Journal of the Korean Data & Information Science Society
2009, **20**(6), 1103–1118

한국데이터정보과학회지

선형추세를 갖는 모집단에 대한 변형계통표집의 일반화와 회귀추정법[†]

김혁주[1] · 김정현[2]

[1]원광대학교 수학·정보통계학부 · [2]원광대학교 정보통계학과

접수 2009년 4월 23일, 수정 2009년 9월 25일, 게재확정 2009년 10월 14일

요 약

유한모집단의 평균 또는 합계를 추정하고자 하는 경우 모집단 단위들의 배열순서는 중요한 의미를 갖는다. 본 논문에서는 표집률의 역수가 짝수이고 표본 크기가 홀수인 경우 선형추세를 갖는 모집단의 평균 또는 합계를 추정하기 위한 두 가지의 방법을 제시하였다. 첫째 방법은 Singh 등 (1968)의 변형계통표집을 일반화한 방법으로 표본을 뽑은 뒤, 추정량을 정하는 과정에서 보간법을 사용한 것이며, 둘째 방법은 변형계통표집으로 표본을 뽑은 뒤, 회귀추정법으로 모수를 추정하는 것이다. Cochran (1946)의 무한초모집단 모형에 근거를 둔 기대평균제곱오차를 기준으로 하여 기존의 방법들과 제시된 방법들을 비교하였으며, 제시된 두 방법 간의 상호 비교도 시행하였다.

주요용어: 모평균 추정, 무한초모집단 모형, 보간법, 선형추세, 일반화된 변형계통표집, 회귀추정법.

1. 서 론

유한모집단의 평균 또는 합계를 추정하는 문제를 생각하자. 이때 모집단 단위들의 배열순서는 중요한 의미를 갖는다. 모집단 단위들이 우리가 관심을 갖는 특성값과 무관한 순서로 배열되어 있는 경우도 많지만, 배열순서가 어떤 추세 (trend)를 보이는 경우도 종종 있다.

모집단이 추세를 갖는 경우에 자주 언급되는 표집 방법이 계통표집 (systematic sampling)이다. 계통표집은, 모집단이 증가하거나 감소하는 추세를 갖는 경우에는 좋은 방법이지만, 모집단에 주기성이 존재할 때에는 좋지 않은 방법으로 알려져 있다. 계통표집에는 우리가 흔히 얘기하는 보통의 계통표집 (Ordinary Systematic Sampling: OSS) 외에도 Madow (1953)의 중심계통표집 (Centered Systematic Sampling: CSS), Sethi (1965)와 Murthy (1967)의 균형계통표집 (Balanced Systematic Sampling: BSS), Singh 등 (1968)의 변형계통표집 (Modified Systematic Sampling: MSS) 등이 포함된다.

OSS, CSS, MSS, BSS가 어떤 방법인지 간략히 설명한다. 예를 들어 $k = 4$, $n = 5$인 경우를 생각해 보자. 모집단 단위들은 $U_1, U_2, \cdots, U_{20}$이다. 모집단을 다음과 같이 세 가지 방법으로 집락화한다.

1. $C_1 = \{U_1, U_5, U_9, U_{13}, U_{17}\}, C_2 = \{U_2, U_6, U_{10}, U_{14}, U_{18}\}$,
   $C_3 = \{U_3, U_7, U_{11}, U_{15}, U_{19}\}, C_4 = \{U_4, U_8, U_{12}, U_{16}, U_{20}\}$

[†] 이 논문은 2007년도 원광대학교의 교비 지원에 의해서 수행됨.
[1] 교신저자: (570-749) 전북 익산시 신용동 344-2, 원광대학교 수학·정보통계학부 및 기초자연과학연구소, 교수. E-mail: hjkim@wonkwang.ac.kr
[2] (570-749) 전북 익산시 신용동 344-2, 원광대학교 대학원 정보통계학과, 석사 졸업.

2. $C_1' = \{U_1, U_5, U_9, U_{16}, U_{20}\}, C_2' = \{U_2, U_6, U_{10}, U_{15}, U_{19}\},$
   $C_3' = \{U_3, U_7, U_{11}, U_{14}, U_{18}\}, C_4' = \{U_4, U_8, U_{12}, U_{13}, U_{17}\}$
3. $S_1' = \{U_1, U_8, U_9, U_{16}, U_{17}\}, S_2' = \{U_2, U_7, U_{10}, U_{15}, U_{18}\},$
   $S_3' = \{U_3, U_6, U_{11}, U_{14}, U_{19}\}, S_4' = \{U_4, U_5, U_{12}, U_{13}, U_{20}\}$

OSS는 C_1, C_2, C_3, C_4 중 하나를 1/4씩의 확률로 뽑는 방법이며, CSS는 가운데의 두 집락 C_2, C_3 중 하나를 1/2씩의 확률로 뽑는 방법이다 (k가 홀수인 경우에는 한가운데에 있는 $(k+1)/2$번째 집락을 확률 1로 뽑는다). MSS는 C_1', C_2', C_3', C_4' 중 하나를 1/4씩의 확률로 뽑는 방법이며, BSS는 S_1', S_2', S_3', S_4' 중 하나를 1/4씩의 확률로 뽑는 방법이다. 이상의 내용을 일반화하면 일반적인 경우도 쉽게 알 수 있다.

CSS, MSS, BSS는 모집단에 선형추세가 존재하는 경우 모평균이나 모합을 좀 더 효율적으로 추정하기 위하여 제안된 것으로서 OSS를 더욱 개량한 방법들이다. 예를 들어, 정해진 기간에 대하여 어떤 도시나 지역 안에 있는 슈퍼마켓들의 평균 매출액 (또는 매출액의 합계)을 추정하는 경우 슈퍼마켓들을 면적 순으로 배열한다면, 이 모집단에는 선형에 가까운 추세가 존재할 것이다. 표본조사를 통한 추정에 관한 최근의 연구로 김영화와 김기수 (2009), 박종태 (2009) 등의 연구가 있다.

Kim (1998)은 Singh 등 (1968)의 MSS로 표본을 뽑은 뒤 표본의 단순평균이 아닌 가중평균으로 모평균을 추정하는 방법을 제시하였다. 이 방법은 표집률 (sampling fraction)의 역수 k가 짝수이고 표본크기 (sample size) n이 홀수인 경우에 사용하기 위한 것으로서, 선형추세의 특성을 잘 살린 방법인 것으로 밝혀졌다. 그런데 MSS는, n이 홀수인 경우 $(n+1)/2$번째 표집원소의 관점에서 볼 때 일종의 불균형성을 갖는다. 이 불균형성의 의미는 2.1절에서 간단한 예를 통해 설명할 것이다.

본 논문에서는, k가 짝수이고 n이 홀수인 경우 MSS를 기반으로 한 두 가지 방법을 제시하고 연구하고자 한다. 첫째는 MSS에서 $(n+1)/2$번째 표집원소를 랜덤화함으로써 이러한 불균형성을 확률적으로 보완한 뒤 표본의 가중평균으로 모평균을 추정하는 방법이며, 둘째는 MSS에서 회귀추정량 (regression estimator)을 개발하여 모평균을 추정하는 방법이다. k가 짝수이고 n이 홀수인 상황은 실제 조사에서도 종종 일어날 수 있다. 예를 들면, 300개의 점포 중 25개를 뽑아 조사하는 경우, $k = 12,\ n = 25$가 된다.

2. 표집 방법과 모평균 추정 방법

2.1. 일반화된 변형계통표집과 보간법

Singh 등 (1968)은 OSS보다 효율적으로 모수를 추정하기 위하여, OSS에서 뽑힐 수 있는 가능한 집락들 간의 차이를 보정해 주는 MSS를 제안하였는데, 표본 크기 n이 홀수인 경우에는 좀 더 개량된 방법을 사용하는 것이 좋다.

예를 들어 $k = 4$, $n = 5$인 경우를 생각해 보자. 1절에서 설명한 바와 같이 MSS는 1/4씩의 확률로 C_1', C_2', C_3', C_4' 중 하나를 뽑는 방법이다. 그런데 이 4개의 집락에 속한 단위들의 번호를 합하면 각각 51, 52, 53, 54가 되어 약간의 불균형이 존재한다. 이것은 n이 홀수이기 때문에 나타나는 불가피한 현상이다. 따라서 다음과 같은 방법을 생각해 볼 수 있다. 예를 들어 C_1'의 경우 U_1, U_5, U_{16}, U_{20}은 그대로 뽑고 나머지 하나의 표본 단위로 U_9와 U_{12}가 각각 1/2의 확률로 뽑히게 한다. C_2'의 경우에는 U_2, U_6, U_{15}, U_{19}를 그대로 뽑고 U_{10}과 U_{11} 을 1/2씩의 확률로 뽑는다. C_3'과 C_4'의 경우에도 같은 방식으로 한다. 즉 이것은 다음과 같이 4개의 집락 $C_1'', C_2'', C_3'', C_4''$을 더 만들어 8개의 집락 $C_i', C_i''(i =$

$1, 2, 3, 4$) 중 하나를 각각 1/8의 확률로 뽑는 것과 같다.

$$C_1'' = \{U_1, U_5, U_{12}, U_{16}, U_{20}\}$$
$$C_2'' = \{U_2, U_6, U_{11}, U_{15}, U_{19}\}$$
$$C_3'' = \{U_3, U_7, U_{10}, U_{14}, U_{18}\}$$
$$C_4'' = \{U_4, U_8, U_9, U_{13}, U_{17}\}$$

이제 모평균의 추정 방법을 생각해 보자. 8개의 집락 각각에 대하여 집락 안에 있는 단위들의 번호를 합하면 51부터 54까지의 값이 나온다. 이러한 차이는 n이 홀수이기 때문에 생기는 것이다. $C_1'', C_2'', C_3'', C_4''$ 의 경우 각각 $y_{12}, y_{11}, y_{10}, y_9$를 '$y_{10.5}$'로 대체하면 균형이 이루어진다. 여기서 y_i는 단위 U_i의 특성값을 나타내며, $y_{10.5}$는 실제로는 존재하지 않는 가상의 값이다. C_1', C_2', C_3', C_4'의 경우에는 Kim (1998)에서와 같이 각각 $y_9, y_{10}, y_{11}, y_{12}$를 $y_{10.5}$로 대체하면 된다. 예를 들어, C_1''이 뽑혔다고 가정하자. 이 경우 y_5와 y_{12}를 사용하여 $y_{10.5}$를 다음과 같이 추정할 수 있다.

$$\begin{aligned}\widehat{y}_{10.5} &= y_{12} - \frac{12 - 10.5}{12 - 5}(y_{12} - y_5) \\ &= y_{12} - \frac{3}{14}(y_{12} - y_5) \end{aligned} \tag{2.1}$$

이 식은 보간법 (interpolation)을 사용하여 나온 결과이다. 모평균은 y_{12} 대신 $\widehat{y}_{10.5}$를 집어넣어 계산되는 평균값 $\bar{y}_1''^*$로 추정한다.

$$\begin{aligned}\bar{y}_1''^* &= \frac{1}{5}(y_1 + y_5 + \widehat{y}_{10.5} + y_{16} + y_{20}) \\ &= \bar{y}_1'' - \frac{3}{70}(y_{12} - y_5) \end{aligned} \tag{2.2}$$

여기서 $\bar{y}_1''$은 C_1''의 단순평균을 나타낸다.

추출된 집락이 C_2''이라면, $y_{10.5}$ 를 $y_{11} - (1/10)(y_{11} - y_6)$로 추정하여 모평균을

$$\bar{y}_2''^* = \bar{y}_2'' - \frac{1}{50}(y_{11} - y_6) \tag{2.3}$$

로 추정하게 된다. 같은 방식으로 하면, 추출된 집락이 $C_3'', C_4'', C_1', C_2', C_3', C_4'$인 경우 모평균의 추정량은 각각 다음과 같다.

$$\bar{y}_3''^* = \bar{y}_3'' + \frac{1}{40}(y_{14} - y_{10}) \tag{2.4}$$

$$\bar{y}_4''^* = \bar{y}_4'' + \frac{3}{40}(y_{13} - y_9) \tag{2.5}$$

$$\bar{y}_1'^* = \bar{y}_1' + \frac{3}{70}(y_{16} - y_9) \tag{2.6}$$

$$\bar{y}_2'^* = \bar{y}_2' + \frac{1}{50}(y_{15} - y_{10}) \tag{2.7}$$

$$\bar{y}_3'^* = \bar{y}_3' - \frac{1}{40}(y_{11} - y_7) \tag{2.8}$$

$$\bar{y}_4'^* = \bar{y}_4' - \frac{3}{40}(y_{12} - y_8) \tag{2.9}$$

이상의 내용을 일반화해 보자. 먼저 몇 가지 기호를 정의한다.

$$\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} y_i : \text{추정하고자 하는 모평균}$$

$$y'_{ij}(y''_{ij}) : \text{집락 } C'_i\ (C''_i)\text{의 } j\text{번째 단위의 특성값 } (i = 1, 2, \cdots, k; j = 1, 2, \cdots, n)$$

즉, n 이 홀수인 경우

$$y'_{ij} = y_{i+(j-1)k} \quad (j = 1, 2, \cdots, (n+1)/2)$$
$$y'_{ij} = y_{1-i+jk} \quad (j = (n+3)/2, (n+5)/2, \cdots, n)$$
$$y''_{ij} = y_{i+(j-1)k} \quad (j = 1, 2, \cdots, (n-1)/2)$$
$$y''_{ij} = y_{1-i+jk} \quad (j = (n+1)/2, (n+3)/2, \cdots, n)$$
$$\bar{y}'_i = \frac{1}{n}\sum_{j=1}^{n} y'_{ij} \quad \left(\bar{y}''_i = \frac{1}{n}\sum_{j=1}^{n} y''_{ij}\right) : C'_i\ (C''_i)\text{의 평균 } (i = 1, 2, \cdots, k)$$

$2k$개의 집락 $C'_1, \cdots, C'_k, C''_1, \cdots, C''_k$ 중 하나를 $1/2k$씩의 확률로 뽑는다. 뽑힌 집락이 C'_i이면 $\bar{Y}$를 $\bar{y}'^*_i$로 추정하고, C''_i이면 $\bar{Y}$를 $\bar{y}''^*_i$로 추정한다. 여기서 $\bar{y}'^*_i$와 $\bar{y}''^*_i$는 다음의 식으로 표시된다.

$$\bar{y}'^*_i = \bar{y}'_i + \frac{k+1-2i}{2n(2k+1-2i)}(y'_{i,(n+3)/2} - y'_{i,(n+1)/2}) \quad (i = 1, 2, \cdots, k/2) \tag{2.10}$$

$$\bar{y}'^*_i = \bar{y}'_i - \frac{2i-k-1}{2nk}(y'_{i,(n+1)/2} - y'_{i,(n-1)/2}) \quad (i = k/2+1, k/2+2, \cdots, k) \tag{2.11}$$

$$\bar{y}''^*_i = \bar{y}''_i - \frac{k+1-2i}{2n(2k+1-2i)}(y''_{i,(n+1)/2} - y''_{i,(n-1)/2}) \quad (i = 1, 2, \cdots, k/2) \tag{2.12}$$

$$\bar{y}''^*_i = \bar{y}''_i + \frac{2i-k-1}{2nk}(y''_{i,(n+3)/2} - y''_{i,(n+1)/2}) \quad (i = k/2+1, k/2+2, \cdots, k) \tag{2.13}$$

이렇게 $\bar{Y}$를 추정하는 방법을 GMI (Generalized Modified systematic sampling with Interpolation)로 나타내고 그에 따른 추정량을 $\bar{y}_{GMI}$로 나타내자. 즉

$$P(\bar{y}_{GMI} = \bar{y}'^*_i) = P(\bar{y}_{GMI} = \bar{y}''^*_i) = \frac{1}{2k} \quad (i = 1, 2, \cdots, k) \tag{2.14}$$

이다. $\bar{y}_{GMI}$는 $\bar{Y}$에 대한 편향추정량 (biased estimator)이며, 다음과 같은 편향 (bias)과 평균제곱오차 (mean square error)를 갖는다는 것을 쉽게 보일 수 있다.

$$Bias(\bar{y}_{GMI}) = \frac{1}{2k}\sum_{i=1}^{k}(\bar{y}'^*_i + \bar{y}''^*_i) - \bar{Y} \tag{2.15}$$

$$MSE(\bar{y}_{GMI}) = \frac{1}{2k}\sum_{i=1}^{k}\left\{(\bar{y}'^*_i - \bar{Y})^2 + (\bar{y}''^*_i - \bar{Y})^2\right\} \tag{2.16}$$

2.2. 변형계통표집과 회귀추정법

회귀추정 (regression estimation)방법이란, 관심변수 y에 관하여 추정하고자 할 때, y와 밀접한 관계가 있는 보조변수 x에 관한 정보를 이용하여 y를 추정하는 방법이다. 통상적인 단순선형회귀모형은

$$y = \alpha + \beta x + \epsilon \tag{2.17}$$

으로 표시되며, 최소제곱법 (method of least squares)에 의해 적합된 회귀선은

$$\widehat{y} = \widehat{\alpha} + \widehat{\beta} x \tag{2.18}$$

로 나타내진다. 여기서 추정된 회귀계수 $\widehat{\alpha}, \widehat{\beta}$의 식은 다음과 같다.

$$\widehat{\beta} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{2.19}$$

$$\widehat{\alpha} = \bar{y} - \widehat{\beta}\bar{x} \tag{2.20}$$

x, y의 모평균을 각각 $\bar{X}, \bar{Y}$라 하면, $\bar{Y}$에 대한 회귀추정량 (regression estimator)은 다음과 같다.

$$\widehat{\bar{Y}} = \widehat{\alpha} + \widehat{\beta}\bar{X} = \bar{y} + \widehat{\beta}\left(\bar{X} - \bar{x}\right) \tag{2.21}$$

$k = 4, n = 5$인 경우를 생각하자. MSS로 표본을 뽑으면, 2.1절에서 정의된 집락 C_1', C_2', C_3', C_4' 중 하나가 뽑힌다. C_1'이 뽑힌 경우 모평균은 다음과 같이 추정된다. 단위들의 번호인 1, 5, 9, 16, 20을 x변수 (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 y변수 (반응변수) 값들로 보아 최소제곱법을 이용하여 선형추세의 직선을 적합하면 적합된 회귀직선의 방정식은

$$\widehat{y}_1 = \widehat{\alpha}_1 + \widehat{\beta}_1 x_1 \tag{2.22}$$

이다. 단, 여기서 $\widehat{\beta}_1 = S_{(xy)1}/S_{(xx)1} = \sum_{j=1}^{5}(x_{1j} - \bar{x}_1)\left(y_{1j}' - \bar{y}_1'\right) / \sum_{j=1}^{5}(x_{1j} - \bar{x}_1)^2$이고, $\widehat{\alpha}_1 = \bar{y}_1' - \widehat{\beta}_1\bar{x}_1$이며, $\bar{x}_1 = (1+5+9+16+20)/5 = 10.2$이고 $\bar{y}_1'$은 2.1절에서 정의된 바와 같다. 첨자 1은 첫째 집락(C_1')을 나타내는 것이다.

$$\widehat{\beta}_1 = \frac{1}{242.8}\left(-9.2y_{11}' - 5.2y_{12}' - 1.2y_{13}' + 5.8y_{14}' + 9.8y_{15}'\right) \tag{2.23}$$

이므로, 식 (2.21)에 따라

$$\begin{aligned}\widehat{\bar{Y}} &= \bar{y}_1' + \widehat{\beta}_1 (10.5 - 10.2) \\ &= 0.1886y_{11}' + 0.1936y_{12}' + 0.1985y_{13}' + 0.2072y_{14}' + 0.2121y_{15}'\end{aligned} \tag{2.24}$$

가 된다. C_2', C_3' 또는 C_4'이 뽑힌 경우에도 위와 같은 방법으로 $\bar{Y}$를 추정할 수 있다.

이상의 내용을 일반화하면 다음과 같다. k가 짝수이고 n이 3 이상의 홀수인 경우 MSS에 의하여 하나의 집락을 뽑는다. 즉, 각각 $1/k$의 확률로 k개의 집락 $C_1', C_2', \cdots, C_k'$ 중 하나를 뽑는다. 뽑힌 집락을 C_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 최소제곱법을 적용하여 $\bar{Y}$의 값을 추정한다.

$$x_{ij} = \begin{cases} i + (j-1)k & (j = 1, 2, \cdots, (n+1)/2) \\ N + 1 - i - (n-j)k = 1 - i + jk & (j = (n+3)/2, (n+5)/2, \cdots, n) \end{cases} \tag{2.25}$$

이므로,

$$\bar{x}_i = \frac{1}{n}\sum_{j=1}^{n} x_{ij} = \frac{N+1}{2} + \frac{1}{n}\left(i - \frac{k+1}{2}\right) \tag{2.26}$$

이고,

$$\begin{aligned} S_{(xx)i} &= \sum_{j=1}^{n} x_{ij}^2 - n(\bar{x}_i)^2 \\ &= \left(n - \frac{1}{n}\right) i^2 - \left(n - \frac{1}{n}\right)\left\{k\left(\frac{n}{2}+1\right)+1\right\} i \\ &+ \frac{1}{12}\left(n - \frac{1}{n}\right)\left\{\left(n^2+3n+3\right)k^2 + 3\left(n+2\right)k + 3\right\} \end{aligned} \tag{2.27}$$

이다. 회귀직선의 기울기의 추정값은 $\widehat{\beta}_i = \sum_{j=1}^{n} a_{ij} y'_{ij}$ (단, $a_{ij} = (x_{ij} - \bar{x}_i)/S_{(xx)i}, i = 1, 2, \cdots, n$)이며, 모평균 $\bar{Y}$는

$$\begin{aligned} \widehat{\bar{Y}}_{(i)} &= \bar{y}'_i + \widehat{\beta}_i\left(\frac{N+1}{2} - \bar{x}_i\right) \\ &= \left\{\frac{1}{n} - \frac{a_{i1}}{n}\left(i - \frac{k+1}{2}\right)\right\} y'_{i1} + \cdots + \left\{\frac{1}{n} - \frac{a_{in}}{n}\left(i - \frac{k+1}{2}\right)\right\} y'_{in} \end{aligned} \tag{2.28}$$

에 의하여 추정된다.

이 방법을 MREG로 표시하자 (M은 MSS, REG는 Regression을 나타낸다). MREG에 의한 $\bar{Y}$의 추정량을 $\bar{y}_{MREG}$로 나타내면 $\bar{y}_{MREG}$는 다음과 같은 평균제곱오차를 갖는다.

$$MSE\left(\bar{y}_{MREG}\right) = \frac{1}{k}\sum_{i=1}^{k}\left(\widehat{\bar{Y}}_{(i)} - \bar{Y}\right)^2 \tag{2.29}$$

3. 추정량의 기대평균제곱오차

이 절에서는 $\bar{y}_{GMI}$와 $\bar{y}_{MREG}$의 평균제곱오차의 기댓값을 논의한다. 모집단에 추세가 존재한다고 해도 특성값들이 추세대로 정확히 나타나는 것이 아니기 때문에, 평균제곱오차 자체를 기준으로 추정량들을 비교할 수는 없다. 추세를 갖는 모집단의 경우 추정량들을 비교하기 위한 훌륭한 이론적 근거가 되는 것이 Cochran (1946)의 무한초모집단 모형 (infinite superpopulation model)이다. 이 모형은 주어진 유한모집단을 무한한 초모집단으로부터 뽑힌 표본으로 간주하는 것이다. 먼저 일반적인 경우로서 모형을 다음과 같이 설정한다.

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

여기서 μ_i는 i의 함수이며, e_i들은 $E(e_i) = 0$, $E(e_i^2) = \sigma^2$, $E(e_i e_j) = 0$ $(i \neq j)$이 성립하는 랜덤오차항이다. 연산자 E는 무한초모집단에 걸쳐 취한 기댓값을 나타낸다.

이제부터 μ와 e에 관해서도 y에 관해 사용한 것과 같은 양식의 기호를 사용한다. 예를 들면,

$$\begin{aligned}
\bar{\mu} &= \frac{1}{N}\sum_{i=1}^{N}\mu_i \\
\mu'_{ij} &= \mu_{i+(j-1)k} \quad (j=1,2,\cdots,(n+1)/2)(n:\text{홀수}) \\
\bar{\mu}'_i &= \frac{1}{n}\sum_{j=1}^{n}\mu'_{ij} \\
\bar{\mu}'^{*}_i &= \bar{\mu}'_i + \frac{k+1-2i}{2n(2k+1-2i)}(\mu'_{i,(n+3)/2} - \mu'_{i,(n+1)/2}) \quad (i=1,2,\cdots,(k-1)/2)
\end{aligned}$$

등이다.

$\bar{y}_{GMI}$와 $\bar{y}_{MREG}$의 효율성을 평가함에 있어서 다음의 정리가 매우 중요한 역할을 한다.

정리 3.1 식 (3.1)로 주어진 모형을 가정할 때, k가 짝수이고 n이 3 이상의 홀수인 경우 $\bar{y}_{GMI}$와 $\bar{y}_{MREG}$의 기대평균제곱오차는 다음과 같다.

$$EMSE(\bar{y}_{GMI}) = \frac{1}{2k}\sum_{i=1}^{k}\left\{(\bar{\mu}'^{*}_i-\bar{\mu})^2 + (\bar{\mu}''^{*}_i-\bar{\mu})^2\right\} + \frac{\sigma^2}{n}\frac{N-n}{N} + \frac{\sigma^2}{12n^2}S_k \tag{3.2}$$

$$EMSE\ (\bar{y}_{MREG}) = \frac{1}{k}\sum_{i=1}^{k}\left(\widehat{\bar{\mu}}_{(i)}-\bar{\mu}\right)^2 + \frac{\sigma^2}{n}\frac{N-n}{N} + \frac{\sigma^2}{Nn}\sum_{i=1}^{k}\frac{\left(i-\frac{k+1}{2}\right)^2}{S_{(xx)i}} \tag{3.3}$$

단, 여기서

$$\begin{aligned}
S_k &= 4 - 12A_k + 6kB_k - \frac{1}{k^2} \\
A_k &= \sum_{i=1}^{k/2}\frac{1}{2k+1-2i} = \frac{1}{2}\left\{\psi\left(k+\frac{1}{2}\right) - \psi\left(\frac{k+1}{2}\right)\right\} \\
B_k &= \sum_{i=1}^{k/2}\frac{1}{(2k+1-2i)^2} = -\frac{1}{4}\left\{\psi^{(1)}\left(k+\frac{1}{2}\right) - \psi^{(1)}\left(\frac{k+1}{2}\right)\right\} \\
\psi(x) &= \frac{d}{dx}\ln\Gamma(x)(x>0) : polygamma\text{함수} \\
\Gamma(x) &= \int_0^{\infty} t^{x-1}e^{-t}dt(x>0) : gamma\text{함수} \\
\psi^{(1)}(x) &= \frac{d}{dx}\psi(x)
\end{aligned}$$

이다.

증명: 식 (3.1)에 의해

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

이며, 또한

$$\begin{aligned} y'_{ij} &= \mu'_{ij} + e'_{ij} \\ y''_{ij} &= \mu''_{ij} + e''_{ij}, \qquad (i = 1, 2, \cdots, k; j = 1, 2, \cdots, n) \end{aligned} \tag{3.5}$$

이다. 따라서

$$\begin{aligned} \bar{y}'^{*}_{i} &= \bar{\mu}'^{*}_{i} + \bar{e}'^{*}_{i} \\ \bar{y}''^{*}_{i} &= \bar{\mu}''^{*}_{i} + \bar{e}''^{*}_{i} \qquad (i = 1, 2, \cdots, k) \end{aligned} \tag{3.6}$$

을 얻는다. 식 (3.4)와 (3.6)을 식 (2.15)에 대입하면 쉽게 식 (3.2)를 얻는다.

한편, 식 (3.4)와 (3.6)을 식 (2.16)에 대입하고 기댓값을 취하면 다음을 얻는다.

$$\begin{aligned} EMSE(\bar{y}_{GMI}) &= \frac{1}{2k}\sum_{i=1}^{k} E\left[\left\{(\bar{\mu}'^{*}_{i} - \bar{\mu}) + (\bar{e}'^{*}_{i} - \bar{e})\right\}^2 + \left\{(\bar{\mu}''^{*}_{i} - \bar{\mu}) + (\bar{e}''^{*}_{i} - \bar{e})\right\}^2\right] \\ &= \frac{1}{2k}\sum_{i=1}^{k}\left[(\bar{\mu}'^{*}_{i} - \bar{\mu})^2 + (\bar{\mu}''^{*}_{i} - \bar{\mu})^2 + E\left\{(\bar{e}'^{*}_{i} - \bar{e})^2\right\} + E\left\{(\bar{e}''^{*}_{i} - \bar{e})^2\right\}\right] \end{aligned} \tag{3.7}$$

그런데 $i = 1, 2, \cdots, k/2$에 대해 다음이 성립한다.

$$\begin{aligned} E\left\{(\bar{e}'^{*}_{i} - \bar{e})^2\right\} &= E\left\{(\bar{e}'_{i} - \bar{e} + P_i)^2\right\} \\ &= E\left\{(\bar{e}'_{i} - \bar{e})^2\right\} + 2E\left\{(\bar{e}'_{i} - \bar{e})P_i\right\} + E\left(P_i^2\right) \end{aligned} \tag{3.8}$$

단, 여기서

$$P_i = \frac{k + 1 - 2i}{2n(2k + 1 - 2i)}\left(e'_{i,(n+3)/2} - e'_{i,(n+1)/2}\right) \tag{3.9}$$

이다. 역시 $i = 1, 2, \cdots, k/2$에 대해

$$E\left\{(\bar{e}'_{i} - \bar{e})^2\right\} = E\left\{(\bar{e}'_{i})^2\right\} - 2E\left\{(\bar{e}'_{i})(\bar{e})\right\} + E\left\{(\bar{e})^2\right\} \tag{3.10}$$

이며,

$$\begin{aligned} E\left\{(\bar{e}'_{i})^2\right\} &= E\left\{\left(\frac{1}{n}\sum_{j=1}^{n} e'_{ij}\right)^2\right\} \\ &= \frac{1}{n^2}E\left\{\sum_{j=1}^{n}(e'_{ij})^2 + 2\sum_{j<j'}(e'_{ij})(e'_{ij'})\right\} \\ &= \frac{1}{n^2}\left[\sum_{j=1}^{n} E\left\{(e'_{ij})^2\right\} + 2\sum_{j<j'} E\left\{(e'_{ij})(e'_{ij'})\right\}\right] \\ &= \frac{1}{n^2}(n\sigma^2 + 0) \\ &= \frac{\sigma^2}{n} \end{aligned} \tag{3.11}$$

이고, 유사한 방법으로

$$E\left\{(\bar{e}_i')(\bar{e})\right\} = E\left\{(\bar{e})^2\right\} = \frac{\sigma^2}{N} \tag{3.12}$$

임을 보일 수 있다.

식 (3.8)의 우측 끝변의 둘째 항은 0이 됨을 쉽게 알 수 있고, 셋째 항은

$$E(P_i^2) = \frac{(k+1-2i)^2\sigma^2}{2n^2(2k+1-2i)^2} \tag{3.13}$$

으로 쉽게 얻어진다. 이 결과들을 식 (3.8)에 넣으면

$$E\left\{(\bar{e}_i'^* - \bar{e})^2\right\} = \frac{\sigma^2}{n}\frac{N-n}{N} + \frac{(k+1-2i)^2\sigma^2}{2n^2(2k+1-2i)^2} \quad (i = 1, 2, \cdots, k/2) \tag{3.14}$$

가 얻어진다.

지금까지 식 (3.14)를 얻은 것과 유사한 과정을 거쳐서 다음 결과가 얻어진다.

$$E\left\{(\bar{e}_i'^* - \bar{e})^2\right\} = \frac{\sigma^2}{n}\frac{N-n}{N} + \frac{(k+1-2i)^2\sigma^2}{2n^2k^2} \quad (i = k/2+1, k/2+2, \cdots, k) \tag{3.15}$$

$$E\left\{(\bar{e}_i''^* - \bar{e})^2\right\} = \frac{\sigma^2}{n}\frac{N-n}{N} + \frac{(k+1-2i)^2\sigma^2}{2n^2(2k+1-2i)^2} \quad (i = 1, 2, \cdots, k/2) \tag{3.16}$$

$$E\left\{(\bar{e}_i''^* - \bar{e})^2\right\} = \frac{\sigma^2}{n}\frac{N-n}{N} + \frac{(k+1-2i)^2\sigma^2}{2n^2k^2} \quad (i = k/2+1, k/2+2, \cdots, k) \tag{3.17}$$

식 (3.14)부터 (3.17)까지를 식 (3.7)에 대입하면

$$\begin{aligned} EMSE(\bar{y}_{GMI}) = {} & \frac{1}{2k}\sum_{i=1}^{k}\left\{(\bar{\mu}_i'^* - \bar{\mu})^2 + (\bar{\mu}_i''^* - \bar{\mu})^2\right\} + \frac{\sigma^2}{n}\frac{N-n}{N} \\ & + \frac{1}{k}\sum_{i=1}^{k/2}\frac{(k+1-2i)^2\sigma^2}{2n^2(2k+1-2i)^2} + \frac{1}{k}\sum_{i=k/2+1}^{k}\frac{(k+1-2i)^2\sigma^2}{2n^2k^2} \end{aligned} \tag{3.18}$$

이 되며, 합에 관한 아래의 결과를 이용하면 식 (3.2)를 얻게 된다.

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

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

$$\sum_{i=k/2+1}^{k} i = \frac{k(3k+2)}{8}$$

$$\sum_{i=k/2+1}^{k} i^2 = \frac{k(k+1)(7k+2)}{24}$$

식 (3.3)도 식 (3.2)와 유사한 개념과 과정에 의해 증명된다. □

이제 모집단이 선형추세를 갖는 경우를 생각하자. 선형추세는 $\mu_i = a + bi$로 표현된다. 여기서 a와 b는 상수이며, $b \neq 0$이다. 즉

$$y_i = a + bi + e_i (i = 1, 2, \cdots, N) \tag{3.19}$$

의 모형을 가정한다. 여기서 e_i들은 식 (3.1)에서와 같이 $E(e_i) = 0$, $E(e_i^2) = \sigma^2$, $E(e_i e_j) = 0$ $(i \neq j)$이 성립하는 랜덤오차항이다.

이 경우 $\bar{y}_{GMI}$의 기대평균제곱오차를 구하기 위한 준비로서 다음의 식들을 얻는다.

$$\bar{\mu} = a + \left(\frac{b}{2}\right)(N+1) \tag{3.20}$$

$$\bar{\mu}_i' = a + \left(\frac{b}{2}\right)(N+1) + \left(\frac{b}{n}\right)\left(i - \frac{k+1}{2}\right) \tag{3.21}$$

$$\mu_{i,(n+3)/2}' = \mu_{1-i+(n+3)k/2} = a + b\left\{1 - i + \frac{(n+3)\,k}{2}\right\} \tag{3.22}$$

$$\mu_{i,(n+1)/2}' = \mu_{i+(n-1)k/2} = a + b\left\{i + \frac{(n-1)\,k}{2}\right\} \tag{3.23}$$

$$\mu_{i,(n-1)/2}' = \mu_{i+(n-3)k/2} = a + b\left\{i + \frac{(n-3)\,k}{2}\right\} \tag{3.24}$$

$$\bar{\mu}_i'^{*} = a + \left(\frac{b}{2}\right)(N+1)\,(i = 1, 2, \cdots, k) \tag{3.25}$$

$$\bar{\mu}_i'' = a + \left(\frac{b}{2}\right)(N+1) - \left(\frac{b}{n}\right)\left(i - \frac{k+1}{2}\right) \tag{3.26}$$

$$\mu_{i,(n+3)/2}'' = \mu_{1-i+(n+3)k/2} = a + b\left\{1 - i + \frac{(n+3)\,k}{2}\right\} \tag{3.27}$$

$$\mu_{i,(n+1)/2}'' = \mu_{1-i+(n+1)k/2} = a + b\left\{1 - i + \frac{(n+1)\,k}{2}\right\} \tag{3.28}$$

$$\mu_{i,(n-1)/2}'' = \mu_{i+(n-3)k/2} = a + b\left\{i + \frac{(n-3)\,k}{2}\right\} \tag{3.29}$$

$$\bar{\mu}_i''^{*} = a + \left(\frac{b}{2}\right)(N+1)\,(i = 1, 2, \cdots, k) \tag{3.30}$$

위의 식들 중 예컨대 식 (3.26)이 나오는 과정을 보이면 다음과 같다.

$$\begin{aligned}\bar{\mu}_i'' &= \frac{1}{n}\sum_{j=1}^{n}\mu_{ij}'' \\ &= \frac{1}{n}\left\{\sum_{j=1}^{(n-1)/2}\mu_{i+(j-1)k} + \sum_{j=(n+1)/2}^{n}\mu_{1-i+jk}\right\} \\ &= \frac{1}{n}\left[\sum_{j=1}^{(n-1)/2}\{a+b\,(i+(j-1)\,k)\} + \sum_{j=(n+1)/2}^{n}\{a+b\,(1-i+jk)\}\right] \\ &= a+\left(\frac{b}{2}\right)(N+1) - \left(\frac{b}{n}\right)\left(i-\frac{k+1}{2}\right)\end{aligned}$$

여기서 $\sum_{j=1}^{(n-1)/2} j = (n+1)\,(n-1)/8$, $\sum_{j=(n+1)/2}^{n} j = (n+1)\,(3n+1)/8$임을 이용했다.

식 (3.20), (3.25), (3.30)에 의해 $\bar{\mu}_i'^{*} - \bar{\mu} = 0$, $\bar{\mu}_i''^{*} - \bar{\mu} = 0$이므로, <정리 3.1>의 결과를 이용하면 다음의 정리를 얻는다. $\bar{y}_{MREG}$의 기대평균제곱오차가 얻어지는 과정도 유사하다.

정리 3.2 모집단에 식 (3.19)로 주어지는 선형추세가 존재하는 경우 $\bar{y}_{GMI}$와 $\bar{y}_{MREG}$의 기대평균제곱오차는 다음과 같다. 여기서 k는 짝수이고 n은 3 이상의 홀수이다.

$$EMSE\,(\bar{y}_{GMI}) = \frac{\sigma^2}{n}\frac{N-n}{N} + \frac{\sigma^2}{12n^2}S_k \tag{3.31}$$

$$EMSE\,(\bar{y}_{MREG}) = \frac{\sigma^2}{n}\frac{N-n}{N} + \frac{\sigma^2}{Nn}\sum_{i=1}^{k}\frac{\left(i-\frac{k+1}{2}\right)^2}{S_{(xx)i}} \tag{3.32}$$

4. 효율성 비교

4.1. 전통적인 방법들과의 효율성 비교

이 절에서는 $\bar{y}_{GMI}$의 효율성을 전통적인 방법에 의한 추정량들의 효율성과 비교한다. 선형추세의 경우 전통적인 방법에 의한 $\bar{Y}$의 추정량들은 다음과 같다.

(1) 단순랜덤표집 (Simple Random Sampling: SRS)

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

(2) 보통의 계통표집 (Ordinary Systematic Sampling: OSS)

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

(3) 변형계통표집 (Modified Systematic Sampling: MSS) (Singh 등, 1968)

$$EMSE\,(\bar{y}_{MSS}) = \left(\frac{b^2}{12n^2}\right)(k+1)\,(k-1) + \frac{\sigma^2}{n}\frac{N-n}{N} \qquad (n:\text{홀수}) \tag{4.3}$$

(4) 균형계통표집 (Balanced Systematic Sampling: BSS) (Sethi, 1965; Murthy, 1967)

$$EMSE\left(\bar{y}_{BSS}\right)=\left(\frac{b^2}{12n^2}\right)(k+1)(k-1)+\frac{\sigma^2}{n}\frac{N-n}{N}\qquad (n:\text{홀수}) \tag{4.4}$$

(5) 중심계통표집 (Centered Systematic Sampling: CSS) (Madow, 1953)

$$EMSE\left(\bar{y}_{CSS}\right)=\frac{b^2}{4}+\frac{\sigma^2}{n}\frac{N-n}{N}\qquad (k:\text{짝수}) \tag{4.5}$$

식 (3.31)과 식 (4.1)부터 (4.5)까지를 근거로 하여, GMI를 포함한 여러 방법들의 효율성을 비교하면 다음과 같다. 물론 기대평균제곱오차가 작은 쪽이 더 효율적이다. 예컨대 '$GMI<OSS$'라는 표현은 $EMSE\left(\bar{y}_{GMI}\right)<EMSE\left(\bar{y}_{OSS}\right)$, 즉 GMI가 OSS보다 효율적임을 의미한다. 참고로, OSS, BSS, MSS와 CSS의 효율성 비교는 Bellhouse와 Rao (1975)에도 주어져 있다.

정리 4.1 식 (3.19)로 표현되는 선형추세가 존재하는 경우 다음 내용이 성립한다.

(1) $k=2$이고 $n=3,5,7,\cdots$ 인 경우

1. $\sigma^2<36b^2/13$이면, $GMI<MSS=BSS<CSS=OSS<SRS$.
2. $36b^2/13\le\sigma^2<36b^2n^2/13$이면, $MSS=BSS\le GMI<CSS=OSS<SRS$.
3. $36b^2n^2/13\le\sigma^2<12b^2n^2\left(N+1\right)/13$이면, $MSS=BSS<CSS=OSS\le GMI<SRS$.
4. $12b^2n^2\left(N+1\right)/13\le\sigma^2$이면, $MSS=BSS<CSS=OSS<SRS\le GMI$.

(2) $k=4,6,8,\cdots$이고 $n=3,5,7,\cdots$이며 $n<\sqrt{\left(k^2-1\right)/3}$인 경우

1. $\sigma^2<3b^2n^2/S_k$이면, $GMI<CSS<MSS=BSS<OSS<SRS$.
2. $3b^2n^2/S_k\le\sigma^2<b^2\left(k^2-1\right)/S_k$이면, $CSS\le GMI<MSS=BSS<OSS<SRS$.
3. $b^2\left(k^2-1\right)/S_k\le\sigma^2<b^2n^2\left(k^2-1\right)/S_k$이면, $CSS<MSS=BSS\le GMI<OSS<SRS$.
4. $b^2n^2\left(k^2-1\right)/S_k\le\sigma^2<b^2n^2\left(N+1\right)(k-1)/S_k$이면, $CSS<MSS=BSS<OSS\le GMI<SRS$.
5. $b^2n^2\left(N+1\right)(k-1)/S_k\le\sigma^2$이면, $CSS<MSS=BSS<OSS<SRS\le GMI$.

(3) $k=4,6,8,\cdots$이고 $n=3,5,7,\cdots$이며 $n=\sqrt{\left(k^2-1\right)/3}$인 경우 (예를 들면, $k=26$, $n=15$)

1. $\sigma^2<3b^2n^2/S_k$이면, $GMI<CSS=MSS=BSS<OSS<SRS$.
2. $3b^2n^2/S_k\le\sigma^2<b^2n^2\left(k^2-1\right)/S_k$이면, $CSS=MSS=BSS\le GMI<OSS<SRS$.
3. $b^2n^2\left(k^2-1\right)/S_k\le\sigma^2<b^2n^2\left(N+1\right)(k-1)/S_k$이면, $CSS=MSS=BSS<OSS<GMI\le SRS$.
4. $b^2n^2\left(N+1\right)(k-1)/S_k\le\sigma^2$이면, $CSS=MSS=BSS<OSS<SRS\le GMI$.

(4) $k=4,6,8,\cdots$이고 $n=3,5,7,\cdots$이며 $n>\sqrt{\left(k^2-1\right)/3}$인 경우

1. $\sigma^2<b^2\left(k^2-1\right)/S_k$이면, $GMI<MSS=BSS<CSS<OSS<SRS$.

2. $b^2\left(k^2-1\right)/S_k \le \sigma^2 < 3b^2n^2/S_k$이면, $MSS = BSS \le GMI < CSS < OSS < SRS$.
3. $3b^2n^2/S_k \le \sigma^2 < b^2n^2\left(k^2-1\right)/S_k$이면, $MSS = BSS < CSS \le GMI < OSS < SRS$.
4. $b^2n^2\left(k^2-1\right)/S_k \le \sigma^2 < b^2n^2(N+1)(k-1)/S_k$이면, $MSS = BSS < CSS < OSS \le GMI < SRS$.
5. $b^2n^2(N+1)(k-1)/S_k \le \sigma^2$이면, $MSS = BSS < CSS < OSS < SRS \le GMI$.

예제 4.1 모집단 크기가 $N = 450$이고 표본 크기가 $n = 15$라 하자. 따라서 $k = 30$이며, k는 짝수이고 n은 홀수이다. 기울기 $b = 0.5$인 선형추세가 모집단에 존재한다고 하자. 이 경우 필요한 polygamma 함수와 그의 도함수의 값들은 다음과 같이 계산된다.

$$\psi(30.5) = 3.401244, \quad \psi(15.5) = 2.708235$$
$$\psi^{(1)}(30.5) = 0.0333302, \quad \psi^{(1)}(15.5) = 0.0666420$$
$$A_{30} = \frac{1}{2}\{\psi(30.5) - \psi(15.5)\} = 0.3465045$$
$$B_{30} = -\frac{1}{4}\left\{\psi^{(1)}(30.5) - \psi^{(1)}(15.5)\right\} = 0.00832795$$
$$S_{30} = 4 - 12A_{30} + 180B_{30} - \frac{1}{(30)^2} = 1.339866$$

따라서 정리 4.1의 (2)에 의하여 여러 방법들의 효율성은 다음과 같이 비교된다.

(i) $\sigma^2 < 125.9454$이면, $GMI < CSS < MSS = BSS < OSS < SRS$.
(ii) $125.9454 \le \sigma^2 < 167.7407$이면, $CSS \le GMI < MSS = BSS < OSS < SRS$.
(iii) $167.7407 \le \sigma^2 < 37741.6473$이면, $CSS < MSS = BSS \le GMI < OSS < SRS$.
(iv) $37741.6473 \le \sigma^2 < 549080.0946$이면, $CSS < MSS = BSS < OSS \le GMI < SRS$.
(v) $549080.0946 \le \sigma^2$이면, $CSS < MSS = BSS < OSS < SRS \le GMI$.

위의 예에서 볼 수 있듯이 무한초모집단 모형에서의 오차항의 분산 σ^2이 작을수록 GMI는 전통적인 방법들에 비해 효율적이다. σ^2이 비현실적으로 크면 GMI는 비효율적인 방법이 되는데, 이런 경우는 선형추세 자체가 무의미한 경우가 되므로, 현실적인 경우에는 GMI가 다른 방법들에 비해 우수하다고 할 수 있다.

MREG와 전통적 방법들 간의 효율성 비교도 똑같은 패턴을 보이므로 상세한 내용은 생략한다.

4.2. GMI와 MREG 간의 효율성 비교

선형추세의 경우 식 (3.31)과 (3.32)를 근거로 하여 GMI와 MREG의 효율성을 비교할 수 있다. 참고로, Yates (1948)의 끝값수정법 (End Corrections: EC), 그리고 김혁주 (2004)에서 소개된 MLS라는 방법의 효율성도 함께 비교한다. EC는, OSS로 표본을 뽑은 뒤 표본의 첫 단위와 끝단위에 통상적인 가중치 $1/n$대신 각각 $1/n + (2i-k-1)/2k(n-1)$과 $1/n - (2i-k-1)/2k(n-1)$이라는 가중치를 주어 모평균 $\bar{Y}$를 추정하는 방법이다. MLS는, MSS로 표본을 뽑고 나서 최소제곱법을 사용한 뒤 한 단계를 더 거쳐서 모평균을 추정하는 방법으로서, 본 논문에서 제시된 MREG와는 다르며 더 복잡한 방법이다. GMI, MREG, EC와 MLS 이 네 방법의 공통적인 특징은 기대평균제곱오차가 k, n, σ^2의 값에만 의

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

n	GMI	MREG	EC	MLS
5	0.1794	0.1754	0.1853	0.2071
25	0.0352	0.0350	0.0353	0.0365
55	0.0159	0.0159	0.0160	0.0162
105	0.0083	0.0083	0.0083	0.0084

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

n	GMI	MREG	EC	MLS
5	0.1878	0.1837	0.1937	0.2155
25	0.0368	0.0367	0.0370	0.0382
55	0.0167	0.0167	0.0167	0.0170
105	0.0087	0.0087	0.0087	0.0088

표 **4.3** $k = 16$인 경우 $EMSE(\cdot)/\sigma^2$의 값들

n	GMI	MREG	EC	MLS
5	0.1920	0.1879	0.1979	0.2196
25	0.0377	0.0375	0.0378	0.0390
55	0.0171	0.0170	0.0171	0.0174
105	0.0089	0.0089	0.0089	0.0090

표 **4.4** $k = 20$인 경우 $EMSE(\cdot)/\sigma^2$의 값들

n	GMI	MREG	EC	MLS
5	0.1945	0.1904	0.2004	0.2221
25	0.0382	0.0380	0.0383	0.0395
55	0.0173	0.0173	0.0173	0.0176
105	0.0091	0.0090	0.0091	0.0091

존하며 기울기 b의 값에 무관하다는 것이다. 표 4.1부터 표 4.4까지는 몇 개의 k값과 n값에 대하여 네 가지 방법에 의한 $EMSE(\cdot)/\sigma^2$의 값들을 구한 것이다.

위의 표들에서 볼 수 있듯이, 본 논문에서 제시된 GMI와 MREG는 기존의 EC와 MLS보다 효율적이다. 그리고 GMI와 MREG를 비교하면, MREG가 GMI보다 다소 효율적인데, 표본 크기 n이 커짐에 따라 차이가 거의 없어진다.

5. 결 론

본 논문에서는 선형추세를 갖는 모집단의 평균 또는 합계를 효율적으로 추정하기 위한 두 가지 방법을 제시하였다. 제시된 방법들 (GMI와 MREG)은 표집 방법과 모평균 (또는 모합) 추정 방법의 두 단계로 나누어 설명할 수 있다. 먼저, GMI의 표집 방법은 Singh 등 (1968)의 변형계통표집 (MSS)을 일반화한 것으로서, 표본 크기 n이 홀수인 경우 $(n+1)/2$번째 추출 단위를 고정하지 않고 랜덤화함으로써 확률적으로 균형을 이루게 하였다. 결과적으로, 가능한 표본의 가짓수를 다양화한 것이다. 다음으로 GMI의 모평균 또는 모합 추정 방법은 Kim (1998)에서 사용된 보간법을 확장, 적용한 것이다. 한편, MREG는 MSS를 표집 방법으로 하여 표본을 뽑은 뒤 회귀추정법을 사용하여 모평균 또는 모합을 추정하는 것이다.

기대평균제곱오차를 기준으로 할 때, GMI는 Kim (1998)의 추정법과 동등한 효율성을 갖는다. 또한

GMI와 MREG는 랜덤오차항의 분산인 σ^2이 작을수록 전통적인 방법들에 비해 효율적이며, σ^2의 값에 관계없이 Yates (1948)의 EC와 김혁주 (2004)의 MLS보다 효율적인 것으로 나타났다. 또한 MREG는 GMI보다 약간 더 효율적인데, 표본 크기가 증가함에 따라 차이가 거의 없어지는 것으로 밝혀졌다. 따라서 표본이 크면 GMI와 MREG가 실질적으로 차이가 없고, 소표본 (n이 20 미만)인 경우에는 MREG가 GMI보다 선호될 수 있을 것이다.

참고문헌

김영화, 김기수 (2009). 고객집단별 보험금에 대한 소지역 추정. <한국데이터정보과학회지>, **20**, 77-87.

김혁주(2004). 변형된 계통추출과 최소제곱법을 이용한 모평 균 추정. <응용통계연구>, **17**, 105-117.

박종태 (2009). 소지역의 실업률에 대한 상대위험도의 추정에 관한 비교연구. <한국데이터정보과학회지>, **20**, 349-356.

Bellhouse, D. R. and Rao, J. N. K. (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.

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

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,* Series B, **27**, 315-320.

Singh, D., Jindal, K. K. and Garg, J. N. (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.

Journal of the Korean Data & Information Science Society
2009, **20**(6), 1103–1118

한국데이터정보과학회지

Generalization of modified systematic sampling and regression estimation for population with a linear trend [†]

Hyuk Joo Kim[1] · Jeong Hyeon Kim[2]

[1]Division of Mathematics and Informational Statistics, Wonkwang University
[2] Department of Informational Statistics, Wonkwang University

Received 23 April 2009, revised 25 September 2009, accepted 14 October 2009

Abstract

When we wish to estimate the mean or total of a finite population, the numbering of the population units is of importance. In this paper, we have proposed two methods for estimating the mean or total of a population having a linear trend, for the case when the reciprocal of the sampling fraction is an even number and the sample size is an odd number. The first method involves drawing a sample by using a method which is a generalization of Singh et al's (1968) modified systematic sampling, and using interpolation in determining the estimator. The second method involves selecting a sample by modified systematic sampling, and estimating the population parameters by the regression estimation method. Under the criterion of the expected mean square error based on Cochran's (1946) infinite superpopulation model, the proposed methods have been compared with existing methods. We have also made a comparison between the two proposed methods.

Keywords: Estimation of population mean, generalized modified systematic sampling, infinite superpopulation model, interpolation, linear trend, regression estimation.

[†] This paper was supported by Wonkwang University in 2007.
[1] Corresponding author: Professor, Division of Mathematics and Informational Statistics and Institute of Basic Natural Sciences, Wonkwang University, Jeonbuk 570-749, Korea.
E-mail: hjkim@wonkwang.ac.kr
[2] Master, Department of Informational Statistics, Graduate School, Wonkwang University, Jeonbuk 570-749, Korea.