The Korean Communications
in Statistics Vol. 6, No. 2, 1999
pp. 457-466

On Centralizing the Modified Systematic Sampling Method for Populations with Linear Trends

Hyuk Joo Kim[1)]

Abstract

'Centered modified systematic sampling (CMSS)' was proposed by Kim (1985) for estimating the mean of a population with a linear trend. In the present paper, a version of this sampling method is suggested. This version turns out to be efficient in the same degree 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. An illustrative example is given.

1. Introduction

When we perform sampling inspections or 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, it is expected that there exists a linear trend in this population.

The mean of such a population can be efficiently estimated by using a well-devised sampling method. Various sampling methods have been proposed by several researchers so far. In particular, Kim (1985) proposed *centered balanced systematic sampling* (CBSS) and *centered modified systematic sampling* (CMSS). These sampling methods proved to be quite efficient as compared with existing methods. Recently, a second type of CBSS was suggested by Kim (1997).

In the present paper, some modification will be made with CMSS when n (the sample size) is an odd number and k (the reciprocal of the sampling fraction) is an even number. The resultant method will be compared with the original method and other methods. In comparing various methods we will use the expected mean square error criterion based on Cochran's (1946) infinite superpopulation model.

1) Associate Professor, Division of Mathematical Science, Wonkwang University, Iksan, 570-749, Korea.

2. CMSS1 and CMSS2

It is assumed that we have a population of size $N=kn$. The units of this population are denoted by $U_1, U_2, \cdots, U_N$. We wish to select a sample of size n from this population.

2.1 CMSS1

CMSS proposed by Kim (1985), which will be called *CMSS1* from now on, is briefly described.

If $k=N/n$ is an odd number, CMSS1 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) CMSS1 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 S'_1 or S'_2 is selected with respective probability 1/2. Here, the clusters S'_1 and S'_2 are as follows :

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

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

for n even, and

$$S'_1=\{U_{(j-1/2)k}:\ j=1,2,\cdots,(n+1)/2\}\cup\{U_{1+(j-1/2)k}:\ j=(n+3)/2,(n+5)/2,\cdots,n\}$$

$$S'_2=\{U_{1+(j-1/2)k}:\ j=1,2,\cdots,(n+1)/2\}\cup\{U_{(j-1/2)k}:\ j=(n+3)/2,(n+5)/2,\cdots,n\}$$

for n odd.

For example, if $N=24$, $n=6$, $k=4$, then $S'_1=\{U_2, U_6, U_{10}, U_{15}, U_{19}, U_{23}\}$ and $S'_2=\{U_3, U_7, U_{11}, U_{14}, U_{18}, U_{22}\}$, and if $N=20$, $n=5$, $k=4$, then $S'_1=\{U_2, U_6, U_{10}, U_{15}, U_{19}\}$ and $S'_2=\{U_3, U_7, U_{11}, U_{14}, U_{18}\}$. It is to be noted that CMSS1 is obtained by combining the ideas of CSS and modified systematic sampling (MSS), which was proposed by Singh et al. (1968).

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 S'_i be denoted by y'_{ij} ($i=1,2$; $j=1,2,\cdots,n$), and let the

mean value for the units in S'_i be denoted by $\bar{y}'_i$ $(i=1,2)$. For example, if $N=20$, $n=5$, $k=4$, then $y'_{12}=y_6$, $y'_{24}=y_{14}$, $\bar{y}'_1=\sum_{j=1}^{5} y'_{1j}/5 = (y_2+y_6+y_{10}+y_{15}+y_{19})/5$, etc.

The population mean $\bar{Y}=\sum_{i=1}^{N} y_i/N$ is estimated by $\bar{y}'_i$ if the cluster S'_i is selected $(i=1,2)$. That is, if we let $\bar{y}_{cm1}$ denote the estimator of $\bar{Y}$ by CMSS1, then $\bar{y}_{cm1}$ has the probability distribution

$$P(\bar{y}_{cm1}=\bar{y}'_1)=P(\bar{y}_{cm1}=\bar{y}'_2)=\frac{1}{2}.$$

The mean square error of $\bar{y}_{cm1}$ is

$$MSE(\bar{y}_{cm1})=\frac{1}{2}\{(\bar{y}'_1-\bar{Y})^2+(\bar{y}'_2-\bar{Y})^2\}.$$

2.2 CMSS2

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, CMSS1 selects either $S'_1=\{U_2, U_6, U_{10}, U_{15}, U_{19}\}$ or $S'_2=\{U_3, U_7, U_{11}, U_{14}, U_{18}\}$ with respective probability 1/2. We notice that the sums of the numbers assigned to the units in S'_1 and S'_2 are, respectively, 52 and 53, showing a difference of 1. Such a difference is unavoidable in the case when n is odd. Suppose now that U_{10} in S'_1 is replaced by U_{11}, and instead U_{11} in S'_2 is replaced by U_{10}. Let us denote the resultant clusters as S''_1 and S''_2, that is, $S''_1=\{U_2, U_6, U_{11}, U_{15}, U_{19}\}$ and $S''_2=\{U_3, U_7, U_{10}, U_{14}, U_{18}\}$. The numbers of the units in S''_1 and S''_2 now sum to 53 and 52, respectively, giving difference of 1, which is the same as before.

Motivated by the above reasoning, we can introduce the following method, which we expect to be efficient, on the average, in the same degree as CMSS1. Let us define two clusters S''_1 and S''_2 as follows :

$$S''_1=(S'_1-\{U_{N/2}\})\cup\{U_{1+N/2}\}$$
$$S''_2=(S'_2-\{U_{1+N/2}\})\cup\{U_{N/2}\}$$

We suggest a sampling method such that either S''_1 or S''_2 is selected with respective probability 1/2. From now on, this sampling method will be called *CMSS2*.

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 S''_i and the mean value for the units in S''_i. The population mean $\bar{Y}$ is estimated by $\bar{y}''_1$ or $\bar{y}''_2$ according as S''_1 or S''_2 is selected. If we let $\bar{y}_{cm2}$ denote the estimator of $\bar{Y}$ by CMSS2, then $\bar{y}_{cm2}$ has the probability distribution

$$P(\bar{y}_{cm2}=\bar{y}''_1)=P(\bar{y}_{cm2}=\bar{y}''_2)=\frac{1}{2}.$$

$\bar{y}_{cm2}$ is generally a biased estimator for $\bar{Y}$ and has bias

$$\text{Bias}\ (\bar{y}_{cm2})=\frac{1}{2}(\bar{y}''_1+\bar{y}''_2)-\bar{Y}.$$

It is easily checked that $\bar{y}_{cm2}$ has mean square error

$$MSE(\bar{y}_{cm2})=\frac{1}{2}\{(\bar{y}''_1-\bar{Y})^2+(\bar{y}''_2-\bar{Y})^2\}.$$

3. Expected mean square error for CMSS2

In this section, we derive the expected mean square error of $\bar{y}_{cm2}$ on the theoretical basis of Cochran's (1946) infinite superpopulation model.

3.1 General case

We regard the finite population as a sample drawn 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 $\varepsilon(e_i)=0$, $\varepsilon(e_i^2)=\sigma^2$, and $\varepsilon(e_i e_j)=0$ $(i\neq j)$. The operator ε 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 S''_i, e''_{ij} denotes the random error for the jth unit in S''_i, and so on.

The following theorem is very important in evaluating the efficiency of CMSS2.

The proof of this theorem is omitted, because its method is quite similar to that used in the theorem in Kim (1997, p.746), where CBSS was considered.

Theorem 1. Assuming the model (3.1), the expected mean square error of $\bar{y}_{cm2}$ for k even and n odd is

$$\varepsilon MSE(\bar{y}_{cm2}) = \frac{1}{2}\{(\bar{\mu}''_1 - \bar{\mu})^2 + (\bar{\mu}''_2 - \bar{\mu})^2\} + \frac{\sigma^2}{n}\frac{N-n}{N}. \tag{3.2}$$

This theorem is of a similar type to Theorem 8.5 of Cochran (1977, p.213), which is about the expected variance of the estimator by ordinary systematic sampling in the case when $\mu_i = \mu\ (i = 1, 2, \cdots, N)$, that is, when there is no trend.

3.2 Population with a linear trend

Now, let us consider the case when the population has a linear trend. For such a population, μ_i, the expected value of y_i over the infinite superpopulation, is expressed as $\mu_i = a + bi$, where a and b are constants with $b \neq 0$. In other words, the assumed model is

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

In this case, we obtain the following theorem :

Theorem 2. For a population characterized by (3.3), the expected mean square error of $\bar{y}_{cm2}$ for k even and n odd is

$$\varepsilon MSE(\bar{y}_{cm2}) = \frac{b^2}{4n^2} + \frac{\sigma^2}{n}\frac{N-n}{N}. \tag{3.4}$$

This theorem also can be proved by a method similar to that used in Kim (1997, p.752), using the following formulas:

$$\bar{\mu} = a + \left(\frac{b}{2}\right)(N+1), \tag{3.5}$$

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

$$= a + \left(\frac{b}{2}\right)(N+1) + \frac{b}{2n}, \tag{3.6}$$

$$\bar{\mu}''_2 = a + \left(\frac{b}{2}\right)(N+1) - \frac{b}{2n}, \tag{3.7}$$

where we used $\sum_{i=1}^{N} i = N(N+1)/2$, $\sum_{j=1}^{(n-1)/2} j = (n+1)(n-1)/8$ and $\sum_{j=(n+1)/2}^{n} j = (n+1)(3n+1)/8$.

From the above theorem, we can see that $\varepsilon MSE(\bar{y}_{cm2})$ does not depend on the value of a, the intercept of the linear trend, whereas the value of b, the slope, has a crucial effect on $\varepsilon MSE(\bar{y}_{cm2})$.

4. Comparison of efficiency with other methods

In this section, the efficiency of CMSS2 is compared with that of other methods.

First, let us compare the expected mean square error of $\bar{y}_{cm2}$ with that of $\bar{y}_{cm1}$ from CMSS1. It was obtained in Kim (1985) that

$$\varepsilon MSE(\bar{y}_{cm1}) = \frac{b^2}{4n^2} + \frac{\sigma^2}{n}\frac{N-n}{N} \quad (k : \text{even}, \ n : \text{odd}), \tag{4.1}$$

which is equal to (3.4). Consequently, we can state that $\bar{y}_{cm1}$ and $\bar{y}_{cm2}$ are equally efficient from the viewpoint of the expected mean square error criterion.

Now let us consider simple random sampling (SRS), stratified random sampling (StRS), ordinary systematic sampling (OSS), modified systematic sampling (MSS), centered systematic sampling (CSS), balanced systematic sampling (BSS) proposed by Sethi (1965) and named by Murthy (1967), and two types of centered balanced systematic sampling (CBSS) proposed by Kim (1985, 1997). StRS is such that the jth stratum ($j=1, 2, \cdots, n$) consists of units $U_{1+(j-1)k}$, $U_{2+(j-1)k}$, $\cdots$, U_{jk}. From each stratum one unit is selected at random. Since all strata are of equal size and one unit is selected from each stratum, the estimator $\bar{y}_{st}$ of $\bar{Y}$ simplifies to the sample mean. 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 (3.3), the following were obtained in Kim (1985, 1997):

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

$$\varepsilon MSE(\bar{y}_{st}) = \left(\frac{b^2}{12n}\right)(k+1)(k-1) + \frac{\sigma^2}{n}\frac{N-n}{n} \tag{4.3}$$

$$\varepsilon MSE(\bar{y}_{oss}) = \left(\frac{b^2}{12}\right)(k+1)(k-1) + \frac{\sigma^2}{n}\frac{N-n}{N} \tag{4.4}$$

$$\varepsilon MSE(\bar{y}_{mss}) = \varepsilon MSE(\bar{y}_{bss}) = \left(\frac{b^2}{12n^2}\right)(k+1)(k-1) + \frac{\sigma^2}{n}\frac{N-n}{N} \ (n : \text{odd}) \tag{4.5}$$

$$\varepsilon MSE(\bar{y}_{css}) = \frac{b^2}{4} + \frac{\sigma^2}{n}\frac{N-n}{N} \quad (k : \text{even}) \tag{4.6}$$

$$\varepsilon MSE(\bar{y}_{cb1}) = \varepsilon MSE(\bar{y}_{cb2}) = \frac{b^2}{4n^2} + \frac{\sigma^2}{n}\frac{N-n}{N} \quad (k : \text{even}, n : \text{odd}) \tag{4.7}$$

Here $\bar{y}_{srs}$, $\bar{y}_{st}$, $\bar{y}_{oss}$, $\bar{y}_{mss}$, $\bar{y}_{css}$, $\bar{y}_{bss}$, $\bar{y}_{cb1}$ and $\bar{y}_{cb2}$ denote the sample mean, which is used as the estimator of the population mean, obtained from SRS, StRS, OSS, MSS, CSS, BSS, CBSS1 and CBSS2, respectively.

On the basis of formulas (3.4) and (4.1) through (4.7), the methods under consideration can be arranged according to the magnitude of the expected mean square error as follows. For the sake of simplicity, $\varepsilon MSE(\bar{y}_{cm2})$ is denoted as " $cm2$ ", $\varepsilon MSE(\bar{y}_{oss})$ as " oss ", and so on. Thus, for example, " $oss > cm2$" means that CMSS2 is more efficient than OSS. We only consider the case of $n = 3, 5, 7, \cdots$ since the case of $n = 1$ does not have practical meaning. Note that this ordering is true regardless of the value of b, the slope of the linear trend.

Theorem 3. For a population with a linear trend of the form (3.3), the following holds:

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

$$srs > oss = css > st > bss = mss = cb1 = cb2 = cm1 = cm2.$$

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

$$srs > oss > st > bss = mss \ge css > cb1 = cb2 = cm1 = cm2.$$

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

$$srs > oss > st > css > bss = mss > cb1 = cb2 = cm1 = cm2.$$

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

$$srs > oss > css \ge st > bss = mss > cb1 = cb2 = cm1 = cm2.$$

As we see from the above theorem, CMSS1 and CMSS2, together with CBSS1 and CBSS2, are quite efficient as compared with other methods in each case.

Example. The following data are for a small artificial population that exhibits a steady decreasing trend. We have N=40, k=8 and n=5.

93	90	91	88	82	85	79	78
76	80	80	78	75	73	68	69
64	62	60	59	57	55	52	53
45	47	43	37	35	39	40	34
32	31	28	29	22	19	17	17

The mean and the variance of this population are $\bar{Y}=56.55$ and $\sigma_y^2=523.148$, respectively.

The possible samples and MSEs of the estimators of $\bar{Y}$ by various sampling methods are given in Table 1. For example, $MSE(\bar{y}_{cm2})$ is computed as follows :

$$\begin{aligned} MSE(\bar{y}_{cm2}) &= \frac{1}{2}\{(\bar{y}''_1-\bar{Y})^2+(\bar{y}''_2-\bar{Y})^2\} \\ &= \frac{1}{2}\{(56.0-56.55)^2+(56.4-56.55)^2\} \\ &= 0.163 \end{aligned}$$

As we see in Table 1, CMSS2, together with CMSS1, is the most efficient for this population among the ten methods considered.

5. Concluding remarks

Several sampling methods have been introduced so far for the purpose of estimating the mean of a population which has a linear trend. Among them, CMSS1 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 CMSS was suggested and named CMSS2. It was shown that CMSS2, together with CMSS1, CBSS1 and CBSS2, is quite efficient as compared with other sampling methods.

A drawback of CMSS1 and CMSS2 is that we have to suffer some loss of information and reduction of range, because the maximum and minimum values of the population hardly have chance to be included in the sample. For instance, if we apply these methods in investigating sales of enterprises, then the biggest enterprises are likely not to be investigated. This is common to the controlled selection methods such as CSS, CBSS and CMSS. In fact, in the above instance, it would be recommendable to use other sampling method such as, for example, stratified sampling where strata are made according to the size of the enterprises(big, medium and small).

As we saw in Section 4, CMSS2 has the mean square error which is, on the averge, equal to that of CMSS1. This enables CMSS2 to be also a possible choice at the stage of sampling. Like other types of systematic sampling, CMSS2 can be easily applied to practical situations because its sampling procedure is simple.

References

[1] Bellhouse, D. R. and Rao, J. N. K. (1975), "Systematic sampling in the presence of a trend," *Biometrika*, Vol. 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*, Vol. 17, 164-177.

[3] 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.

[4] Kim, H. J. (1997), "A second type of centered balanced systematic sampling method," *The Korean Communications in Statistics*, Vol. 4, 743-752.

[5] Madow, W. G. (1953), "On the theory of systematic sampling, Ⅲ. Comparison of centered and random start systematic sampling," *Annals of Mathematical Statistics*, Vol. 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*, Vol. B27, 315-320.

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

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

Sampling method	Possible samples	MSE
SRS	$\binom{40}{5}$ = 658,008 kinds	93.898
StRS	8^5 = 32,768 kinds	4.669
OSS	{93, 76, 64, 45, 32} {90, 80, 62, 47, 31} {91, 80, 60, 43, 28} {88, 78, 59, 37, 29} {82, 75, 57, 35, 22} {85, 73, 55, 39, 19} {79, 68, 52, 40, 17} {78, 69, 53, 34, 17}	19.618
BSS	{93, 69, 64, 34, 32} {90, 68, 62, 40, 31} {91, 73, 60, 39, 28} {88, 75, 59, 35, 29} {82, 78, 57, 37, 22} {85, 80, 55, 43, 19} {79, 80, 52, 47, 17} {78, 76, 53, 45, 17}	2.638
CSS	{88, 78, 59, 37, 29} {82, 75, 57, 35, 22}	4.123
MSS	{93, 76, 64, 34, 17} {90, 80, 62, 40, 17} {91, 80, 60, 39, 19} {88, 78, 59, 35, 22} {82, 75, 57, 37, 29} {85, 73, 55, 43, 28} {79, 68, 52, 47, 31} {78, 69, 53, 45, 32}	0.778
CBSS1	{88, 75, 59, 35, 29} {82, 78, 57, 37, 22}	1.123
CBSS2	{88, 75, 59, 35, 22} {82, 78, 57, 37, 29}	0.283
CMSS1	{88, 78, 59, 35, 22} {82, 75, 57, 37, 29}	0.163
CMSS2	{88, 78, 57, 35, 22} {82, 75, 59, 37, 29}	0.163