응용통계연구 제 17권 1호, 2004년, pp. 105-117

# 변형된 계통추출과 최소제곱법을 이용한 모평균 추정 *

김혁주 1)

요 약

본 논문에서는 선형추세를 갖는 모집단의 평균을 추정하기 위한 새로운 방법을 제시하였다. 이 방법은 변형계통추출에 의하여 표본을 뽑은 뒤 표본의 단순평균이 아니라 조정된 추정량을 사용하여 모평균을 추정하는 방법이다. 조정된 추정량을 정하는 데에 최소제곱법을 사용하였다. 제시된 방법은 선형추세가 강할수록 효율적이라는 것이 밝혀졌으며, 무한초모집단 모형의 랜덤오차항의 분산인 $\sigma^2$이 매우 크지만 않다면 전통적인 방법들에 비해 상대적으로 효율적인 것으로 나타났다.

주요용어: 선형추세, 모평균, 변형계통추출, 최소제곱법.

## 1. 서론

선형추세(linear trend)를 갖는 유한모집단의 평균을 추정하고자 하는 경우를 생각해 보자. 이 경우 보통의 계통추출(ordinary systematic sampling: OSS)은 단순임의추출(simple random sampling: SRS)보다 훨씬 좋은 결과를 주는 것으로 알려져 있다. 또한 OSS로부터 파생된 여러 표본추출 방법과 추정법들이 많은 연구자들에 의해 연구되어 왔다. Yates(1948)는 선형추세를 갖는 모집단에 대하여 OSS로 표본을 뽑고 끝값수정법(end corrections: EC)으로 모평균을 추정하는 방법을 제안하였다. 선형추세를 갖는 모집단에 대한 표본추출 방법으로 이 밖에도 Madow(1953)의 중심계통추출(centered systematic sampling: CSS), Sethi(1965)와 Murthy(1967)의 균형계통추출(balanced systematic sampling: BSS), Singh 등(1968)의 변형계통추출(modified systematic sampling: MSS), Fountain과 Pathak(1989)의 중심변형추출(centered modified sampling: CMS)과 중심균형추출(centered balanced sampling: CBS) 및 양끝추출(two-end sampling: TES) 등이 있다. 또한 Kim(1985)은 CSS와 MSS의 개념을 결합한 중심변형계통추출(centered modified systematic sampling: CMSS)과, CSS와 BSS의 개념을 결합한 중심균형계통추출(centered balanced systematic sampling: CBSS)을 제안하였다. 그리고 MSS와 보간법의 내삽법(interpolation)을 이용한 방법(MI)과, BSS와 내삽법 및 외삽법(extrapolation)을 이용한 방법(BIE)이 역시 Kim(1998, 1999)에 의하여 소개되었으며, CBSS와 내삽법을 이용한 방법(CBI)이 김혁주와 석은양(2000)에 의하여 제안되었다.

* 본 연구는 한국과학재단의 2001년도 목적기초연구(R05-2001-000-00057-0)지원으로 수행되었습니다.

1) (570-749) 전북 익산시 신용동 344-2, 원광대학교 자연과학대학 수학·정보통계학부 및 기초자연과학연구소, 교수
E-mail: hjkim@wonkwang.ac.kr

모집단에 선형추세가 존재하는 경우는 실제 통계조사에서 종종 접할 수 있다. 예를 들어, 주어진 기간에 대하여 어떤 도시 안에 있는 백화점들의 평균 매출액을 추정하는 경우를 생각해 보자. 이 때 우리는 그 도시 안의 백화점들에 종업원의 수에 따라 증가하거나 감소하는 순서로 번호를 붙임으로써 이 모집단을 직선 모양에 가까운 추세를 갖는 모집단으로 만들 수 있다.

본 연구에서는 선형추세를 갖는 모집단의 평균을 추정하기 위한 새로운 방법이 제시된다. 이 방법은 MSS에 의하여 표본을 뽑은 뒤 회귀분석의 최소제곱법을 도입하여 모평균을 추정하는 것으로서 표본크기 $n$이 3이상의 홀수이고 추출률의 역수인 $k$가 짝수인 경우에 사용하기 위한 것이다.

## 2. 최소제곱법을 이용한 모평균 추정

$N$개의 단위 $U_1, U_2, \cdots, U_N$으로 구성된 모집단을 생각하자. 이 모집단으로부터 크기 $n$인 표본을 뽑으려 한다.

집락 $S_i^{'}$ $(i = 1, 2, \cdots, k)$를 다음과 같이 정의하자.

$n$이 짝수일 때

$S_i^{'} = \{U_{i+(j-1)k} : j = 1, 2, \cdots, n/2\} \bigcup \{U_{N+1-i-(j-1)k} : j = 1, 2, \cdots, n/2\}$ $(i = 1, 2, \cdots, k)$

$n$이 홀수일 때

$S_i^{'} = \{U_{i+(j-1)k} : j = 1, 2, \cdots, (n+1)/2\} \bigcup \{U_{N+1-i-(j-1)k} : j = 1, 2, \cdots, (n-1)/2\}$ $(i = 1, 2, \cdots, k)$

이제 본 논문에서 사용될 기호들을 정의한다.

$y_i$ : 모집단 안의 $i$번째 단위 $U_i$가 가지고 있는 특성값 $(i = 1, 2, \cdots, N)$

$\overline{Y} = \dfrac{1}{N}\sum_{i=1}^{N} y_i$ : 추정하고자 하는 모집단 평균

$y_{ij}^{'}$ : $S_i^{'}$안의 $j$번째 단위의 특성값 $(i = 1, 2, \cdots, k\ ;\ j = 1, 2, \cdots, n\ )$

즉, $n$이 짝수일 때

$$y_{ij}^{'} = y_{i+(j-1)k} \quad (j = 1, 2, \cdots, n/2)$$
$$y_{ij}^{'} = y_{N+1-i-(n-j)k} = y_{1-i+jk} \quad (j = n/2+1, n/2+2, \cdots, n)$$

$n$이 홀수일 때

$$y_{ij}^{'} = y_{i+(j-1)k} \quad (j = 1, 2, \cdots, (n-1)/2, (n+1)/2)$$
$$y_{ij}^{'} = y_{N+1-i-(n-j)k} = y_{1-i-jk} \quad (j = (n+3)/2, (n+5)/2, \cdots, n)$$

$\overline{y}_i^{'} = \dfrac{1}{n}\sum_{j=1}^{n} y_{ij}^{'}$ : $S_i^{'}$안의 단위들의 특성값의 평균 $(i = 1, 2, \cdots, k)$

이제 모평균 $\overline{Y}$를 추정하는 새로운 방법을 제시하고자 한다. 이 방법은 $k$가 짝수이고 $n$이 3이상의 홀수인 경우에 사용할 수 있는 방법으로서 Singh 등(1968)의 MSS를 사용하여 표본을 추출한 뒤 표본평균 $\overline{y}_{MSS}$대신 수정된 추정량을 써서 $\overline{Y}$를 추정하는 것이다.

$N=20,\ n=5,\ k=4$인 경우를 예로 들어 새로운 추정방법을 설명하기로 한다. 우선 $S_1', S_2', S_3', S_4'$중 하나의 집락을 뽑는다. 이 추출방법이 MSS이다. $S_1'=\{U_1,U_5,U_9,U_{16},U_{20}\}$, $S_2'=\{U_2,U_6,U_{10},U_{15},U_{19}\}$, $S_3'=\{U_3,U_7,U_{11},U_{14},U_{18}\}$, $S_4'=\{U_4,U_8,U_{12},U_{13},U_{17}\}$이므로, 네 집락 안에 있는 단위들의 번호를 합하면 각각 $51,52,53,54$이다. 모집단에 선형추세가 존재하는 경우에는 이렇게 미세한 차이도 제거해 주는 것이 좋을 것이다. 왜냐하면, 증가하는 선형추세의 경우에는 앞번호를 가진 단위보다 뒷번호를 가진 단위가 더 큰 값을 갖는 경향이 있으므로 단위들의 번호의 합이 큰 집락이 작은 집락보다 큰 평균값을 갖기 쉽고, 감소하는 선형추세의 경우에는 그 반대일 것이기 때문이다. 따라서 단순한 표본평균 $\overline{y}_i'$를 쓰는 것보다 수정된 추정량을 써서 $\overline{Y}$를 추정하는 것이 바람직할 것이다. $S_1'$이 뽑힌 경우에는 $y_9$대신 $y_{10.5}$를 쓰고, $S_2'$가 뽑힌 경우에는 $y_{10}$대신 $y_{10.5}$를 쓰며 $S_3'$, $S_4'$가 뽑힌 경우에는 각각 $y_{11}$, $y_{12}$대신 $y_{10.5}$를 쓰면 균형이 이루어진다. 물론 여기서 $y_{10.5}$는 실제로는 존재하지 않는 가상적인 값이다.

$S_1'$이 뽑힌 경우 $y_{10.5}$는 다음과 같은 방법에 의하여 추정할 수 있다. 단위들의 번호인 $1,5,9,16,20$을 $x$변수(설명변수)값들$(x_{11},x_{12},x_{13},x_{14},x_{15})$로 보고 $y_1,y_5,y_9,y_{16},y_{20}$(즉, $y_{11}',y_{12}',y_{13}',y_{14}',y_{15}'$)을 $y$변수(반응변수)값들로 보아 회귀분석의 최소제곱법을 이용하여 선형추세의 직선을 적합하면 적합된 회귀직선의 방정식은

$$\hat{y}_1=\hat{\alpha}_1+\hat{\beta}_1x_1 \tag{2.1}$$

이다. 단, 여기서 $\hat{\beta}_1=S_{(xy)1}/S_{(xx)1}=\sum_{j=1}^{5}(x_{1j}-\overline{x}_1)(y_{1j}'-\overline{y}_1')/\sum_{j=1}^{5}(x_{1j}-\overline{x}_1)^2$이고, $\hat{\alpha}_1=\overline{y}_1'-\hat{\beta}_1\overline{x}_1$이며, $\overline{x}_1=(1+5+9+16+20)/5=10.2$이고, $\overline{y}_1'$은 앞에서 정의된 바와 같다. 첨자 1은 첫 번째 집락$(S_1')$을 나타내는 것이다. 식 (2.1)은

$$\hat{y}_1=\overline{y}_1'+\hat{\beta}_1(x_1-\overline{x}_1) \tag{2.2}$$

로 나타낼 수 있고

$$\hat{\beta}_1=\sum_{j=1}^{5}a_{1j}y_{1j}'\quad (\text{단, } a_{1j}=(x_{1j}-\overline{x}_1)/S_{(xx)1},\ j=1,2,\cdots,n)$$

로 쓸 수 있으므로 $y_{10.5}$는 다음의 값으로 추정된다.

$$\begin{aligned}\hat{y}_{10.5}^{(1)} &= \overline{y}_1'+(10.5-10.2)\left(-\frac{9.2}{242.8}y_{11}'-\frac{5.2}{242.8}y_{12}'-\frac{1.2}{242.8}y_{13}'+\frac{5.8}{242.8}y_{14}'+\frac{9.8}{242.8}y_{15}'\right)\\ &= 0.1886y_{11}'+0.1936y_{12}'+0.1985y_{13}'+0.2072y_{14}'+0.2121y_{15}'\end{aligned} \tag{2.3}$$

따라서, $S_1'$이 뽑힌 경우 $y_{13}'$ 즉 $y_9$대신 이 값을 사용하여 모평균 $\overline{Y}$를 다음의 값으로 추정한다.

$$\overline{y}_1'^{*}=\frac{1}{5}(y_{11}'+y_{12}'+\hat{y}_{10.5}^{(1)}+y_{14}'+y_{15}')$$

$$= 0.2377y_{11}' + 0.2387y_{12}' + 0.0397y_{13}' + 0.2414y_{14}' + 0.2424y_{15}' \quad (2.4)$$

$S_2'$, $S_3'$ 또는 $S_4'$가 뽑힌 경우에도 위와 같은 방법으로 $\overline{Y}$를 추정할 수 있다.

이상의 내용을 일반화하면 다음과 같다. $k$가 짝수이고 $n$이 3이상의 홀수인 경우 MSS에 의하여 하나의 집락을 뽑는다. 즉, 각각 $1/k$의 확률로 $k$개의 집락 $S_1', S_2', \cdots, S_k'$중 하나를 뽑는다. 뽑힌 집락을 $S_i'$라 하자. $x_{i1}, x_{i2}, \cdots, x_{in}$을 설명변수의 값들로 보고 $y_{i1}', y_{i2}', \cdots, y_{in}'$을 반응변수의 값들로 보아 회귀분석의 최소제곱법을 적용하여 $y_{(N+1)/2}$의 값을 추정한다.

$$x_{ij} = \begin{cases} i + (j-1)k \quad (j = 1, 2, \cdots, (n+1)/2) \\ N + 1 - i - (n-j)k = 1 - i + jk \quad (j = (n+3)/2, (n+5)/2, \cdots, n) \end{cases} \quad (2.5)$$

이므로,

$$\overline{x}_i = \frac{1}{n}\sum_{j=1}^{n} x_{ij} = \frac{N+1}{2} + \frac{1}{n}\left(i - \frac{k+1}{2}\right) \quad (2.6)$$

이고,

$$\begin{aligned} S_{(xx)i} &= \sum_{j=1}^{n} x_{ij}^2 - n(\overline{x}_i)^2 \\ &= \left(n - \frac{1}{n}\right)i^2 - \left(n - \frac{1}{n}\right)\left\{k\left(\frac{n}{2}+1\right)+1\right\}i \\ &\quad + \frac{1}{12}\left(n - \frac{1}{n}\right)\left\{(n^2+3n+3)k^2 + 3(n+2)k + 3\right\} \end{aligned} \quad (2.7)$$

이다. 회귀직선의 기울기의 추정값은 $\hat{\beta}_i = \sum_{j=1}^{n} a_{ij}y_{ij}'$ (단, $a_{ij} = (x_{ij} - \overline{x}_i)/S_{(xx)i}$, $i = 1, 2, \cdots, n$)이며, $y_{(N+1)/2}$는

$$\begin{aligned} \hat{y}_{(N+1)/2}^{(i)} &= \overline{y}_i' + \hat{\beta}_i\left(\frac{N+1}{2} - \overline{x}_i\right) \\ &= \left\{\frac{1}{n} - \frac{a_{i1}}{n}\left(i - \frac{k+1}{2}\right)\right\}y_{i1}' + \cdots + \left\{\frac{1}{n} - \frac{a_{in}}{n}\left(i - \frac{k+1}{2}\right)\right\}y_{in}' \end{aligned} \quad (2.8)$$

에 의하여 추정된다.

따라서, 모평균 $\overline{Y}$는 다음의 값으로 추정할 수 있다.

$$\begin{aligned} \overline{y}_i'^{*} &= \frac{1}{n}(y_{i1}' + \cdots + y_{i,(n-1)/2}' + \hat{y}_{(N+1)/2}^{(i)} + y_{i,(n+3)/2}' + \cdots + y_{in}') \\ &= \sum_{j=1}^{n} c_{ij}y_{ij}' \end{aligned} \quad (2.9)$$

단, 여기서

$$c_{ij} = \begin{cases} \frac{1}{n}\left\{1 + \frac{1}{n} - \frac{a_{ij}}{n}\left(i - \frac{k+1}{2}\right)\right\} & (j = 1, 2, \cdots, (n-1)/2, (n+3)/2, \cdots, n) \\ \frac{1}{n}\left\{\frac{1}{n} - \frac{a_{i,(n+1)/2}}{n}\left(i - \frac{k+1}{2}\right)\right\} & (j = (n+1)/2) \end{cases} \tag{2.10}$$

이다.

MSS의 M과 최소제곱(Least Squares)법의 L, S를 따서 이 방법을 MLS로 표시하자. MLS에 의한 $\overline{Y}$의 추정량을 $\overline{y}_{MLS}$로 나타내면 $\overline{y}_{MLS}$는 다음과 같은 편향과 평균제곱오차를 갖는다.

$$Bias(\overline{y}_{MLS}) = \frac{1}{k}\sum_{i=1}^{k} \overline{y}_i^{'*} - \overline{Y} \tag{2.11}$$

$$MSE(\overline{y}_{MLS}) = \frac{1}{k}\sum_{i=1}^{k} (\overline{y}_i^{'*} - \overline{Y})^2 \tag{2.12}$$

## 3. 무한초모집단 모형 하에서의 평균제곱오차의 기대값

추세를 갖는 모집단에 대한 표본추출 및 추정 방법을 논할 때 이론적 도구로 유용하게 쓰이는 것이 Cochran(1946)의 무한초모집단 모형(infinite superpopulation model)이다. 무한초모집단 모형이란, 주어진 유한모집단을 무한초모집단으로부터 뽑힌 하나의 표본으로 간주하는 것으로서 다음과 같이 세워진다.

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

여기서 $a + bi$는 선형추세를 나타내며, $e_i$는 오차항으로서 $E(e_i) = 0$, $E(e_i^2) = \sigma^2$, $E = (e_i e_j) = 0$ $(i \neq j$ 일 때)이다. $E$는 무한초모집단에 걸친 기대값을 나타낸다.

정리 3.1 식 (3.1)을 가정할 때, 여러 방법들에 의한 모평균 $\overline{Y}$의 추정량의 평균제곱오차의 기대값들은 다음과 같다. 단, 여기서 $A = \sigma^2(1/n - 1/N)$이다.

1) OSS : $$EMSE(\overline{y}_{OSS}) = \frac{b^2(k^2-1)}{12} + A \tag{3.2}$$

2) EC : $$EMSE(\overline{y}_{EC}) = A + \frac{\sigma^2(k^2-1)}{6k^2(n-1)^2} \tag{3.3}$$

3) MSS : $$EMSE(\overline{y}_{MSS}) = \begin{cases} A & (n : \text{짝수}) \\ \frac{b^2(k^2-1)}{12n^2} + A & (n : \text{홀수}) \end{cases} \tag{3.4}$$

$$4)\ \text{BSS} : EMSE(\overline{y}_{BSS}) = \begin{cases} A & (n : \text{짝수}) \\ \dfrac{b^2(k^2-1)}{12n^2} + A & (n : \text{홀수}) \end{cases} \tag{3.5}$$

$$5)\ \text{CSS} : EMSE(\overline{y}_{CSS}) = \begin{cases} \dfrac{b^2}{4} + A & (k : \text{짝수}) \\ A & (k : \text{홀수}) \end{cases} \tag{3.6}$$

$$6)\ \text{CMSS} : EMSE(\overline{y}_{CMSS}) = \begin{cases} A & (k : \text{짝수},\ n : \text{짝수}) \\ \dfrac{b^2}{4n^2} + A & (k : \text{짝수},\ n : \text{홀수}\ ) \end{cases} \tag{3.7}$$

$$7)\ \text{CBSS} : EMSE(\overline{y}_{CBSS}) = \begin{cases} A & (k : \text{짝수},\ n : \text{짝수}) \\ \dfrac{b^2}{4n^2} + A & (k : \text{짝수},\ n : \text{홀수}\ ) \end{cases} \tag{3.8}$$

8) CMS, CBS, TES :

$$EMSE(\overline{y}_{CMS}) = EMSE(\overline{y}_{CBS}) = EMSE(\overline{y}_{TES})$$

$$= \begin{cases} A & (n : \text{짝수}) \\ A & (k : \text{홀수},\ n : \text{홀수}) \\ \dfrac{b^2}{4n^2} + A & (k : \text{짝수},\ n : \text{홀수}\ ) \end{cases} \tag{3.9}$$

$$9)\ \text{MI} : EMSE(\overline{y}_{MI}) = A + \frac{\sigma^2}{12n^2}\left(4 - 12A_k + 6kB_k - \frac{1}{k^2}\right)\ (k : \text{짝수},\ n : 3\text{이상의 홀수}) \tag{3.10}$$

단, $A_k = \dfrac{1}{2}\left\{\psi\left(k+\dfrac{1}{2}\right) - \psi\left(\dfrac{k+1}{2}\right)\right\}$

$B_k = -\dfrac{1}{4}\left\{\psi^{(1)}\left(k+\dfrac{1}{2}\right) - \psi^{(1)}\left(\dfrac{k+1}{2}\right)\right\}$

$\psi(x) = \dfrac{d}{dx}\ln\Gamma(x)\ \ (x>0)$ : polygamma 함수

$\Gamma(x) = \displaystyle\int_0^\infty t^{x-1}e^{-t}dt\ \ (x>0)$ : gamma 함수

$\psi^{(1)}(x) = \dfrac{d}{dx}\psi(x)$

$$10)\ \text{BIE} : EMSE(\overline{y}_{BIE}) = A + \frac{\sigma^2}{2n^2}(1 - \gamma - 2\ln 2 + C_k)\ (k : \text{짝수},\ n : 3\text{이상의 홀수}) \tag{3.11}$$

단, $\gamma = 0.577215...$ : Euler 상수

$C_k = \dfrac{k}{8}\left\{\pi^2 - 2\psi^{(1)}\left(k+\dfrac{1}{2}\right)\right\} - \psi\left(k+\dfrac{1}{2}\right)$

$$11)\ \text{CBI} : EMSE(\overline{y}_{CBI}) = A + \frac{\sigma^2}{2n^2(k+1)^2}\ \ (k : \text{짝수},\ n : 5\text{이상의 홀수}) \tag{3.12}$$

$$12)\ \text{MLS} : EMSE(\overline{y}_{MLS}) = A + \sigma^2 T(k,n)\ (k : \text{짝수},\ n : 3\text{이상의 홀수}) \tag{3.13}$$

$$\text{단, } T(k,n) = \frac{n-1}{n^3} + \frac{6n-3}{kn^3(n^2-1)} S(k,n)$$

$$S(k,n) = \sum_{i=1}^{k} \frac{(2i-k-1)^2}{(2i-k-1)(6i-3kn-3k-3)+k^2n^2}$$

위의 정리에서 $EMSE(\cdot) = A$인 방법이 가장 효율적인 방법이다. 참고로 $k$와 $n$이 홀수인지 짝수인지에 따라 $EMSE(\cdot) = A$인 방법을 표 3.1에 정리하였다.

표 3.1: $EMSE(\cdot) = A$인 방법들

| $n$ | $k$ | |
|---|---|---|
| | 홀수 | 짝수 |
| 홀수 | CSS, CMS, CBS, TES | |
| 짝수 | CSS, MSS, BSS, CMS, CBS, TES | MSS, BSS, CMSS, CBSS, CMS, CBS, TES |

## 4. 기존의 방법들과의 비교

모집단에 선형추세가 존재하는 경우 식 (3.2)부터 식 (3.13)까지의 식들을 사용하여 여러 방법들의 효율성을 비교할 수 있다. 본 연구에서 제시된 방법이 $k$가 짝수이고 표본크기 $n$이 3이상의 홀수인 경우에 사용하기 위한 것이므로, 이 절에서 고려하는 경우는 모두 이러한 경우이다.

### 4.1. 표본의 단순평균으로 모평균을 추정하는 방법들과의 비교

표본의 단순평균으로 모평균 $\overline{Y}$를 추정하는 방법으로 OSS, MSS, BSS, CSS, CMSS, CBSS, CMS, CBS, TES 등이 있다. 제3절의 정리에서 볼 수 있듯이 이 방법들에 의한 추정량들의 기대평균제곱오차 (expected mean square error: EMSE)는 무한초모집단 모형의 오차항의 분산 $\sigma^2$과 선형추세의 기울기 $b$의 영향을 받는다. $k$와 $n$이 주어진 상태에서 생각할 때, 기울기가 일정하면 $\sigma^2$이 클수록 EMSE가 크며, $\sigma^2$이 일정하면 EMSE는 $b = 0$일 때 최소이고 $b$의 절대값이 클수록 증가한다.

본 연구에서 제시된 방법인 MLS를 이 방법들과 비교해 보자. 첫 번째로 식 (3.2)와 식 (3.13)으로부터, MLS가 OSS보다 효율적일 조건, 즉

$$EMSE(\overline{y}_{MLS}) < EMSE(\overline{y}_{OSS}) \tag{4.1}$$

일 필요충분조건은

$$\sigma^2 < \frac{b^2(k^2-1)}{12T(k,n)} \tag{4.2}$$

임을 얻게 된다. 이것은 무한초모집단 모형의 오차항의 분산 $\sigma^2$이 터무니없이 크지만 않으면 성립한다.

두 번째로 MLS와 Singh 등(1968)의 MSS를 비교해 보면, 식 (3.4)와 식 (3.13)으로부터, $\overline{y}_{MLS}$가 $\overline{y}_{MSS}$보다 효율적일 필요충분조건은

$$\sigma^2 < \frac{b^2(k^2-1)}{12n^2T(k,n)} \tag{4.3}$$

이다. 이것은 $\sigma^2$이 작을수록(즉 선형추세가 뚜렷할수록) $\overline{y}_{MLS}$의 효율성이 특히 우수하다는 것을 의미한다.

이러한 방식으로 모든 경우에 대하여 여러 방법들을 비교한 결과를 다음과 같이 정리하였다. 표현을 간결하게 하기 위하여 $EMSE(\overline{y}_{OSS}), EMSE\ (\overline{y}_{MLS})$ 등을 각각 $OSS$, $MLS$ 등으로 나타냈다. 예컨대 "$MLS < OSS$"는 $\overline{y}_{MLS}$가 $\overline{y}_{OSS}$보다 효율적이라는 것을 뜻하는 표현이다. 그리고 $k$가 짝수이고 $n$이 홀수인 모든 경우에 $CMSS = CBSS = CMS = CBS = TES$이므로 간편성을 위하여 $CMSS$ 한 가지만 대표로 표시하였다.

1) $k=2$이고 $n$이 3이상의 홀수인 경우

i ) $\sigma^2 < b^2/\{4n^2T(2,n)\}$이면 $MLS < CMSS = MSS = BSS < CSS = OSS$

ii ) $b^2/\{4n^2T(2,n)\} \le \sigma^2 < b^2/\{4T(2,n)\}$이면
$CMSS = MSS = BSS \le MLS < CSS = OSS$

iii ) $b^2/\{4T(2,n)\} \le \sigma^2$이면 $CMSS = MSS = BSS < CSS = OSS \le MLS$

2) $k$가 4이상의 짝수, $n$이 3이상의 홀수이고 $n < \sqrt{(k^2-1)/3}$인 경우

i ) $\sigma^2 < b^2/\{4n^2T(k,n)\}$이면 $MLS < CMSS < CSS < MSS = BSS < OSS$

ii ) $b^2/\{4n^2T(k,n)\} \le \sigma^2 < b^2/\{4T(k,n)\}$이면
$CMSS \le MLS < CSS < MSS = BSS < OSS$

iii ) $b^2/\{4T(k,n)\} \le \sigma^2 < b^2(k^2-1)/\{12n^2T(k,n)\}$이면
$CMSS < CSS \le MLS < MSS = BSS < OSS$

iv ) $b^2(k^2-1)/\{12n^2T(k,n)\} \le \sigma^2 < b^2(k^2-1)/\{12T(k,n)\}$이면
$CMSS < CSS < MSS = BSS \le MLS < OSS$

v ) $b^2(k^2-1)/\{12T(k,n)\} \le \sigma^2$이면 $CMSS < CSS < MSS = BSS < OSS \le MLS$

3) $k$가 4이상의 짝수, $n$이 3이상의 홀수이고 $n = \sqrt{(k^2-1)/3}$인 경우
(예를 들면 $k=26,\ n=15$)

i ) $\sigma^2 < b^2/\{4n^2T(k,n)\}$이면 $MLS < CMSS < CSS = MSS = BSS < OSS$

ii ) $b^2/\{4n^2T(k,n)\} \le \sigma^2 < b^2/\{4T(k,n)\}$이면
$CMSS \le MLS < CSS = MSS = BSS < OSS$

iii ) $b^2/\{4T(k,n)\} \le \sigma^2 < b^2(k^2-1)/\{12T(k,n)\}$이면
$CMSS < CSS = MSS = BSS \le MLS < OSS$

iv ) $b^2(k^2-1)/\{12T(k,n)\} \le \sigma^2$이면 $CMSS < CSS = MSS = BSS < OSS \le MLS$

4) $k$가 4이상의 짝수, $n$이 3이상의 홀수이고 $n > \sqrt{(k^2-1)/3}$인 경우

i ) $\sigma^2 < b^2/\{4n^2T(k,n)\}$이면 $MLS < CMSS < MSS = BSS < CSS < OSS$
ii ) $b^2/\{4n^2T(k,n)\} \le \sigma^2 < b^2(k^2-1)/\{12n^2T(k,n)\}$이면
$CMSS \le MLS < MSS = BSS < CSS < OSS$
iii ) $b^2(k^2-1)/\{12n^2T(k,n)\} \le \sigma^2 < b^2/\{4T(k,n)\}$이면
$CMSS < MSS = BSS \le MLS < CSS < OSS$
iv ) $b^2/\{4T(k,n)\} \le \sigma^2 < b^2(k^2-1)/\{12T(k,n)\}$이면
$CMSS < MSS = BSS < CSS \le MLS < OSS$
v ) $b^2(k^2-1)/\{12T(k,n)\} \le \sigma^2$이면 $CMSS < MSS = BSS < CSS < OSS \le MLS$

예제 4.1: 모집단의 크기가 $N = 300$이고 표본크기가 $n = 25$이며 따라서 $k = 12$인 경우를 생각해 보자. 선형추세의 기울기는 $b = 0.5$라고 하자. 이 경우 여러 방법의 효율성을 비교해 보면 다음과 같다.

i ) $\sigma^2 < 0.0654$이면 $MLS < CMSS < MSS = BSS < CSS < OSS$
ii ) $0.0654 \le \sigma^2 < 3.1155$이면 $CMSS \le MLS < MSS = BSS < CSS < OSS$
iii ) $3.1155 \le \sigma^2 < 40.8497$이면 $CMSS < MSS = BSS \le MLS < CSS < OSS$
iv ) $40.8497 \le \sigma^2 < 1947.1678$이면 $CMSS < MSS = BSS < CSS \le MLS < OSS$
v ) $1947.1678 \le \sigma^2$이면 $CMSS < MSS = BSS < CSS < OSS \le MLS$

여기서 볼 수 있듯이, $\sigma^2$의 값이 아주 크지만 않다면 MLS는 전통적인 방법들에 비해 효율적인 추정량이 된다. 사실 $\sigma^2$이 너무 큰 값을 갖는 경우는 선형추세의 의미가 거의 없어지기 때문에 논의할 필요가 없게 된다.

### 4.2. 표본의 가중평균으로 모평균을 추정하는 방법들과의 비교

표본값들의 단순평균이 아닌 가중평균으로 모평균을 추정하는 방법으로는 Ya-tes(1948)의 EC, Kim(1998)의 MI, Kim(1999)의 BIE, 그리고 김혁주와 석은양(2000)의 CBI 등이 있다. 본 논문의 MLS도 이에 속한다. 이 방법들의 공통적인 특징은 식 (3.3), (3.10), (3.11), (3.12), (3.13)에서 볼 수 있듯이 EMSE가 $k$, $n$, $\sigma^2$에만 의존하며 기울기 $b$의 값에 무관하다는 것이다. $k$와 $n$이 주어진 상태에서 생각하면, $\sigma^2$이 클수록 EMSE가 증가한다. 표 4.1부터 표 4.3까지는 몇 개의 $k$값과 $n$값에 대하여 다섯 가지 방법에 의한 $EMSE(\cdot)/\sigma^2$의 값들을 구해 놓은 것이다.

표 4.1: $k = 8$인 경우 $EMSE(\cdot)/\sigma^2$의 값들

| $n$ | EC | MI | BIE | CBI | MLS |
|---|---|---|---|---|---|
| 5 | 0.1853 | 0.1794 | 0.3065 | 0.1752 | 0.2071 |
| 25 | 0.0353 | 0.0352 | 0.0403 | 0.0350 | 0.0365 |
| 55 | 0.0160 | 0.0160 | 0.0170 | 0.0159 | 0.0162 |
| 105 | 0.0084 | 0.0083 | 0.0086 | 0.0083 | 0.0084 |

표 4.2: $k = 12$인 경우 $EMSE(\cdot)/\sigma^2$의 값들

| $n$ | EC | MI | BIE | CBI | MLS |
|---|---|---|---|---|---|
| 5 | 0.1937 | 0.1878 | 0.4054 | 0.1835 | 0.2155 |
| 25 | 0.0370 | 0.0368 | 0.0456 | 0.0367 | 0.0382 |
| 55 | 0.0167 | 0.0167 | 0.0185 | 0.0167 | 0.0170 |
| 105 | 0.0088 | 0.0087 | 0.0092 | 0.0087 | 0.0088 |

표 4.3: $k = 20$인 경우 $EMSE(\cdot)/\sigma^2$의 값들

| $n$ | EC | MI | BIE | CBI | MLS |
|---|---|---|---|---|---|
| 5 | 0.2004 | 0.1945 | 0.5993 | 0.1900 | 0.2221 |
| 25 | 0.0383 | 0.0382 | 0.0544 | 0.0380 | 0.0395 |
| 55 | 0.0173 | 0.0173 | 0.0207 | 0.0173 | 0.0176 |
| 105 | 0.0091 | 0.0091 | 0.0100 | 0.0090 | 0.0091 |

위의 표들에서 볼 수 있는 바와 같이 MLS는 항상 BIE보다 효율적이다. 그리고 표본크기 $n$이 작을 때에는 MLS의 EMSE가 EC, MI, CBI의 EMSE보다 근소하게 크나 $n$이 커짐에 따라 EC, MI, CBI의 EMSE와 거의 같아짐을 알 수 있다. 더욱이 MLS는 최소제곱법을 사용하기 때문에 모집단에 존재하는 선형추세의 회귀방정식을 추정할 수 있다는 장점이 있다. 이같은 장점은 EC, MI, BIE와 CBI는 갖지 못한 장점이므로, 특히 표본크기 $n$이 클 때에는 MLS의 가치가 크다고 할 수 있다.

### 4.3. 모의실험을 이용한 설명

예제 4.2: 모형

$$y_i = 5 + 0.8i + e_i \quad (i = 1, 2, \cdots, 36) \tag{4.4}$$

를 설정하자(즉 $a = 5$, $b = 0.8$). 오차항 $e_i$의 값들을 실제로 발생시킴으로써 무한초모집단으로부터 크기 $N = 36$인 모집단을 생성한 다음, 다시 이 모집단으로부터 크기 $n = 9$인 표본을 뽑아 모평균을 추정하는 문제를 생각해 보자. $k = 4$이며, 오차항 $e_i$는 3절에서와 같은 조건을 만족시킨다. $e_i$의 분산 $\sigma^2$의 값은 0.16으로 하였고(즉 4.1절에서 4)의 i )에 해당하는 경우이다) $e_i$의 분포의 형태는 정규분포로 정하였으며, 미니탭(MINITAB)의 RANDOM 명령문을 이용하여 $e_i$의 값들을 발생시켰다. 생성된 모집단은 다음과 같다.

| | | | | | | |
|---|---|---|---|---|---|---|
| 5.8845 | 6.3701 | 7.2784 | 7.9600 | 8.5229 | 9.8625 | 10.9136 |
| 10.9442 | 12.2332 | 12.6402 | 13.3558 | 14.5838 | 15.5889 | 16.1484 |
| 16.2267 | 17.6055 | 18.8280 | 19.7193 | 20.4190 | 20.3251 | 21.2957 |
| 22.3068 | 23.0911 | 23.1101 | 25.2562 | 25.8478 | 26.6090 | 27.5326 |
| 28.0149 | 29.5544 | 29.2740 | 30.3364 | 31.3012 | 32.2578 | 33.2673 |
| 33.6672 | | | | | | |

이 모집단의 평균은 $\overline{Y} = 19.6703$이며, 이 모집단은 증가하는 선형추세를 가지고 있다. 기존의 방법들에 의한 $\overline{Y}$의 추정량들의 평균제곱오차는 아래와 같다.

$MSE(\overline{y}_{OSS})$=0.6195, $MSE(\overline{y}_{EC})$=0.0169, $MSE(\overline{y}_{MSS})$=0.0096,
$MSE(\overline{y}_{BSS})$=0.0190, $MSE(\overline{y}_{CSS})$=0.1048, $MSE(\overline{y}_{CMSS})$=0.0079,
$MSE(\overline{y}_{CBSS})$=0.0223, $MSE(\overline{y}_{CMS})$=0.0037, $MSE(\overline{y}_{CBS})$=0.0088,
$MSE(\overline{y}_{TES})$=0.0144, $MSE(\overline{y}_{MI})$=0.0037, $MSE(\overline{y}_{BIE})$=0.0079,
$MSE(\overline{y}_{CBI})$=0.0123.

한편 MLS를 사용하면, 모평균 $\overline{Y}$를 다음과 같은 네 개의 값 중 하나로 추정하게 된다(확률은 각각 1/4). 이 네 개의 값들은 식 (2.9)와 식 (2.10)에 따라 컴퓨터를 사용하여 구한 것이다.

$$\overline{y}_1^{'*} = 19.6145 \qquad \overline{y}_2^{'*} = 19.6635$$
$$\overline{y}_3^{'*} = 19.7213 \qquad \overline{y}_4^{'*} = 19.6139$$

따라서 MLS에 의한 $\overline{Y}$의 추정량 $\overline{y}_{MLS}$의 평균제곱오차는

$$MSE(\overline{y}_{MLS}) = 0.0022$$

이다. 이 값은 위의 열세 가지를 포함한 열네 가지의 방법에 의한 평균제곱오차 중 가장 작은 값이므로, MLS가 열네 가지의 방법 중 가장 효율적이라는 것을 보여 준다.

## 5. 결론

모수를 추정할 때 결정해야 할 중요한 두 가지는 표본추출 방법과 추정량이다. 선형추세를 갖는 모집단의 평균을 추정하는 경우에도 선형추세라는 특성을 살려서 이 두 가지를 결정할 필요가 있다. 본 논문에서는 이 점을 염두에 두고 MSS(변형계통추출)라는 표본추출 방법과 회귀분석의 최소제곱법을 도입한 추정 방법을 제시하였다.

제시된 추정량을 $\overline{y}_{MLS}$로 표시했으며, Cochran(1946)의 무한초모집단 모형에 근거를 둔 기대평균제곱오차를 기준으로 하여 $\overline{y}_{MLS}$의 효율을 기존의 방법에 의한 추정량들과 비교하였다. 그 결과 $\overline{y}_{MLS}$는 무한초모집단 모형의 오차항의 분산 $\sigma^2$이 작을수록 (즉 선형추세

가 강할수록) 효율적이라는 것이 밝혀졌으며, $\sigma^2$이 아주 크지 않은 값을 갖는 대부분의 현실적인 경우에 전통적인 추정량들에 비해서 효율적인 것으로 나타났다.

## 참고문헌

[1] 김혁주, 석은양 (2000), 선형추세를 갖는 모집단에 대한 효율적인 모평균 추정: 계통추출의 확장, <응용통계연구>, **13**, 457-476.

[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] Fountain, R. L. and Pathak, P. K. (1989), Systematic and nonrandom sampling in the presence of linear trends, *Communications in Statistics - Theory and Methods*, **18**, 2511-2526.

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

[6] 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 & Information Science Society*, **10**, 91-102.

[7] 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.

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

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

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

[11] Yates, F. (1948), Systematic sampling, *Philosophical Transactions of the Royal Society of London*, **A241**, 345-377.

[ 2003년 4월 접수, 2003년 10월 채택 ]

# Estimation of Population Mean Using Modified Systematic Sampling and Least Squares Method *

Hyuk Joo Kim [1)]

ABSTRACT

In this paper, a new method is developed for estimating the mean of a population which has a linear trend. This method involves drawing a sample by the modified systematic sampling, and then estimating the population mean with an adjusted estimator, not with the sample mean itself. We use the method of least squares in determining the adjusted estimator. The proposed method is shown to be more and more efficient as the linear trend becomes stronger. It turns out to be relatively efficient as compared with the conventional methods if $\sigma^2$, the variance of the random error term in the infinite superpopulation model, is not very large.

*Keywords:* linear trend, population mean, modified systematic sampling, least squares method.

* This work was supported by grant No. R05-2001-000-00057-0 from the Basic Research program of the KOSEF.

1) Professor, Division of Mathematics & Informational Statistics and Institute of Basic Natural Sciences, Wonkwang University, Iksan, Jeonbuk 570-749, Korea.
E-mail: hjkim@wonkwang.ac.kr