wonkwang University, Iksan, 570-749, Korea.

2.1. Description of CBSS1

CBSS proposed by Kim (1985), which will be called *CBSS1* from now on, is briefly described.

If $k=N/n$ is an odd number, CBSS1 selects the units $U_{(k+1)/2+(j-1)k}$ for $j=1,2,\cdots,n$. For example, if $N=25$, $n=5$ and $k=5$, then $U_3, U_8, U_{13}, U_{18}, U_{23}$ are selected. Thus in this case (odd k) CBSS1 is the same as centered systematic sampling (CSS) proposed by Madow (1953).

Let us concentrate on the case when k is an even number in the remaining part of this section. In this case, either C'_1 or C'_2 is selected with respective probabilities 1/2. Here, the clusters C'_1 and C'_2 are as follows :

$$C'_1=\{U_{(k/2)(4j-3)} : j=1,2,\cdots,n/2\}\cup\{U_{1+(k/2)(4j-1)} : j=1,2,\cdots,n/2\}$$

$$C'_2=\{U_{1+(k/2)(4j-3)} : j=1,2,\cdots,n/2\}\cup\{U_{(k/2)(4j-1)} : j=1,2,\cdots,n/2\}$$

for n even, and

$$C'_1=\{U_{(k/2)(4j-3)} : j=1,2,\cdots,(n+1)/2\}\cup\{U_{1+(k/2)(4j-1)} : j=1,2,\cdots,(n-1)/2\}$$

$$C'_2=\{U_{1+(k/2)(4j-3)} : j=1,2,\cdots,(n+1)/2\}\cup\{U_{(k/2)(4j-1)} : j=1,2,\cdots,(n-1)/2\}$$

for n odd.

For example, if $N=24$, $n=6$, $k=4$, then $C'_1=\{U_2, U_7, U_{10}, U_{15}, U_{18}, U_{23}\}$ and $C'_2=\{U_3, U_6, U_{11}, U_{14}, U_{19}, U_{22}\}$, and if $N=20$, $n=5$, $k=4$, then $C'_1=\{U_2, U_7, U_{10}, U_{15}, U_{18}\}$ and $C'_2=\{U_3, U_6, U_{11}, U_{14}, U_{19}\}$. It is to be noted that CBSS1 is obtained by combining the ideas of CSS and balanced systematic sampling (BSS), which was proposed by Sethi (1965) and named by Murthy (1967).

Let y_i denote the value for U_i (the ith unit in the population)($i=1,2,\cdots,N$). Also let the value for the jth unit in C'_i be denoted by y'_{ij} ($i=1,2$; $j=1,2,\cdots,n$), and let the mean value for the units in C'_i be denoted by $\bar{y}'_i$ ($i=1,2$). For example, if $N=20$, $n=5$, $k=4$, then $y'_{13}=y_{10}$, $y'_{25}=y_{19}$, $\bar{y}'_1=\sum_{j=1}^{5}y'_{1j}/5 = (y_2+y_7+y_{10}+y_{15}+y_{18})/5$, etc.

The population mean $\bar{Y}=\sum_{i=1}^{N}y_i/N$ is estimated by $\bar{y}'_1$ or $\bar{y}'_2$ according as C'_1 or C'_2 is selected. That is, the estimator $\bar{y}_{cb1}$ of $\bar{Y}$ by CBSS1 has

$$P(\bar{y}_{cb1}=\bar{y}'_1)=P(\bar{y}_{cb1}=\bar{y}'_2)=\frac{1}{2}$$

and mean square error

$$MSE(\bar{y}_{cb1}) = \frac{1}{2}\{(\bar{y}'_1 - \bar{Y})^2 + (\bar{y}'_2 - \bar{Y})^2\}.$$

2.2. Proposition of CBSS2

Consider the case when k is even and n is odd. For example, suppose that N=20, n=5 and k=4. As was seen in Section 2.1, CBSS1 selects either $C'_1 = \{U_2, U_7, U_{10}, U_{15}, U_{18}\}$ and $C'_2 = \{U_3, U_6, U_{11}, U_{14}, U_{19}\}$ with respective probabilities 1/2. We notice that the sums of the numbers assigned to the units in C'_1 and C'_2 are, respectively, 52 and 53, showing a difference of 1. Such a difference is inevitable in the case of odd n. Suppose now that U_{18} in C'_1 is replaced by U_{19}, and instead U_{19} in C'_2 is replaced by U_{18}. Let us denote the resultant clusters as C^*_1 and C^*_2, that is, $C^*_1 = \{U_2, U_7, U_{10}, U_{15}, U_{19}\}$ and $C^*_2 = \{U_3, U_6, U_{11}, U_{14}, U_{18}\}$. The sums of the unit numbers in C^*_1 and C^*_2 now become 53 and 52, respectively, giving a difference of 1, which is the same as before.

Having this motivation in mind, we can introduce the following method, which we expect to be efficient, on the average, in the same degree as CBSS1. Let us define two clusters C^*_1 and C^*_2 as follows :

$$C^*_1 = (C'_1 - \{U_{(k/2)(2n-1)}\}) \cup \{U_{1+(k/2)(2n-1)}\}$$

$$C^*_2 = (C'_2 - \{U_{1+(k/2)(2n-1)}\}) \cup \{U_{(k/2)(2n-1)}\}$$

We propose a sampling method such that either C^*_1 or C^*_2 is selected with respective probabilities $1/2$. This sampling method will henceforth be referred to as *CBSS2*.

Let y^*_{ij} $(i=1,2\ ;\ j=1,2,\cdots,n)$ and $\bar{y}^*_i$ $(i=1,2)$ denote, respectively, the value for the jth unit in C^*_i and the mean value for the units in C^*_i . The population mean $\bar{Y}$ is estimated by $\bar{y}^*_1$ or $\bar{y}^*_2$ according as C^*_1 or C^*_2 is selected. If we let $\bar{y}_{cb2}$ denote the estimator of $\bar{Y}$ by CBSS2, then

$$P(\bar{y}_{cb2} = \bar{y}^*_1) = P(\bar{y}_{cb2} = \bar{y}^*_2) = \frac{1}{2}.$$

$\bar{y}_{cb2}$ is generally a biased estimator for $\bar{Y}$ and has mean square error

$$MSE(\bar{y}_{cb2}) = \frac{1}{2}\{(\bar{y}^*_1 - \bar{Y})^2 + (\bar{y}^*_2 - \bar{Y})^2\}.$$

3. Expected mean square error for CBSS2

In this section, the efficiency of $\bar{y}_{cb2}$ is compared with that of $\bar{y}_{cb1}$ using the expected mean square error criterion based on 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

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

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

From now on, with regard to μ and e also we will use the same notation as adopted for y. That is, $\bar{\mu}^*_i$ denotes the mean μ value for the units in C^*_i, e^*_{ij} denotes the random error for the jth unit in C^*_i, and so on.

The following theorem is very important in evaluating the efficiency of CBSS2.

Theorem. Assuming the model (3.1), the expected mean square error of $\bar{y}_{cb2}$ for k even and n odd is

$$EMSE(\bar{y}_{cb2}) = \frac{1}{2}\{(\bar{\mu}^*_1 - \bar{\mu})^2 + (\bar{\mu}^*_2 - \bar{\mu})^2\} + \sigma^2\left(\frac{1}{n} - \frac{1}{N}\right). \tag{3.2}$$

Proof. We know that

$$MSE(\bar{y}_{cb2}) = \frac{1}{2}\{(\bar{y}^*_1 - \bar{Y})^2 + (\bar{y}^*_2 - \bar{Y})^2\}, \tag{3.3}$$

and by (3.1) we obtain

$$\bar{Y} = \bar{\mu} + \bar{e}. \tag{3.4}$$

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

$$y^*_{ij} = \mu^*_{ij} + e^*_{ij} \quad (i = 1, 2;\ j = 1, 2, \cdots, n), \tag{3.5}$$

from which we obtain

$$\bar{y}^*_i = \bar{\mu}^*_i + \bar{e}^*_i \quad (i = 1, 2). \tag{3.6}$$

Substituting (3.4) and (3.6) into (3.3) and taking expectation, we have

$$\begin{aligned} EMSE(\bar{y}_{cb2}) &= \frac{1}{2}E[\{(\bar{\mu}^*_1 - \bar{\mu}) + (\bar{e}^*_1 - \bar{e})\}^2 + \{(\bar{\mu}^*_2 - \bar{\mu}) + (\bar{e}^*_2 - \bar{e})\}^2] \\ &= \frac{1}{2}[(\bar{\mu}^*_1 - \bar{\mu})^2 + E\{(\bar{e}^*_1 - \bar{e})^2\} + (\bar{\mu}^*_2 - \bar{\mu})^2 + E\{(\bar{e}^*_2 - \bar{e})^2\}] \\ &= \frac{1}{2}\{(\bar{\mu}^*_1 - \bar{\mu})^2 + (\bar{\mu}^*_2 - \bar{\mu})^2\} + \frac{1}{2}\sum_{i=1}^{2} E\{(\bar{e}^*_i - \bar{e})^2\}. \end{aligned} \tag{3.7}$$

We further have, for $i=1,2$,

$$E\{(\bar{e}^*_i-\bar{e})^2\}=E\{(\bar{e}^*_i)^2\}-2E\{(\bar{e}^*_i)(\bar{e})\}+E\{(\bar{e})^2\} \tag{3.8}$$

and

$$\begin{aligned} E\{(\bar{e}^*_i)^2\} &= E\{(\frac{1}{n}\sum_{j=1}^{n} e^*_{ij})^2\} \\ &= \frac{1}{n^2}E\{\sum_{j=1}^{n}(e^*_{ij})^2+2\sum_{j<j'}(e^*_{ij})(e^*_{ij'})\} \\ &= \frac{1}{n^2}[\sum_{j=1}^{n}E\{(e^*_{ij})^2\}+2\sum_{j<j'}E\{(e^*_{ij})(e^*_{ij'})\}] \\ &= \frac{1}{n^2}(n\sigma^2+0) \quad \text{(by the assumptions of the model)} \\ &= \frac{\sigma^2}{n}, \end{aligned} \tag{3.9}$$

and similarly

$$E\{(\bar{e}^*_i)(\bar{e})\}=E\{(\bar{e})^2\}=\frac{\sigma^2}{N}. \tag{3.10}$$

Substitution of (3.9) and (3.10) into (3.8) gives

$$E\{(\bar{e}^*_i-\bar{e})^2\}=\sigma^2\left(\frac{1}{n}-\frac{1}{N}\right) \qquad (i=1,2), \tag{3.11}$$

and (3.2) is immediately obtained from (3.7) and (3.11). This completes the proof. ■

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

$$y_i=a+bi+e_i \quad (i=1,2,\cdots,N). \tag{3.12}$$

This is the case of a population with a linear trend.

In this case, we can obtain the following formulas (See Appendix for derivation) :

$$\bar{\mu}=a+(\frac{b}{2})(N+1) \tag{3.13}$$

$$\bar{\mu}^*_1=a+(\frac{b}{2})(N+1)+\frac{b}{2n} \tag{3.14}$$

$$\bar{\mu}^*_2=a+(\frac{b}{2})(N+1)-\frac{b}{2n} \tag{3.15}$$

Substituting these formulas into (3.2), we obtain

$$EMSE(\bar{y}_{c\ell 2})=\frac{b^2}{4n^2}+\sigma^2(\frac{1}{n}-\frac{1}{N}) \quad (k: \text{even},\ n: \text{odd}). \tag{3.16}$$

(3.16) now needs to be compared with the expected mean square error of $\bar{y}_{cb1}$ by CBSS1. It was obtained in Kim (1985) that

$$EMSE(\bar{y}_{cb1}) = \frac{b^2}{4n^2} + \sigma^2\left(\frac{1}{n} - \frac{1}{N}\right) \quad (k : \text{even}, \; n : \text{odd}), \tag{3.17}$$

which is equal to (3.16). Consequently, we can state that $\bar{y}_{cb1}$ and $\bar{y}_{cb2}$ are equally efficient from the viewpoint of the expected mean square error criterion.

4. Comparison with other methods

In this section, the efficiency of CBSS2 is compared with that of other methods. Let us consider simple random sampling (SRS), ordinary systematic sampling (OSS), balanced systematic sampling (BSS), centered systematic sampling (CSS), modified systematic sampling (MSS) proposed by Singh et al. (1968), and centered modified systematic sampling (CMSS) proposed by Kim (1985). Discussions on comparisons of the performances of BSS, CSS, MSS and OSS are also given in Bellhouse and Rao (1975).

For a population characterized by the model (3.12), the following were obtained in Kim (1985):

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

$$EMSE(\bar{y}_{sy}) = \left(\frac{b^2}{12}\right)(k+1)(k-1) + \sigma^2\left(\frac{1}{n} - \frac{1}{N}\right) \tag{4.2}$$

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

$$EMSE(\bar{y}_{cen}) = \frac{b^2}{4} + \sigma^2\left(\frac{1}{n} - \frac{1}{N}\right) \quad (k : \text{even}) \tag{4.4}$$

$$EMSE(\bar{y}_{cm}) = \frac{b^2}{4n^2} + \sigma^2\left(\frac{1}{n} - \frac{1}{N}\right) \quad (k : \text{even}, \; n : \text{odd}) \tag{4.5}$$

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

On the basis of formulas (3.16), (3.17), and (4.1) through (4.5), we can arrange the methods under consideration according to the magnitude of the expected mean square error as follows. For simplicity's sake, $EMSE(\bar{y}_{cb2})$ is denoted as " $cb2$ ", $EMSE(\bar{y}_{cen})$ as " cen ", and so on. Thus, for example, " $cen > cb2$ " means that CBSS2 is more efficient than CSS. We only consider the case of $n = 3, 5, 7, \cdots$ since the case of $n = 1$ does not have practical meaning.

(1) If $k=2$ and $n=3,5,7,\cdots$, then

$$ran > sy = cen > bal = mod = cm = cb1 = cb2.$$

(2) If $k=4,6,8,\cdots$, $n=3,5,7,\cdots$, and $n \le \sqrt{(k^2-1)/3}$, then

$$ran > sy > bal = mod \ge cen > cm = cb1 = cb2.$$

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

$$ran > sy > cen > bal = mod > cm = cb1 = cb2.$$

As we see from the above result, CBSS2, together with CBSS1 and CMSS, is quite efficient as compared with other methods in each case.

Example. The following data, adopted from Cochran (1977, p.211), are for a small artificial population that exhibits a fairly steady rising trend. We have N=40, k=8, n=5.

0	1	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	31	33	32	35	37	38	38

The mean and the variance of this population are $\overline{Y}=18.175$ and $S^2=136.251$, respectively. The possible samples and MSEs of the estimators of $\overline{Y}$ by various sampling methods are given in Table 1. For example, $MSE(\overline{y}_{cb2})$ is calculated as follows :

$$\begin{aligned} MSE(\overline{y}_{cb2}) &= \frac{1}{2}\ \{(\overline{y}^*_1-\overline{Y})^2+(\overline{y}^*_2-\overline{Y})^2\} \\ &= \frac{1}{2}\ \{(18.4-18.175)^2+(18.2-18.175)^2\} \\ &= 0.026 \end{aligned}$$

As we see in Table 1, CBSS2, together with CMSS, is the most efficient for this population among the eight methods considered.

Remark. As stated in Cochran (1977, p.213), the results proved in Sections 3 and 4 do not apply to any single finite population (i.e., to any specific set of values $y_1, y_2, \cdots, y_N$) but to the average of all finite populations that can be drawn from the infinite superpopulation. It is to be understood from this viewpoint that CBSS1 is less efficient than CBSS2 and CMSS in the above example.

Table 1. Possible samples and MSEs of the estimators of $\overline{Y}$ by various sampling methods (for population in Example)

Sampling method	Possible samples	MSE
SRS	$\binom{40}{5}$ kinds	23.844
OSS	{0, 8, 15, 24, 31} {1, 6, 16, 23, 31} {1, 6, 16, 25, 33} {2, 8, 17, 28, 32} {5, 9, 18, 29, 35} {4, 10, 19, 27, 37} {7, 13, 20, 26, 38} {7, 12, 20, 30, 38}	4.834
BSS	{0, 12, 15, 30, 31} {1, 13, 16, 26, 31} {1, 10, 16, 27, 33} {2, 9, 17, 29, 32} {5, 8, 18, 28, 35} {4, 6, 19, 25, 37} {7, 6, 20, 23, 38} {7, 8, 20, 24, 38}	0.494
CSS	{2, 8, 17, 28, 32} {5, 9, 18, 29, 35}	0.826
MSS	{0, 8, 15, 30, 38} {1, 6, 16, 26, 38} {1, 6, 16, 27, 37} {2, 8, 17, 29, 35} {5, 9, 18, 28, 32} {4, 10, 19, 25, 33} {7, 13, 20, 23, 31} {7, 12, 20, 24, 31}	0.254
CMSS	{2, 8, 17, 29, 35} {5, 9, 18, 28, 32}	0.026
CBSS1	{2, 9, 17, 29, 32} {5, 8, 18, 28, 35}	0.266
CBSS2	{2, 9, 17, 29, 35} {5, 8, 18, 28, 32}	0.026

5. Concluding remarks

Several sampling methods have been introduced so far in order to estimate the mean of a population which has a linear trend. Among them, CBSS1 proposed by Kim (1985) was seen to be a desirable method for such a type of population.

In this paper, for the case of k even and n odd, a second type of CBSS was proposed and named CBSS2. It can be regarded as a kind of controlled selection method which is derived from balanced systematic sampling. It restricts the number of possible samples to two, among the k possible samples that can be drawn by BSS. It was shown that CBSS2 gives an estimator with the expected mean square error which is equal to that resulting from CBSS1. It was also shown that CBSS2, together with CBSS1, is quite efficient as compared with other sampling methods. Like other types of systematic sampling, CBSS2 can be easily used in real situations because the sampling procedure is simple.

References

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

[2] 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.

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

[4] 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.

[5] 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.

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

[7] Sethi, V. K. (1965). On optimum pairing of units, *Sankhya*, **B27**, 315-320.

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

Appendix : Derivation of formulas (3.13) through (3.15)

(1) $$\bar{\mu} = \frac{1}{N} \sum_{i=1}^{N} \mu_i$$
$$= \frac{1}{N} \sum_{i=1}^{N} (a + bi)$$
$$= \frac{1}{N} (Na + b \sum_{i=1}^{N} i)$$
$$= a + (\frac{b}{2})(N+1)$$

where we used $\sum_{i=1}^{N} i = N(N+1)/2$.

(2) $$\bar{\mu}^*_1 = \frac{1}{n} \sum_{j=1}^{n} \mu^*_{1j}$$
$$= \frac{1}{n} \left\{ \sum_{j=1}^{(n-1)/2} \mu_{(k/2)(4j-3)} + \sum_{j=1}^{(n-1)/2} \mu_{1+(k/2)(4j-1)} + \mu_{1+(k/2)(2n-1)} \right\}$$
$$= \frac{1}{n}\left[\sum_{j=1}^{(n-1)/2} \left\{a + b(\frac{k}{2})(4j-3)\right\} + \sum_{j=1}^{(n-1)/2} \left\{a + b(1 + (\frac{k}{2})(4j-1))\right\} + a + b\left\{1 + (\frac{k}{2})(2n-1)\right\}\right]$$

By using $\sum_{j=1}^{(n-1)/2} j = (n+1)(n-1)/8$, we obtain formula (3.14) by straightforward algebra.

(3) Formula (3.15) is derived by quite a similar method to (3.14).