The Korean Communications in Statistics **Vol. 14** No. 3, 2007, pp. 667–677

An Estimator of Population Mean Based on Balanced Systematic Sampling When Both the Sample Size and the Reciprocal of the Sampling Fraction are Odd Numbers*

Hyuk Joo Kim[1)]

Abstract

In this paper, we propose a method for estimating the mean of a population which has a linear trend, when both n, the sample size, and k, the reciprocal of the sampling fraction, are odd numbers. The proposed method, not having the drawbacks of centered systematic sampling, centered modified sampling and centered balanced sampling, consists of selecting a sample by balanced systematic sampling and estimating the population mean by using interpolation. We compare the efficiency of the proposed method and existing methods under the criterion of the expected mean square error based on the infinite superpopulation model.

Keywords: Balanced systematic sampling; estimation of population mean; infinite superpopulation model; interpolation; linear trend.

1. 서론

우리는 종종 표본조사를 통하여 유한모집단의 평균을 효율적으로 추정하는 데에 관심을 갖는다. 그런데 모집단 단위들이 특성값에 무관한 순서로 배열되어 있지 않고 어떤 추세 (trend)를 갖고 있는 경우가 있다. 추세 중 대표적인 것이 선형추세 (linear trend)인데, 이 경우 유용하게 쓰이는 표집 방법이 계통표집 (systematic sampling)이다. 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.

표집률 (sampling fraction)의 역수 k가 짝수이고 표본 크기 (sample size) n이 홀수인 경우, 선형추세를 갖는 모집단의 평균을 추정하는 문제를 Kim (1998, 2000)이 연구하였

* This paper was supported by Wonkwang University in 2006.

1) Professor, Division of Mathematics and Informational Statistics and Institute of Basic Natural Sciences, Wonkwang University, Shinyong-dong 344-2, Iksan, Jeonbuk 570-749, Korea.
E-mail : hjkim@wonkwang.ac.kr

다. MSS 및 BSS와 보간법 (interpolation)을 결합하여 모평균을 효율적으로 추정할 수 있는 것으로 밝혀졌다.

그런데 k가 홀수인 경우에는 Kim (1998, 2000)에서 제시된 방법들의 수학적 내용이 성립하지 않는다. 이때 n이 짝수이면 연구할 가치가 있는 내용이 거의 없으나, n이 홀수이면 모집단의 집락화에서 불균형성이 존재하기 때문에 이를 보정해 주는 방법이 필요하다. 이 경우 수학적으로는 CSS와, Fountain과 Pathak (1989)의 중심변형표집 (centered modified sampling: CMS)과 중심균형표집 (centered balanced sampling: CBS)이 최적인 것으로 나타나지만, CSS, CMS, CBS에는 표본으로 뽑힐 기회가 전혀 없는 모집단 단위들이 존재하며, 상당수의 학자들은 이 점을 큰 단점으로 본다.

이러한 단점을 갖지 않는 방법으로, 김혁주 (2005)는 MSS와 보간법을 결합하여 모평균을 추정하는 방법을 제시하였다. 이 방법은 무한초모집단 모형 (infinite superpopulation model)에서 가정하는 오차항의 분산 σ^2이 작을수록 전통적인 방법들에 비해 효율적인 것으로 밝혀졌다.

그런데 계통표집을 응용하여 모평균을 추정하는 방법에는 크게 나눠 MSS계열과 BSS계열이 있으므로, BSS를 응용한 방법을 연구하여 기존의 방법들과 비교하는 것도 상당한 의미가 있을 것이다. 본 논문에서는 k가 3 이상의 홀수이고 n이 5 이상의 홀수인 경우, CSS, CMS, CBS의 단점을 갖지 않으면서 선형추세라는 특성을 잘 살려서 BSS를 응용하여 모평균을 효율적으로 추정하는 방법을 제시하고자 한다.

2. 모평균 추정 방법

모집단의 크기를 $N = kn$이라 하고 모집단의 단위들을 $U_1, U_2, \ldots, U_N$으로 나타내자. 표본 크기는 n으로 나타낸다.

먼저 BSS는 어떠한 표집 방법인지 간략히 알아보자. BSS는 k개의 집락 (cluster) $S_1', S_2', \ldots, S_k'$중 하나를 등확률 (즉, $1/k$)로 뽑는 방법이다. 그리고 뽑힌 집락의 평균 $\bar{y}_{BSS}$로 모평균을 추정한다. 여기서 집락 $S_i'(i = 1, 2, \ldots, k)$는 다음과 같이 정의된다.

n이 짝수일 때

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

n이 홀수일 때

$$S_i' = \{U_{i+2(j-1)k} \ : j = 1, 2, \ldots, (n+1)/2\} \bigcup \{U_{2jk+1-i} \ : j = 1, 2, \ldots, (n-1)/2\}.$$

예를 들어 $N = 35$, $n = 7$, $k = 5$이면 $S_1' = \{U_1, U_{10}, U_{11}, U_{20}, U_{21}, U_{30}, U_{31}\}$, $S_2' = \{U_2, U_9, U_{12}, U_{19}, U_{22}, U_{29}, U_{32}\}$, $S_3' = \{U_3, U_8, U_{13}, U_{18}, U_{23}, U_{28}, U_{33}\}$, $S_4' = \{U_4, U_7, U_{14}, U_{17}, U_{24}, U_{27}, U_{34}\}$, $S_5' = \{U_5, U_6, U_{15}, U_{16}, U_{25}, U_{26}, U_{35}\}$이다.

이제 본 논문에서 사용될 기호들을 정의한다.

y_i : 모집단 안의 i번째 단위 U_i가 가지고 있는 특성값 $(i = 1, 2, \ldots, N)$,

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

y'_{ij} : S'_i 안의 j번째 단위의 특성값 $(i = 1, 2, \ldots, k\ ;\ j = 1, 2, \ldots, n)$,

즉, n이 짝수일 때

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

n이 홀수일 때

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

$\bar{y}'_i = \dfrac{1}{n}\sum_{j=1}^{n} y'_{ij}$: S'_i 안의 단위들의 특성값의 평균 $(i = 1, 2, \ldots, k)$.

이제 k가 3 이상의 홀수이고 n이 5 이상의 홀수인 경우 모평균 $\bar{Y}$를 추정하는 방법을 제시한다. $N = 35, n = 7, k = 5$인 경우를 다시 예로 들어 생각하자. 우선 $S'_1, S'_2, S'_3, S'_4, S'_5$ 중 하나의 집락을 뽑는다 (확률은 각각 1/5). 즉, BSS로 하나의 집락을 뽑는다. $S'_1, S'_2, S'_3, S'_4, S'_5$안의 단위들의 번호를 합하면 각각 124, 125, 126, 127, 128이다. 이렇게 다섯 집락 사이에는 약간의 차이가 존재하는데, 이것은 표본 크기 n이 홀수이기 때문이다. 모집단에 선형추세가 있는 경우에는 이런 차이를 제거해 주는 것이 좋으며, 따라서 $\bar{Y}$를 추정할 때 단순한 표본평균 $\bar{y}'_i$를 쓰는 것보다 수정된 추정량을 사용하는 것이 좋을 것이다. $S'_1, S'_2, S'_3, S'_4, S'_5$가 뽑힌 경우 각각 $y_{11}, y_{12}, y_{13}, y_{14}, y_{15}$을 $\widehat{y}_{13}$으로 대체하거나, 각각 $y_{21}, y_{22}, y_{23}, y_{24}, y_{25}$를 $\widehat{y}_{23}$으로 대체하면 균형이 이루어진다. 여기서 $\widehat{y}_{13}$은 (i) S'_1, S'_2, S'_4, S'_5 중 하나가 뽑힌 경우 U_{13}의 양쪽으로 인접한 두 단위의 값을 사용하여 보간법으로 y_{13}을 추정한 것이며, (ii) S'_3가 뽑힌 경우에는 y_{13} 자체를 의미하는 것이다. 그리고 $\widehat{y}_{23}$은 바로 앞의 문장에서 13만 23으로 바꿔서 쓰면 정의된다. $\widehat{y}_{13}$과 $\widehat{y}_{23}$을 구체적인 식으로 표시하면 다음과 같다. 여기서 $i = 1, 2$이다.

$$\begin{aligned} \widehat{y}_{10i+3} &= \frac{1}{9}(7y_{10i+1} + 2y_{10i+10}), && (S'_1\text{이 뽑힌 경우}) \\ &= \frac{1}{7}(6y_{10i+2} + y_{10i+9}), && (S'_2\text{이 뽑힌 경우}) \\ &= y_{10i+3}, && (S'_3\text{이 뽑힌 경우}) \\ &= \frac{1}{7}(y_{10i-3} + 6y_{10i+4}), && (S'_4\text{이 뽑힌 경우}) \\ &= \frac{1}{9}(2y_{10i-4} + 7y_{10i+5}), && (S'_5\text{이 뽑힌 경우}) \end{aligned}$$

S_i'가 뽑힌 경우 $\bar{Y}$를 $\bar{y}_i'^*(3)$ 또는 $\bar{y}_i'^*(5)$로 추정하는데, $\bar{y}_i'^*(3)$과 $\bar{y}_i'^*(5)$는 다음과 같다.

$$\begin{aligned}
\bar{y}_1'^*(3) &= \frac{1}{7}(y_1 + y_{10} + \widehat{y}_{13} + y_{20} + y_{21} + y_{30} + y_{31}) \\
&= \bar{y}_1{}' + \frac{2}{63}(y_{20} - y_{11}) = \bar{y}_1{}' + \frac{2}{63}(y_{14}' - y_{13}'), \\
\bar{y}_2'^*(3) &= \bar{y}_2' + \frac{1}{49}(y_{24}' - y_{23}'), \\
\bar{y}_3'^*(3) &= \bar{y}_3', \\
\bar{y}_4'^*(3) &= \bar{y}_4' - \frac{1}{49}(y_{43}' - y_{42}'), \\
\bar{y}_5'^*(3) &= \bar{y}_5' - \frac{2}{63}(y_{53}' - y_{52}'), \\
\bar{y}_1{}'^*(5) &= \frac{1}{7}(y_1 + y_{10} + y_{11} + y_{20} + \widehat{y}_{23} + y_{30} + y_{31}) \\
&= \bar{y}_1{}' + \frac{2}{63}(y_{30} - y_{21}) = \bar{y}_1{}' + \frac{2}{63}(y_{16}' - y_{15}'), \\
\bar{y}_2{}'^*(5) &= \bar{y}_2{}' + \frac{1}{49}(y_{26}' - y_{25}'), \\
\bar{y}_3'^*(5) &= \bar{y}_3', \\
\bar{y}_4'^*(5) &= \bar{y}_4' - \frac{1}{49}(y_{45}' - y_{44}'), \\
\bar{y}_5'^*(5) &= \bar{y}_5' - \frac{2}{63}(y_{55}' - y_{54}').
\end{aligned}$$

위의 식들에서 괄호안의 3과 5는 각각 집락의 셋째 원소와 다섯째 원소를 중심으로 보간법이 이루어졌다는 것을 나타내는 숫자이다.

이상의 내용을 일반화해 보자. k가 3 이상의 홀수이고 n이 5 이상의 홀수인 경우 BSS에 의해 하나의 집락을 뽑는다. 즉, 각각 $1/k$의 확률로 $S_1', S_2', \ldots, S_k'$ 중 하나를 뽑는다. 뽑힌 집락이 S_i'이면 모평균 $\bar{Y}$는 $\bar{y}_i'^*(3), \bar{y}_i'^*(5), \ldots, \bar{y}_i'^*(n-2)$ 중 하나로 추정된다 (확률은 각각 $2/(n-3)$). 여기서 $i = 1, 2, \ldots, (k-1)/2$인 경우

$$\bar{y}_i'^*(m) = \bar{y}_i' + \frac{k+1-2i}{2n(2k+1-2i)}(y_{i,m+1}' - y_{i,m}') \quad (m = 3, 5, \ldots, n-2) \tag{2.1}$$

이고

$$\bar{y}_{(k+1)/2}'^*(m) = \bar{y}_{(k+1)/2}' \quad (m = 3, 5, \ldots, n-2) \tag{2.2}$$

이며, $i = (k+3)/2, (k+5)/2, \ldots, k$인 경우

$$\bar{y}_i'^*(m) = \bar{y}_i' - \frac{2i-k-1}{2n(2i-1)}(y_{i,m}' - y_{i,m-1}') \quad (m = 3, 5, \ldots, n-2) \tag{2.3}$$

이다.

이렇게 $\bar{Y}$를 추정하는 방법을 BI로 나타내고 BI에 의한 $\bar{Y}$의 추정량을 $\bar{y}_{BI}$로 나타내자. 즉

$$P(\bar{y}_{BI} = \bar{y_i}'^*(m)) = \frac{2}{k(n-3)} \quad (i = 1, 2, \ldots, k\ ;\ m = 3, 5, \ldots, n-2) \tag{2.4}$$

이다. $\bar{y}_{BI}$는 $\bar{Y}$의 편향추정량 (biased estimator)이며, 편향

$$B(\bar{y}_{BI}) = \frac{2}{k(n-3)} \sum_{i=1}^{k} \sum_{m} \bar{y}_i'^*(m) - \bar{Y} \tag{2.5}$$

와 평균제곱오차

$$MSE(\bar{y}_{BI}) = \frac{2}{k(n-3)} \sum_{i=1}^{k} \sum_{m} \left\{ \bar{y}_i'^*(m) - \bar{Y} \right\}^2 \tag{2.6}$$

를 갖는다. 여기서 $\sum_m$은 $m = 3, 5, \ldots, n-2$에 대해 합한 것을 의미한다.

3. $\bar{y}_{BI}$의 기대평균제곱오차

추세를 갖는 모집단에 관한 추론의 효율성을 논할 때 훌륭한 이론적 근거로서 사용되는 것이 Cochran (1946)의 무한초모집단 모형이다. 이 모형은 주어진 유한모집단을 무한초모집단으로부터 뽑힌 하나의 표본으로 간주하는 것으로서 다음 식으로 표시된다.

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

여기서 μ_i는 i의 함수이며, e_i는 랜덤오차항으로서 $E(e_i) = 0,\ E(e_i^2) = \sigma^2, E(e_i e_j) = 0$ ($i \neq j$ 일 때)이다. E는 무한초모집단에 걸친 기대값을 의미한다.

이제부터 μ에 관해서도 y에 관해서 정의된 것과 같은 양식의 기호를 사용하기로 한다. 예컨대

$$\begin{aligned}
\bar{\mu} &= \frac{1}{N} \sum_{i=1}^{N} \mu_i, \\
\mu'_{ij} &= \mu_{i+(j-1)k} \quad (j = 1, 3, 5, \ldots, n)(n : \text{홀수}), \\
\bar{\mu_i}' &= \frac{1}{n} \sum_{j=1}^{n} \mu'_{ij}, \\
\bar{\mu}_i'^*(m) &= \bar{\mu_i}' + \frac{k+1-2i}{2n(2k+1-2i)} (\mu'_{i,m+1} - \mu'_{i,m}) \quad (i = 1, 2, \ldots, (k-1)/2)
\end{aligned}$$

등이다. 이러한 기호들과 위의 가정들, 그리고 식 (2.5)와 (2.6)을 이용하면 다음 정리를 얻는다. 증명은 Kim (2000)에 있는 정리 1의 증명과 유사한 방법으로 할 수 있으므로 생략한다.

정리 3.1 모형 (3.1)을 가정할 때, $\bar{y}_{BI}$의 편향과 평균제곱오차의 기대값은 다음과 같다.

$$EB(\bar{y}_{BI}) = \frac{1}{k}\sum_{i=1}^{k}\sum_{m}\bar{\mu}_i^{'*}(m) - \bar{\mu}, \tag{3.2}$$

$$EMSE(\bar{y}_{BI}) = \frac{2}{k(n-3)}\sum_{i=1}^{k}\sum_{m}\left\{\bar{\mu}_i^{'*}(m) - \bar{\mu}\right\}^2 + Q + \frac{\sigma^2}{2nN}(k-1-4kA_k+2k^2B_k)$$
$$(k:3\text{이상의 홀수}, n:5\text{이상의 홀수}) \tag{3.3}$$

단, 여기서

$$\begin{aligned}
Q &= \frac{\sigma^2}{n}\frac{N-n}{N},\\
A_k &= \sum_{i=1}^{(k-1)/2}\frac{1}{2k+1-2i} = \frac{1}{2}\left\{\psi\left(k+\frac{3}{2}\right) - \psi\left(\frac{k+2}{2}\right)\right\} - \frac{1}{2k+1},\\
B_k &= \sum_{i=1}^{(k-1)/2}\frac{1}{(2k+1-2i)^2} = -\frac{1}{4}\left\{\psi^{(1)}\left(k+\frac{3}{2}\right) - \psi^{(1)}\left(\frac{k+2}{2}\right)\right\} - \frac{1}{(2k+1)^2},\\
\psi(x) &= \frac{d}{dx}\ln\Gamma(x)(x>0) : \text{polygamma 함수},\\
\Gamma(x) &= \int_0^\infty t^{x-1}e^{-t}dt\ (x>0) : \text{gamma 함수},\\
\psi^{(1)}(x) &= \frac{d}{dx}\psi(x)
\end{aligned}$$

이다.

이제 모집단에 선형추세가 존재하는 경우를 생각하자. 선형추세는 식 (3.1)에서 $\mu_i = a + bi$ (a와 b는 상수; $b \neq 0$)인 경우를 말한다. 즉 가정된 무한초모집단 모형이

$$y_i = a + bi + e_i \quad (i = 1, 2, \ldots, N) \tag{3.4}$$

인 경우이다. 이 경우 $\bar{\mu} = a + (b/2)(N+1)$임을 쉽게 알 수 있으며 또한 $i = 1, 2, \ldots, k$와 $m = 3, 5, \ldots, n-2$에 대해 $\bar{\mu}_i^{'*}(m) = a + (b/2)(N+1)$이라는 것을 유도할 수 있으므로 정리 3.1에 의하여 다음 정리를 얻는다.

정리 3.2 모집단이 식 (3.4)의 선형추세를 가질 때, $\bar{y}_{BI}$의 기대편향과 기대평균제곱오차는 다음과 같다.

$$EB(\bar{y}_{BI}) = 0, \tag{3.5}$$

$$EMSE(\bar{y}_{BI}) = Q + \frac{\sigma^2}{2nN}(k-1-4kA_k+2k^2B_k),$$
$$(k:3\text{이상의 홀수}, n:5\text{이상의 홀수}). \tag{3.6}$$

4. 기존의 방법들과의 효율성 비교

이 절에서는 본 논문에서 제시된 방법인 BI의 효율성을 기존의 여러 방법과 비교해 보자. 모평균 $\bar{Y}$를 추정하는 방법으로 크게 나눠 두 종류가 있다. 첫째는 표본의 단순평균으로 $\bar{Y}$를 추정하는 방법이며, 둘째는 표본의 단순평균이 아니라 가중평균으로 $\bar{Y}$를 추정하는 방법이다.

4.1. 표본의 단순평균으로 모평균을 추정하는 방법들과의 비교

표본의 단순평균으로 $\bar{Y}$를 추정하는 방법들로는 단순임의표집 (simple random sampling: SRS), 층화임의표집 (stratified random sampling: StRS), 보통의 계통표집 (OSS), 균형계통표집 (BSS), 변형계통표집 (MSS) 등이 있다. 여기서 고려하는 StRS는 j번째 층 ($j = 1,2,\ldots,n$)이 k개의 단위 $U_{1+(j-1)k}, U_{2+(j-1)k}, \ldots, U_{jk}$로 이루어지도록 모집단을 n개의 층으로 분할한 후 각 층에서 1개씩의 단위를 랜덤하게 뽑는 방법이다. 모형 (3.4) 하에서, k와 n이 홀수인 경우 전통적인 방법들에 의한 $\bar{Y}$의 추정량들의 기대평균제곱오차의 공통적인 형태는 다음 식과 같다.

$$EMSE(\widehat{\bar{Y}}) = Q + b^2 f(n,k). \tag{4.1}$$

여기서 $f(n,k)$는 n과 k의 함수로서 구체적으로는

$$\begin{aligned} f(n,k) &= \frac{1}{12}(N+1)(k-1), \quad (SRS\text{의 경우}) \\ &= \frac{1}{12n}(k+1)(k-1), \quad (StRS\text{의 경우}) \\ &= \frac{1}{12}(k+1)(k-1), \quad (OSS\text{의 경우}) \\ &= \frac{1}{12n^2}(k+1)(k-1), \quad (BSS, MSS\text{의 경우}) \end{aligned} \tag{4.2}$$

이다. 이 방법들의 경우 선형추세의 기울기 b의 값이 기대평균제곱오차에 중요한 영향을 미친다는 것을 식 (4.1)에서 알 수 있다.

식 (3.6), (4.1), (4.2)를 근거로 하여 BI를 포함한 여러 방법들을 비교한 결과를 다음과 같이 정리하였다. 표현의 간편성을 위하여 $EMSE(\bar{y}_{BI}), EMSE(\bar{y}_{BSS})$ 등을 각각 BI, BSS 등으로 나타낸다. 따라서, 예컨대 "BI<BSS"라는 표현은 BI가 BSS보다 기대평균제곱오차의 관점에서 효율적이라는 것을 뜻한다.

정리 4.1 모집단이 식 (3.4)의 선형추세를 가질 때, 다음 내용이 성립한다. 여기서 $S_k = k - 1 - 4kA_k + 2k^2B_k$이다.

(i) $\sigma^2 < b^2k(k^2-1)/6S_k$이면 $BI < MSS = BSS < StRS < OSS < SRS$

(ii) $b^2k(k^2-1)/6S_k \le \sigma^2 < b^2N(k^2-1)/6S_k$이면 $MSS = BSS \le BI < StRS < OSS < SRS$

(iii) $b^2N(k^2-1)/6S_k \le \sigma^2 < b^2nN(k^2-1)/6S_k$이면 $MSS = BSS < StRS \le BI < OSS < SRS$

(iv) $b^2nN(k^2-1)/6S_k \le \sigma^2 < b^2nN(N+1)(k-1)/6S_k$이면 $MSS = BSS < StRS < OSS \le BI < SRS$

(v) $b^2nN(N+1)(k-1)/6S_k \le \sigma^2$이면 $MSS = BSS < StRS < OSS < SRS \le BI$

예제 4.1 $N = 135$개의 단위 중에서 $n = 15$개의 단위를 뽑는 경우를 생각해 보자. 이때 $k = 9$가 된다. 선형추세의 기울기는 $b = 0.7$이라 하자. 절편 a는 임의의 상수이다. 이때 $A_9 = 0.2933, B_9 = 0.0221, S_9 = 1.0184$이며, 여러 방법들의 효율성을 비교하면 아래와 같다.

(i) $\sigma^2 \le 57.7392$이면 $BI < MSS = BSS < StRS < OSS < SRS$

(ii) $57.7392 \le \sigma^2 < 866.0887$이면 $MSS = BSS \le BI < StRS < OSS < SRS$

(iii) $866.0887 \le \sigma^2 < 12991.3305$이면 $MSS = BSS < StRS \le BI < OSS < SRS$

(iv) $12991.3305 \le \sigma^2 < 176682.0944$이면 $MSS = BSS < StRS < OSS \le BI < SRS$

(v) $176682.0944 \le \sigma^2$이면 $MSS = BSS < StRS < OSS < SRS \le BI$

위의 예에서 볼 수 있듯이, σ^2이 터무니없이 크지만 않으면 BI는 매우 효율적인 방법이 된다 ((i)과 (ii)의 경우). 사실 σ^2이 너무 크면 선형추세의 의미가 거의 없어지므로 논의할 필요가 거의 없어진다.

4.2. 표본의 가중평균으로 모평균을 추정하는 방법들과의 비교

k와 n이 모두 홀수인 경우 표본의 가중평균으로 모평균 $\bar{Y}$를 추정하는 방법으로 Yates (1948)의 끝값수정법 (end corrections; EC)과, 1절에서 언급한 김혁주 (2005)의 방법이 있다. 김혁주의 방법은 MSS와 보간법을 결합한 것으로 편의상 MI로 표기한다. EC에 의한 $\bar{Y}$의 추정량 $\bar{y}_{EC}$와 MI에 의한 $\bar{Y}$의 추정량 $\bar{y}_{MI}$의 기대평균제곱오차는 각각 다음과 같이 알려져 있다.

$$EMSE(\bar{y}_{EC}) = Q + \frac{\sigma^2(k^2-1)}{6k^2(n-1)^2}, \tag{4.3}$$

$$EMSE(\bar{y}_{MI}) = Q + \frac{\sigma^2}{12N^2}\left\{(k-1)(4k+1) - 12k^2A_k + 6k^3B_k\right\}. \tag{4.4}$$

본 논문에서 제시된 방법인 BI를 EC 및 MI와 비교한 결과는 다음의 정리에 주어진다.

정리 4.2 모집단이 식 (3.4)의 선형추세를 가질 때, 다음의 대소 관계가 성립한다.

$$EMSE(\bar{y}_{BI}) < EMSE(\bar{y}_{MI}) < EMSE(\bar{y}_{EC}),$$
$$(k: 3\text{이상의 홀수}, n: 5\text{이상의 홀수}) \tag{4.5}$$

증명:

(i) (BI와 MI의 비교)

식 (3.6)과 식 (4.4)에 의해

$$\begin{aligned} EMSE(\bar{y}_{MI}) - EMSE(\bar{y}_{BI}) &= \frac{\sigma^2}{12N^2}\left\{12k^2A_k - 6k^3B_k - (2k-1)(k-1)\right\} \\ &= \frac{\sigma^2}{12n^2}\left(12A_k - 6kB_k - 2 + \frac{3}{k} - \frac{1}{k^2}\right) \end{aligned}$$

이다. 그런데

$$\begin{aligned} A_k &= \sum_{i=1}^{(k-1)/2} \frac{1}{2k+1-2i} = \sum_{i=(k+3)/2}^{k} \frac{1}{2i-1}, \\ B_k &= \sum_{i=1}^{(k-1)/2} \frac{1}{(2k+1-2i)^2} = \sum_{i=(k+3)/2}^{k} \frac{1}{(2i-1)^2} \end{aligned}$$

이므로, 괄호 안의 식은

$$\begin{aligned} &12A_k - 6kB_k - 2 + \frac{3}{k} - \frac{1}{k^2} \\ &= \sum_{i=(k+3)/2}^{k} \frac{12}{2i-1} - \sum_{i=(k+3)/2}^{k} \frac{6k}{(2i-1)^2} - \frac{(2k-1)(k-1)}{k^2} \\ &= \sum_{i=(k+3)/2}^{k} \left\{\frac{12}{2i-1} - \frac{6k}{(2i-1)^2} - \frac{2(2k-1)}{k^2}\right\} \\ &= \sum_{i=(k+3)/2}^{k} \frac{(-16k+8)i^2 + (24k^2+16k-8)i - 6k^3 - 12k^2 - 4k + 2}{k^2(2i-1)^2} \end{aligned}$$

이 된다. 여기서 분자만 생각하면, i^2항의 계수가 음수이므로 최소값은 $i=(k+3)/2$ 또는 $i=k$에서 생기는데, 그때의 값은 $2k^3+10k^2-8k+8$ ($i=(k+3)/2$일 때)과 $2k^3+12k^2-12k+2$($i=k$일 때)이다. 이들은 $k \geq 3$일 때 모두 양수이므로 $12A_k - 6kB_k - 2 + 3/k - 1/k^2 > 0$이라는 것을 알 수 있다. 따라서 $EMSE(\bar{y}_{MI}) > EMSE(\bar{y}_{BI})$라는 결론이 나온다.

(ii) (MI와 EC의 비교)

식 (4.3)과 (4.4)에 의해

$$\begin{aligned} &EMSE(\bar{y}_{EC}) - EMSE(\bar{y}_{MI}) \\ &= \frac{\sigma^2(k^2-1)}{6k^2(n-1)^2} - \frac{\sigma^2}{12N^2}\left\{(k-1)(4k+1) - 12k^2A_k + 6k^3B_k\right\} \\ &= \frac{\sigma^2(k^2-1)}{6k^2(n-1)^2} - \frac{\sigma^2(k^2-1)}{6k^2n^2} + \frac{\sigma^2(k^2-1)}{6N^2} \end{aligned}$$

$$-\frac{\sigma^2}{12N^2}\left\{(k-1)(4k+1)-12k^2A_k+6k^3B_k\right\}$$
$$=\frac{\sigma^2(k^2-1)}{6k^2}\left\{\frac{1}{(n-1)^2}-\frac{1}{n^2}\right\}+\frac{\sigma^2}{12n^2}\left(12A_k-6kB_k-2+\frac{3}{k}-\frac{1}{k^2}\right)$$

이다. 그런데 여기서 첫째 항은 0보다 크다는 것이 명백하고, 둘째 항의 괄호 안에 있는 식은 이미 (i)에서 0보다 크다는 것이 증명되었다. 따라서 $EMSE(\bar{y}_{EC}) > EMSE(\bar{y}_{MI})$임을 알 수 있다.

(i)과 (ii)를 종합하여 식 (4.5)가 증명되었다. □

k의 여러 값에 대하여, $\bar{y}_{EC}$, $\bar{y}_{MI}$, $\bar{y}_{BI}$의 기대평균제곱오차를 나타내는 식 (4.3), (4.4), (3.6)의 둘째 항의 값들을 비교한 내용이 표 4.1에 주어져 있다.

표 4.1: EC, MI, BI의 비교 (기대평균제곱오차의 둘째 항의 값)

k	EC	MI	BI
3	$0.1481\sigma^2/(n-1)^2$	$0.1007\sigma^2/n^2$	$0.0533\sigma^2/n^2$
5	$0.1600\sigma^2/(n-1)^2$	$0.1079\sigma^2/n^2$	$0.0558\sigma^2/n^2$
7	$0.1633\sigma^2/(n-1)^2$	$0.1098\sigma^2/n^2$	$0.0564\sigma^2/n^2$
9	$0.1646\sigma^2/(n-1)^2$	$0.1106\sigma^2/n^2$	$0.0566\sigma^2/n^2$
11	$0.1653\sigma^2/(n-1)^2$	$0.1110\sigma^2/n^2$	$0.0567\sigma^2/n^2$
13	$0.1657\sigma^2/(n-1)^2$	$0.1112\sigma^2/n^2$	$0.0567\sigma^2/n^2$
15	$0.1659\sigma^2/(n-1)^2$	$0.1113\sigma^2/n^2$	$0.0568\sigma^2/n^2$
25	$0.1664\sigma^2/(n-1)^2$	$0.1116\sigma^2/n^2$	$0.0568\sigma^2/n^2$
∞	$0.1667\sigma^2/(n-1)^2$	$0.1118\sigma^2/n^2$	$0.0569\sigma^2/n^2$

5. 결론

추세를 갖는 모집단에 관하여 추론할 때 이 추세의 특성을 잘 활용하면 모수를 좀 더 효율적으로 추정할 수 있다. 본 논문에서는, 모집단에 선형추세가 존재하는 경우 표본크기 n과 표집률의 역수 k가 모두 홀수일 때 모평균 $\bar{Y}$를 추정하는 새로운 방법 (BI)을 제시하였다. 이 방법은 Sethi (1965)와 Murthy (1967)가 제시한 BSS로 표본을 뽑은 뒤 보간법을 써서 표본평균 $\bar{y}_{BSS}$보다 수정된 추정량인 $\bar{y}_{BI}$를 사용하여 모평균을 추정하는 것이다. 따라서 BI는 모집단의 모든 단위들의 표본추출 확률이 동일하게 표본을 뽑은 뒤 표본의 단순평균이 아닌 가중평균으로 모평균을 추정하는 방법이다.

Cochran (1946)의 무한초모집단 모형에 근거를 둔 기대평균제곱오차를 기준으로 하여 BI와 기존 방법들의 효율성을 비교하였다. 그 결과 BI는 오차항의 분산 σ^2이 작을수록 (즉, 선형추세가 강할수록), 표본의 단순평균으로 모평균을 추정하는 방법들에 비해 효율적이라는 것이 밝혀졌으며, σ^2이 아주 크지 않은 대부분의 현실적인 경우에 이 방법들보다 효율적인 방법인 것으로 나타났다. 한편, 표본의 가중평균으로 모평균을 추정

하는 기존 방법들과 비교한 결과, BI는 Yates (1948)의 EC와 김혁주 (2005)의 MI보다 효율적인 것으로 밝혀졌다.

참고문헌

김혁주 (2005). 표본의 크기와 표집률의 역수가 모두 홀수인 경우 변형계통표집과 보간법을 이용한 모평균 추정. *Journal of the Korean Data Analysis Society,* **7,** 2057–2066.

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.

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.

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

Kim, H. J. (2000). Estimation of mean using balanced systematic sampling and interpolation for population with linear trend. *Journal of the Korean Statistical Society,* **29,** 455–471.

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. *Sankhyā,* Ser. 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,* Ser. A, **241,** 345–377.

[Received September 2007, Accepted November 2007]