응용통계연구 제 13권 2호, 2000년, pp. 457-476

선형추세를 갖는 모집단에 대한 효율적인 모평균 추정 : 계통추출의 확장 *

김혁주[1] 석은양[2]

요 약

본 연구에서는 선형추세를 갖는 모집단에 대한 효율적인 표본추출방법과 모평균 추정법을 제안하였다. 이 방법은 계통추출을 확장한 중심균형계통추출을 써서 표본을 뽑은 뒤 표본평균보다 수정된 추정량을 써서 모평균을 추정하는 것이다. 수정된 추정량을 정하는 데에 보간법의 개념을 사용하였다. 제안된 추정량과 기존의 방법에 의한 추정량들의 효율을 Cochran(1946)의 무한초모집단모형에 근거를 둔 기대평균제곱오차를 기준으로 하여 비교하였다. 제안된 방법은 표본크기 $n(\geq 5)$이 홀수이고 추출률의 역수인 k가 짝수인 경우에 사용하기 위한 것이다. 모의실험을 이용한 예에서도 역시 좋은 결과가 얻어졌다.

주요용어: 선형추세, 중심균형계통추출, 보간법, 무한초모집단모형.

1. 서론

통계조사에서 대하게 되는 모집단이 선형추세(linear trend)를 가지고 있는 경우가 종종 있다. 예를 들어, 주어진 기간에 대하여 어떤 도시 안에 있는 백화점들의 평균 매출액을 추정하고자 하는 경우를 가정해 보자. 만일 그 도시 안의 백화점들이 종업원의 수에 따라 증가 또는 감소하는 순서로 번호를 부여받는다면, 이 모집단에는 직선에 가까운 모양을 갖는 추세가 존재할 것이라는 것을 쉽게 추측할 수 있다.

선형추세를 갖는 모집단의 평균을 추정하는 데에 있어서 보통의 계통추출(ordinary systematic sampling : OSS)은 단순임의추출(simple random sampling : SRS)보다 훨씬 좋은 결과를 주는 것으로 알려져 있다. 또한 OSS로부터 파생된 여러 표본추출방법과 추정법들이 많은 연구자들에 의해 연구되어 왔다. Yates(1948)는 선형추세를 갖는 모집단에 대하여 OSS에 근거한 끝값수정법(end corrections)을 제안하였다. 이 방법은 선형추세를 갖는 모집단의 특성을 잘 살린 방법이자 효율적인 방법으로 인정받아 왔다. 선형추세를 갖는 모집단에 대한 표본추출방법으로 이 밖에도 Madow(1953)의 중심계통추출(centered systematic sampling : CSS), Sethi(1965)와 Murthy(1967)의 균형계통추출(balanced systematic

* 본 연구는 한국과학재단의 1998년도 목적기초연구 (981-0105-034-1)지원으로 수행되었습니다.

1) (570-749) 전북 익산시 신용동 344-2, 원광대학교 자연과학대학 수리과학부, 부교수
E-mail: hjkim@wonnms.wonkwang.ac.kr
2) (570-749) 전북 익산시 신용동 344-2, 원광대학교 자연과학대학 수리과학부, 강사
E-mail: seok1221@hanmail.net

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.

한편 Cochran(1946)은 여러 방법들의 효율성을 연구하는 데에 훌륭한 이론적 근거가 되는 무한초모집단모형(infinite superpopulation model)을 소개하였으며, Bellhouse와 Rao(1975)는 OSS, CSS, BSS, MSS의 효율성을 비교하고 논의하였다. Iachan(1982)은 Buckland(1951)의 논의 이후의 계통추출의 발전 과정을 광범위하게 논의하였다.

본 논문에서는 선형추세를 갖는 모집단의 평균을 효율적으로 추정하기 위한 새로운 방법이 제안된다. 이 방법은 표본크기 $n(\geq 5)$이 홀수이고 추출률의 역수인 k가 짝수인 경우에 사용하기 위한 것이다. 본 논문에서 제안되는 방법과 기존의 여러 방법들과의 비교가 많은 관심의 대상이 되는데, 이러한 비교는 Cochran(1946)이 소개한 무한초모집단모형에 바탕을 둔 기대평균제곱오차(expected mean square error)를 기준으로 하여 이루어진다.

2. 모집단 단위들의 집락화와 기호의 정의

크기가 $N = kn$인 모집단을 생각하자. 모집단을 구성하는 N개의 단위에는 1부터 N까지의 번호가 붙어 있다고 하고, 이 N개의 단위들을 $U_1, U_2, \cdots, U_N$으로 나타내기로 하자. 이 모집단으로부터 크기 n인 표본을 뽑으려 한다.

이제 여러 표본추출방법들과 밀접한 관련이 있으며 아주 중요한 의미를 갖는, 모집단 단위들의 집락화(clustering)에 관하여 생각하기로 한다. 다음과 같은 세 가지 방법의 집락화를 생각하자. 각각의 경우에 모집단은 n 개씩의 단위들을 갖는 k 개의 집락으로 나누어진다.

1) $S_1, S_2, \cdots, S_k$
   단, $S_i = \{U_{i+(j-1)k} \ : \ j = 1, 2, \cdots, n\}$ $(i = 1, 2, \cdots, k)$
2) $S'_1, S'_2, \cdots, S'_k$
   단, n이 짝수일 때
   $S'_i = \{U_{i+(j-1)k} \ : \ j = 1, 2, \cdots, n/2\} \bigcup \{U_{N+1-i-(j-1)k} \ : \ j = 1, 2, \cdots, n/2\}$
   $(i = 1, 2, \cdots, k)$
   n이 홀수일 때
   $S'_i = \{U_{i+(j-1)k} \ : \ j = 1, 2, \cdots, (n+1)/2\} \bigcup \{U_{N+1-i-(j-1)k} \ : \ j = 1, 2, \cdots, (n-1)/2\}$
   $(i = 1, 2, \cdots, k)$
3) $S''_1, S''_2, \cdots, S''_k$
   단, n이 짝수일 때
   $S''_i = \{U_{i+2(j-1)k} \ : \ j = 1, 2, \cdots, n/2\} \bigcup \{U_{2jk+1-i} \ : \ j = 1, 2, \cdots, n/2\} (i = 1, 2, \cdots, k)$,
   n이 홀수일 때

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

예제 2.1: N=20, n=5이고 따라서 k=4인 경우를 생각하자. 이 경우 위에서 설명한 세 가지 방법의 집락화는 다음과 같다.

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

이제 본 논문에서 사용될 기호들을 정의한다.

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

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

y_{ij} : S_i 안의 j번째 단위의 특성값 $(i = 1, 2, \cdots, k \ ; \ j = 1, 2, \cdots, n)$
즉, $y_{ij} = y_{i+(j-1)k}$

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

y_{ij}' : S_i' 안의 j번째 단위의 특성값 $(i = 1, 2, \cdots, k \ ; \ j = 1, 2, \cdots, n)$
즉, n이 짝수일 때
$y_{ij}' = y_{i+(j-1)k}$ $(j = 1, 2, \cdots, n/2)$
$y_{ij}' = y_{N+1-i-(n-j)k} = y_{1-i+jk}$ $(j = n/2+1, n/2+2, \cdots, n)$
n이 홀수일 때
$y_{ij}' = y_{i+(j-1)k}$ $(j = 1, 2, \cdots, (n-1)/2, (n+1)/2)$
$y_{ij}' = y_{N+1-i-(n-j)k} = y_{1-i+jk}$ $(j = (n+3)/2, (n+5)/2, \cdots, n)$

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

y_{ij}'' : S_i'' 안의 j번째 단위의 특성값 $(i = 1, 2, \cdots, k \ ; \ j = 1, 2, \cdots, n)$
즉, n이 짝수일 때
$y_{ij}'' = y_{i+(j-1)k}$ $(j = 1, 3, 5, \cdots, n-1)$
$y_{ij}'' = y_{1-i+jk}$ $(j = 2, 4, 6, \cdots, n)$
n이 홀수일 때
$y_{ij}'' = y_{i+(j-1)k}$ $(j = 1, 3, 5, \cdots, n)$
$y_{ij}'' = y_{1-i+jk}$ $(j = 2, 4, 6, \cdots, n-1)$

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

위에서 각각의 $y_i(i=1,2,\cdots,N)$는 2차원 첨자를 써서 세 가지 방법으로 나타내질 수 있음을 볼 수 있다. 예컨대 앞의 예제 2.1에서 다음과 같은 네 가지 표현은 모두 같은 것을 나타낸다.

$$y_8,\ y_{42},\ y'_{42},\ y''_{12}$$

여기서 첨자 8은 1차원 첨자이고 나머지 첨자들은 2차원 첨자이다.

3. 여러 가지의 표본추출방법들과 추정방법들에 관한 고찰

크기 $N=kn$인 모집단으로부터 크기 n인 표본을 뽑는 방법은 여러 가지가 있다. 그 중 선형추세를 갖는 모집단의 경우에 쓰일 수 있는 기존의 몇 가지 방법들에 관하여 간단히 살펴보기로 한다.

1) 보통의 계통추출 (ordinary systematic sampling: OSS)

이 방법은 2절에서 정의된 S_i들 중 하나를 무작위로 뽑아 표본평균으로 모평균 $\bar{Y}$를 추정하는 방법이다. 표본평균 $\bar{y}_{sy}$는 $\bar{Y}$의 비편향추정량이며 분산

$$V(\bar{y}_{sy})=\frac{1}{k}\sum_{i=1}^{k}(\bar{y}_i-\bar{Y})^2 \tag{3.1}$$

을 갖는다.

2) 끝값수정법 (end corrections: EC) (Yates, 1948)

이 방법은 OSS와 동일한 방법으로 표본을 뽑지만, 표본의 첫단위와 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 추정한다. 즉, 뽑히는 집락을 S_i라 하면 $\bar{Y}$의 추정량은 다음과 같다.

$$\begin{aligned}\bar{y}_{ec} &= \left\{\frac{1}{n}+\frac{2i-k-1}{2k(n-1)}\right\}y_{i1}+\frac{1}{n}y_{i2}+\cdots+\frac{1}{n}y_{i,\,n-1}+\left\{\frac{1}{n}-\frac{2i-k-1}{2k(n-1)}\right\}y_{in}\\ &= \bar{y}_i+\frac{2i-k-1}{2k(n-1)}(y_{i1}-y_{in})\end{aligned} \tag{3.2}$$

이 $\bar{y}_{ec}$는 $\bar{Y}$의 편향추정량이라는 것을 쉽게 보일 수 있다. $\bar{y}_{cc}$의 편향(bias)과 평균제곱오차(mean square error)는 각각 다음과 같다.

$$B(\bar{y}_{cc})=\frac{1}{k}\sum_{i=1}^{k}\left\{(\bar{y}_i-\bar{Y})+\frac{2i-k-1}{2k(n-1)}(y_{i1}-y_{in})\right\} \tag{3.3}$$

$$MSE(\bar{y}_{cc})=\frac{1}{k}\sum_{i=1}^{k}\left\{(\bar{y}_i-\bar{Y})+\frac{2i-k-1}{2k(n-1)}(y_{i1}-y_{in})\right\}^2 \tag{3.4}$$

3) 변형계통추출 (modified systematic sampling: MSS) (Singh et al., 1968)

S_i'들 중 하나를 무작위로 뽑아 표본평균 $\bar{y}_{mod}$로 $\bar{Y}$를 추정한다. $\bar{y}_{mod}$는 $\bar{Y}$의 비편향추정량이며, 분산

$$V(\bar{y}_{mod}) = \frac{1}{k}\sum_{i=1}^{k}(\bar{y}_i' - \bar{Y})^2 \tag{3.5}$$

을 갖는다.

4) 균형계통추출 (balanced systematic sampling: BSS) (Sethi, 1965; Murthy, 1967)

S_i'' 들 중 하나를 무작위로 뽑아 표본평균 $\bar{y}_{bal}$로 $\bar{Y}$를 추정한다. $\bar{y}_{bal}$은 $\bar{Y}$의 비편향추정량이며 , 분산

$$V(\bar{y}_{bal}) = \frac{1}{k}\sum_{i=1}^{k}(\bar{y}_i'' - \bar{Y})^2 \tag{3.6}$$

을 갖는다.

5) 중심계통추출 (centered systematic sampling: CSS) (Madow, 1953)

k가 홀수이면 집락 $S_{(k+1)/2}$를 추출한다. $\bar{Y}$의 추정량 $\bar{y}_{ccn}$은 $\bar{y}_{(k+1)/2}$이며, 편향

$$B(\bar{y}_{ccn}) = \bar{y}_{(k+1)/2} - \bar{Y} \tag{3.7}$$

과 평균제곱오차

$$MSE(\bar{y}_{ccn}) = (\bar{y}_{(k+1)/2} - \bar{Y})^2 \tag{3.8}$$

을 갖는다.

k 가 짝수이면 두 집락 $S_{k/2}$와 $S_{k/2+1}$ 중 하나를 뽑는다(확률은 각각 1/2). $\bar{y}_{ccn}$은 $\bar{y}_{k/2}$ 또는 $\bar{y}_{k/2+1}$이 되며, 다음과 같은 편향과 평균제곱오차를 갖는다.

$$B(\bar{y}_{ccn}) = \frac{1}{2}(\bar{y}_{k/2} + \bar{y}_{k/2+1}) - \bar{Y} \tag{3.9}$$

$$MSE(\bar{y}_{ccn}) = \frac{1}{2}\left\{(\bar{y}_{k/2} - \bar{Y})^2 + (\bar{y}_{k/2+1} - \bar{Y})^2\right\} \tag{3.10}$$

6) 중심변형계통추출 (centered modified systematic sampling: CMSS) (Kim, 1985)

이 방법은 MSS와 CSS의 개념을 결합한 것으로, MSS를 "중심화"한 것이라고 할 수 있다.

k가 홀수이면 집락 $S'_{(k+1)/2}$을 뽑는다. 그런데 이 경우 $S'_{(k+1)/2}$은 $S_{(k+1)/2}$과 정확히 일치하므로 CMSS는 CSS와 같은 방법이 된다.

k가 짝수이면 각각 1/2의 확률로 두 집락 $S'_{k/2}$와 $S'_{k/2+1}$ 중 하나를 뽑는다. $\bar{Y}$의 추정량 $\bar{y}_{cm}$은 $\bar{y}'_{k/2}$ 또는 $\bar{y}'_{k/2+1}$이 되며, 편향

$$B(\bar{y}_{cm}) = \frac{1}{2}(\bar{y}'_{k/2} + \bar{y}'_{k/2+1}) - \bar{Y} \tag{3.11}$$

와 평균제곱오차

$$MSE(\bar{y}_{cm}) = \frac{1}{2}\left\{(\bar{y}'_{k/2} - \bar{Y})^2 + (\bar{y}'_{k/2+1} - \bar{Y})^2\right\} \tag{3.12}$$

을 갖는다.

7) 중심균형계통추출 (centered balanced systematic sampling: CBSS) (Kim, 1985)

이 방법은 BSS와 CSS의 개념을 결합한 것으로, BSS를 중심화한 것이다.

k가 홀수이면 집락 $S''_{(k+1)/2}$을 뽑는다. 그런데 이 경우 $S''_{(k+1)/2}$은 $S_{(k+1)/2}$과 정확히 일치하므로 CBSS는 CSS와 같은 방법이 된다.

k가 짝수이면 각각 1/2의 확률로 두 집락 $S''_{k/2}$와 $S''_{k/2+1}$ 중 하나를 뽑는다. $\bar{Y}$의 추정량 $\bar{y}_{cb}$는 $\bar{y}''_{k/2}$ 또는 $\bar{y}''_{k/2+1}$이 되며, 편향

$$B(\bar{y}_{cb}) = \frac{1}{2}(\bar{y}''_{k/2} + \bar{y}''_{k/2+1}) - \bar{Y} \tag{3.13}$$

와 평균제곱오차

$$MSE(\bar{y}_{cb}) = \frac{1}{2}\left\{(\bar{y}''_{k/2} - \bar{Y})^2 + (\bar{y}''_{k/2+1} - \bar{Y})^2\right\} \tag{3.14}$$

을 갖는다.

8) 중심변형추출(centered modified sampling: CMS) (Fountain과 Pathak, 1989)

이 방법은 6)의 CMSS와 명칭이 비슷한데 내용은 다르며, 다음과 같이 행해진다. n이 짝수이면 3)의 MSS로 표본을 뽑는다. n과 k가 모두 홀수이면

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

를 표본으로 뽑으며, n이 홀수이고 k가 짝수이면

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

을 뽑는다. 여기서 M은 $N/2$ 또는 $N/2+1$이다(확률은 각각 1/2). CMS에 의한 모평균의 추정량(표본평균)을 제안자들의 이름을 따서 $\bar{y}_{cmFP}$로 나타내기로 한다.

9) 중심균형추출(centered balanced sampling: CBS) (Fountain과 Pathak, 1989)

이 방법도 7)의 CBSS와 명칭은 비슷하나 내용은 다르며, 다음과 같이 행해진다. n이 짝수이면 4)의 BSS로 표본을 뽑는다. n과 k가 모두 홀수이면

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

를 표본으로 뽑으며, n이 홀수이고 k가 짝수이면

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

을 뽑는다. 여기서 M은 $N - k/2$ 또는 $N + 1 - k/2$이다(확률은 각각 1/2). CBS에 의한 모평균의 추정량(표본평균)을 제안자들의 이름을 따서 $\bar{y}_{cbFP}$로 나타내기로 한다.

10) 양끝추출(two-end sampling: TES) (Fountain과 Pathak, 1989)
n이 짝수이면

$$\{U_i,\ U_{N+1-i} : i = 1, 2, \cdots, n/2\}$$

를 표본으로 뽑는다. n과 k가 모두 홀수이면

$$\{U_i,\ U_{N+1-i} : i = 1, 2, \cdots, (n-1)/2\} \bigcup \{U_{(N+1)/2}\}$$

을 뽑으며, n이 홀수이고 k가 짝수이면

$$\{U_i,\ U_{N+1-i} : i = 1, 2, \cdots, (n-1)/2\} \bigcup \{U_M\}$$

을 뽑는다. 여기서 M은 $N/2$ 또는 $N/2 + 1$ 이다(확률은 각각 1/2). TES에 의한 모평균의 추정량(표본평균)을 $\bar{y}_{tcs}$로 나타내기로 한다.

4. 새로운 모평균 추정방법

이 절에서는 k가 짝수이고 n이 홀수($n \geq 5$)인 경우 모평균 $\bar{Y}$에 대한 새로운 추정방법을 제시하고자 한다. 이 방법은 Kim(1985)의 CBSS를 사용하여 표본을 추출한 뒤 표본평균 $\bar{y}_{cb}$보다 수정된 추정량을 써서 $\bar{Y}$를 추정하는 것이다.

$N = 28,\ n = 7,\ k = 4$인 경우를 예로 들어 새로운 추정방법을 설명하기로 한다. 우선 CBSS를 써서 S_2''와 S_3'' 중 하나의 집락을 뽑는다. $S_2'' = \{U_2, U_7, U_{10}, U_{15}, U_{18}, U_{23}, U_{26}\}$이고 $S_3'' = \{U_3, U_6, U_{11}, U_{14}, U_{19}, U_{22}, U_{27}\}$ 이므로, S_2'' 안에 있는 단위들의 번호를 합하면 101이고 S_3''의 경우는 102이다. 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 $\bar{y}_2''$ 나 $\bar{y}_3''$를 쓰는 것보다 수정된 추정량을 써서 $\bar{Y}$를 추정

하는 것이 바람직할 것이다. S_2''가 뽑힌 경우에는 y_{10}대신 "$y_{10.5}$"를 쓰거나 y_{18}대신 "$y_{18.5}$"를 쓰고, S_3''가 뽑힌 경우에는 y_{11}대신 "$y_{10.5}$"를 쓰거나 y_{19}대신 "$y_{18.5}$"를 쓰면 균형이 이루어진다. 물론 여기서 $y_{10.5}$와 $y_{18.5}$는 실제로는 존재하지 않는 가상적인 값이다.

S_2''가 뽑힌 경우 $y_{10.5}$는 y_{10}과 y_{15}를 써서 "추정"할 수 있다. 보간법을 쓰면 $y_{10.5}$는 $(1/10)$ $(9y_{10}+y_{15})$로 추정된다. 따라서 y_{10}대신 이 값을 사용하여 $\bar{Y}$를

$$\begin{aligned}\bar{y}_2''^{*}(3) &= \frac{1}{7}\left\{y_2+y_7+\frac{1}{10}(9y_{10}+y_{15})+y_{15}+y_{18}+y_{23}+y_{26}\right\} \\ &= \bar{y}_2''+\frac{1}{70}(y_{15}-y_{10}) \qquad (4.1)\end{aligned}$$

으로 추정할 수 있다. 같은 방식으로 생각하면, $y_{18.5}$를 사용하는 경우 $y_{18.5}$는 $(1/10)$ $(9y_{18}+y_{23})$으로 추정되므로 y_{18}대신 이 값을 사용하여 $\bar{Y}$를

$$\begin{aligned}\bar{y}_2''^{*}(5) &= \frac{1}{7}\left\{y_2+y_7+y_{10}+y_{15}+\frac{1}{10}(9y_{18}+y_{23})+y_{23}+y_{26}\right\} \\ &= \bar{y}_2''+\frac{1}{70}(y_{23}-y_{18}) \qquad (4.2)\end{aligned}$$

로 추정한다.

2차원 첨자를 써서 나타내면, S_2''가 뽑힌 경우 $\bar{Y}$의 추정값을 각각 1/2의 확률로

$$\bar{y}_2''^{*}(3)=\bar{y}_2''+\frac{1}{70}(y_{24}''-y_{23}'') \qquad (4.3)$$

또는

$$\bar{y}_2''^{*}(5)=\bar{y}_2''+\frac{1}{70}(y_{26}''-y_{25}'') \qquad (4.4)$$

로 한다.

S_3''가 뽑힌 경우에는, $y_{10.5}$를 $(1/10)(y_6+9y_{11})$ 로 추정하여 y_{11}대신 사용하거나, $y_{18.5}$를 $(1/10)(y_{14}+9y_{19})$ 로 추정하여 y_{19} 대신 사용한다. 전자의 경우 $\bar{Y}$는

$$\begin{aligned}\bar{y}_3''^{*}(3) &= \frac{1}{7}\left\{y_3+y_6+\frac{1}{10}(y_6+9y_{11})+y_{14}+y_{19}+y_{22}+y_{27}\right\} \\ &= \bar{y}_3''-\frac{1}{70}(y_{11}-y_6) \qquad (4.5)\end{aligned}$$

로 추정되며, 후자의 경우에는

$$\begin{aligned}\bar{y}_3''^{*}(5) &= \frac{1}{7}\left\{y_3+y_6+y_{11}+y_{14}+\frac{1}{10}(y_{14}+9y_{19})+y_{22}+y_{27}\right\} \\ &= \bar{y}_3''-\frac{1}{70}(y_{19}-y_{14}) \qquad (4.6)\end{aligned}$$

로 추정된다. 2차원 첨자를 써서 나타내면, S_3''가 뽑힌 경우 $\bar{Y}$의 추정값은 각각 1/2의 확률로

$$\bar{y}_3''^{*}(3)=\bar{y}_3''-\frac{1}{70}(y_{33}''-y_{32}'') \qquad (4.7)$$

또는

$$\bar{y}_3''^{*}(5) = \bar{y}_3'' - \frac{1}{70}(y_{35}'' - y_{34}'') \tag{4.8}$$

가 된다.

이상의 내용을 일반화하면 다음과 같다. k가 짝수이고 n이 5 이상의 홀수인 경우 CBSS에 의하여 하나의 집락을 뽑는다. 즉, 각각 1/2의 확률로 두 집락 $S_{k/2}''$와 $S_{k/2+1}''$ 중 하나를 뽑는다. $\bar{Y}$의 추정값으로는, $S_{k/2}''$가 뽑힌 경우에는 등확률로 $\bar{y}_{k/2}''^{*}(3), \bar{y}_{k/2}''^{*}(5), \cdots, \bar{y}_{k/2}''^{*}(n-2)$ 중 하나를 택하여 사용하며, $S_{k/2+1}''$이 뽑힌 경우에는 역시 등확률로 $\bar{y}_{k/2+1}''^{*}(3), \bar{y}_{k/2+1}''^{*}(5), \cdots, \bar{y}_{k/2+1}''^{*}(n-2)$ 중 하나를 택하여 사용한다. 단, 여기서

$$\bar{y}_{k/2}''^{*}(m) = \bar{y}_{k/2}'' + \frac{1}{2n(k+1)}(y_{k/2,m+1}'' - y_{k/2,m}'') \tag{4.9}$$

이고

$$\bar{y}_{k/2+1}''^{*}(m) = \bar{y}_{k/2+1}'' - \frac{1}{2n(k+1)}(y_{k/2+1,m}'' - y_{k/2+1,m-1}'') \tag{4.10}$$

이다. 이 방법을 CBSSI(centered balanced systematic sampling with interpolation)로 나타내기로 하자. CBSSI에 의한 $\bar{Y}$의 추정량을 $\bar{y}_{cbi}$라 하면

$$P(\bar{y}_{cbi} = \bar{y}_{k/2}''^{*}(m)) = P(\bar{y}_{cbi} = \bar{y}_{k/2+1}''^{*}(m)) = \frac{1}{n-3} \qquad (m = 3, 5, \cdots, n-2) \tag{4.11}$$

이 되도록 하는 것이다. $\bar{y}_{cbi}$는 다음과 같은 편향과 평균제곱오차를 갖는다는 것을 쉽게 알 수 있다.

$$B(\bar{y}_{cbi}) = \frac{1}{n-3}\sum_{m}\left\{\bar{y}_{k/2}''^{*}(m) + \bar{y}_{k/2+1}''^{*}(m)\right\} - \bar{Y} \tag{4.12}$$

$$MSE(\bar{y}_{cbi}) = \frac{1}{n-3}\sum_{m}\left[\left\{\bar{y}_{k/2}''^{*}(m) - \bar{Y}\right\}^2 + \left\{\bar{y}_{k/2+1}''^{*}(m) - \bar{Y}\right\}^2\right] \tag{4.13}$$

단, 여기서 $\sum_{m}$은 $m = 3, 5, \cdots, n-2$ 에 걸쳐서 취한 합을 의미하며, 앞으로도 마찬가지이다.

이 절에서 제안된 추정방법은 실제 조사에서도 적용될 수 있다. 또한 계통추출법도 집락추출법의 일종이라고 할 수 있는데, 집락추출법의 원칙 중 하나가 집락내 이질적, 집락간 동질적이 되도록(즉 집락간 변동을 작게) 하는 것이므로, 이러한 관점에서 볼 때에도 두 집락의 균형을 맞추는 것이 바람직한 것이라고 할 수 있다.

5. 무한초모집단모형 하에서의 평균제곱오차의 기대값

이 절에서는 Cochran(1946)이 소개한 무한초모집단모형을 사용하여, 3절과 4절에서 논의된 여러 방법들에 의한 모평균 추정량의 평균제곱오차(mean square error)의 기대값들을 구한다.

5.1. 일반적인 경우

무한초모집단모형이란, 주어진 유한모집단을 무한초모집단으로부터 뽑힌 하나의 표본으로 간주하는 것으로서 다음과 같이 세워진다.

$$y_i = \mu_i + e_i \ (i = 1, 2, \cdots, N) \tag{5.1}$$

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

이제부터 μ와 e에 관해서도 2절과 4절에서 정의된 것과 같은 양식의 기호를 사용하기로 한다. 예컨대

$$\bar{\mu} = \frac{1}{N}\sum_{i=1}^{N}\mu_i$$

$$\mu_{ij} = \mu_{i+(j-1)k}$$

$$\bar{\mu}_i = \frac{1}{n}\sum_{j=1}^{n}\mu_{ij}$$

$$\bar{\mu}''^{*}_{k/2}(m) = \bar{\mu}''_{k/2} + \frac{1}{2n(k+1)}(\mu''_{k/2,m+1} - \mu''_{k/2,m})$$

$$\bar{\mu}''^{*}_{k/2+1}(m) = \bar{\mu}''_{k/2+1} - \frac{1}{2n(k+1)}(\mu''_{k/2+1,m} - \mu''_{k/2+1,m-1})$$

등이다.

이러한 기호들과 위의 가정들, 그리고 3절과 4절에서 밝힌 평균제곱오차들을 이용하면 다음과 같은 정리를 얻을 수 있다. 정리의 증명은 부록에 주어져 있다.

정리 5.1 식 (5.1)을 가정할 때, 여러 방법들에 의한 모평균 $\bar{Y}$의 추정량의 평균제곱오차의 기대값들은 다음과 같다. 단, 여기서 $A = \sigma^2(1/n - 1/N)$이다.

$$EMSE(\bar{y}_{sy}) = EV(\bar{y}_{sy}) = \frac{1}{k}\sum_{i=1}^{k}(\bar{\mu}_i - \bar{\mu})^2 + A \tag{5.2}$$

$$EMSE(\bar{y}_{ec}) = \frac{1}{k}\sum_{i=1}^{k}\left\{(\bar{\mu}_i - \bar{\mu}) + \frac{2i-k-1}{2k(n-1)}(\mu_{i1} - \mu_{in})\right\}^2 + A + \frac{\sigma^2(k^2-1)}{6k^2(n-1)^2} \tag{5.3}$$

$$EMSE(\bar{y}_{mod}) = EV(\bar{y}_{mod}) = \frac{1}{k}\sum_{i=1}^{k}(\bar{\mu}'_i - \bar{\mu})^2 + A \tag{5.4}$$

$$EMSE(\bar{y}_{bal}) = EV(\bar{y}_{bal}) = \frac{1}{k}\sum_{i=1}^{k}(\bar{\mu}''_i - \bar{\mu})^2 + A \tag{5.5}$$

$$EMSE(\bar{y}_{ccn}) = \begin{cases} (\bar{\mu}_{(k+1)/2} - \bar{\mu})^2 + A \ (k : \text{홀수}) \\ \frac{1}{2}\left\{(\bar{\mu}_{k/2} - \bar{\mu})^2 + (\bar{\mu}_{k/2+1} - \bar{\mu})^2\right\} + A \ (k : \text{짝수}) \end{cases} \tag{5.6}$$

$$EMSE(\bar{y}_{cm}) = \frac{1}{2}\left\{(\bar{\mu}'_{k/2} - \bar{\mu})^2 + (\bar{\mu}'_{k/2+1} - \bar{\mu})^2\right\} + A \ (k : \text{짝수}) \tag{5.7}$$

$$EMSE(\bar{y}_{cb}) = \frac{1}{2}\left\{(\bar{\mu}''_{k/2} - \bar{\mu})^2 + (\bar{\mu}''_{k/2+1} - \bar{\mu})^2\right\} + A \ (k : \text{짝수}) \tag{5.8}$$

$$EMSE(\bar{y}_{cbi}) = \frac{1}{n-3}\sum_{m}\left[\left\{\bar{\mu}''^{\star}_{k/2}(m) - \bar{\mu}\right\}^2 + \left\{\bar{\mu}''^{\star}_{k/2+1}(m) - \bar{\mu}\right\}^2\right] + A + \frac{\sigma^2}{2n^2(k+1)^2}$$
$$(k : \text{짝수}, \ n : 5 \text{ 이상의 홀수}) \tag{5.9}$$

5.2. 선형추세를 갖는 모집단의 경우

$\mu_i = a + bi$ (a와 b는 상수, $b \neq 0$)인 경우, 즉 가정된 무한초모집단모형이

$$y_i = a + bi + e_i \quad (i = 1, 2, \cdots, N) \tag{5.10}$$

인 경우를 생각해 보자. e_i는 5.1절에서와 같은 조건을 만족시키는 오차항이다. 이 경우가 바로 모집단에 선형추세가 존재하는 경우이다.

여러 방법들에 의한 모평균 추정량의 평균제곱오차의 기대값들을 구하기 위한 준비단계로 다음과 같은 식들을 얻을 수 있다.

$$\mu_i = a + bi \tag{5.11}$$

$$\bar{\mu} = a + \frac{b}{2}(kn + 1) \tag{5.12}$$

$$\mu_{ij} = a + b\{i + (j-1)k\} \tag{5.13}$$

$$\bar{\mu}_i = a + \frac{b}{2}(kn+1) + b\left(i - \frac{k+1}{2}\right) \tag{5.14}$$

$$\bar{\mu}'_i = \begin{cases} a + \frac{b}{2}(kn+1) & (n : \text{짝수}) \\ a + \frac{b}{2}(kn+1) + \frac{b}{n}(i - \frac{k+1}{2}) & (n : \text{홀수}) \end{cases} \tag{5.15}$$

$$\bar{\mu}''_i = \begin{cases} a + \frac{b}{2}(kn+1) & (n : \text{짝수}) \\ a + \frac{b}{2}(kn+1) + \frac{b}{n}(i - \frac{k+1}{2}) & (n : \text{홀수}) \end{cases} \tag{5.16}$$

그리고 본 논문에서 제안된 방법과 관련된 것들로서, k가 짝수이고 n이 5 이상의 홀수일 때 다음 식들이 얻어진다. ($m = 3, 5, \cdots, n-2$)

$$\mu''_{k/2,m+1} = \mu_{1+k/2+(m+1)k} = a + b\left\{1 + (m + \frac{1}{2})k\right\} \tag{5.17}$$

$$\mu''_{k/2,m} = \mu_{k/2+(m-1)k} = a + b(m - \frac{1}{2})k \tag{5.18}$$

$$\mu''_{k/2+1,m} = \mu_{k/2+1+(m-1)k} = a + b\left\{1 + (m - \frac{1}{2})k\right\} \tag{5.19}$$

$$\mu''_{k/2+1,m-1} = \mu_{1-(k/2+1)+(m-1)k} = a + b(m - \frac{3}{2})k \tag{5.20}$$

$$\bar{\mu}''^{*}_{k/2}(m) = \bar{\mu}''_{k/2} + \frac{1}{2n(k+1)}(\mu''_{k/2,m+1} - \mu''_{k/2,m}) = a + \frac{b}{2}(kn+1) \tag{5.21}$$

$$\bar{\mu}''^{*}_{k/2+1}(m) = \bar{\mu}''_{k/2+1} - \frac{1}{2n(k+1)}(\mu''_{k/2+1,m} - \mu''_{k/2+1,m-1}) = a + \frac{b}{2}(kn+1) \tag{5.22}$$

위의 식들과 5.1절의 정리를 이용하면 다음과 같은 결과를 얻게 된다. 앞에서와 같이 $A = \sigma^2(1/n - 1/N)$ 이다. 식 (5.23)부터 식 (5.27)까지는 Kim(1985)과 Fountain과 Pathak (1989)에서 얻어진 것이며, Singh등 (1968)에서도 일부 유사한 식이 얻어진 바 있다. 식 (5.28)과 식 (5.29)는 Kim(1985)에서 얻어진 것이다. 식 (5.30)은 Fountain과 Pathak(1989)이 구한 것이며, 식 (5.31)은 본 연구에 의해서 얻어진 결과이다.

$$EMSE(\bar{y}_{sy}) = \frac{b^2(k^2-1)}{12} + A \tag{5.23}$$

$$EMSE(\bar{y}_{cc}) = A + \frac{\sigma^2(k^2-1)}{6k^2(n-1)^2} \tag{5.24}$$

$$EMSE(\bar{y}_{mod}) = \begin{cases} A & (n : \text{짝수}) \\ \frac{b^2(k^2-1)}{12n^2} + A & (n : \text{홀수}) \end{cases} \tag{5.25}$$

$$EMSE(\bar{y}_{bal}) = \begin{cases} A & (n : \text{짝수}) \\ \frac{b^2(k^2-1)}{12n^2} + A & (n : \text{홀수}) \end{cases} \tag{5.26}$$

$$EMSE(\bar{y}_{ccn}) = \begin{cases} \frac{b^2}{4} + A & (k : \text{짝수}) \\ A & (k : \text{홀수}) \end{cases} \tag{5.27}$$

$$EMSE(\bar{y}_{cm}) = \begin{cases} A & (k : \text{짝수},\ n : \text{짝수}) \\ \frac{b^2}{4n^2} + A & (k : \text{짝수},\ n : \text{홀수}) \end{cases} \tag{5.28}$$

$$EMSE(\bar{y}_{cb}) = \begin{cases} A & (k : \text{짝수},\ n : \text{짝수}) \\ \frac{b^2}{4n^2} + A & (k : \text{짝수},\ n : \text{홀수}) \end{cases} \tag{5.29}$$

$$\begin{aligned} EMSE(\bar{y}_{cmFP}) &= EMSE(\bar{y}_{cbFP}) = EMSE(\bar{y}_{tcs}) \\ &= \begin{cases} A & (n : \text{짝수}) \\ A & (k : \text{홀수}, n : \text{홀수}) \\ \frac{b^2}{4n^2} + A & (k : \text{짝수}, n : \text{홀수}) \end{cases} \end{aligned} \tag{5.30}$$

$$EMSE(\bar{y}_{cbi}) = A + \frac{\sigma^2}{2n^2(k+1)^2} \quad (k : \text{짝수},\ n : 5\text{이상의 홀수}) \tag{5.31}$$

식 (5.31)은 $\bar{y}_{cbi}$의 기대평균제곱오차가 선형추세의 기울기 b의 값에 무관하다는 것을 말해 준다. 이것은 식 (5.24)에서도 볼 수 있듯이 EC에서도 나타나는 바람직한 성질이다.

6. 기존의 방법들과의 효율성 비교

본 논문에서 제안된 추정량인 $\bar{y}_{cbi}$의 기대평균제곱오차가 식 (5.31)과 같이 얻어졌다. 이제 $\bar{y}_{cbi}$의 효율성을 기존의 여러 추정량들과 비교해 보자.

모집단에 선형추세가 존재하는 경우 식 (5.23)부터 식 (5.31)까지의 식들을 사용하여 여러 방법들의 효율성을 비교할 수 있다. 본 논문에서 제안된 방법이 k가 짝수이고 표본크기 n이 5 이상의 홀수인 경우에 사용하기 위한 것이므로, 이 절에서 고려하는 경우는 모두 이러한 경우이다.

먼저 본 논문에서 제안된 방법($\bar{y}_{cbi}$)과 보통의 계통추출법($\bar{y}_{sy}$)을 비교해 보자. 식 (5.23)과 식 (5.31)로부터, $\bar{y}_{cbi}$가 $\bar{y}_{sy}$보다 효율적일 조건, 즉

$$EMSE(\bar{y}_{cbi}) < EMSE(\bar{y}_{sy}) \tag{6.1}$$

일 필요충분조건은

$$\sigma^2 < \frac{b^2 n^2 (k+1)^3 (k-1)}{6} \tag{6.2}$$

임을 얻게 된다. 뒤에 나올 예제 6.1에서 보여지겠지만, 이것은 오차항의 분산 σ^2(모집단의 분산이 아님)이 터무니없이 큰 값을 갖지 않는 한 성립하는 조건이다.

두 번째로 $\bar{y}_{cbi}$와 Kim(1985)의 CBSS의 $\bar{y}_{cb}$를 비교해 보면, 식 (5.29)와 식 (5.31)로부터, $\bar{y}_{cbi}$가 $\bar{y}_{cb}$보다 효율적일 필요충분조건은

$$\sigma^2 < \frac{b^2 (k+1)^2}{2} \tag{6.3}$$

이다. 이것은 선형추세가 뚜렷하여 σ^2이 작을수록 $\bar{y}_{cbi}$의 효율성이 특히 우수하다는 것을 의미한다.

다음으로 $\bar{y}_{cbi}$와 Yates(1948)의 끝값수정법($\bar{y}_{cc}$)을 비교해 보자. 식 (5.24)와 식 (5.31)로부터

$$\begin{aligned} & EMSE(\bar{y}_{cc}) - EMSE(\bar{y}_{cbi}) \\ &= \frac{\sigma^2}{6k^2(k+1)^2 n^2 (n-1)^2} \left\{ (k^4 + 2k^3 - 3k^2 - 2k - 1)n^2 + 6k^2 n - 3k^2 \right\} \end{aligned} \tag{6.4}$$

을 얻게 되는데, 이 값은 k가 짝수이고 n이 홀수이면 항상 0보다 큰 값이라는 것을 쉽게 알 수 있다. k에 $2, 4, 6, \cdots$을 넣어서 생각해 보면 명백하다. 따라서 기대평균제곱오차의 관점에서 보았을 때 $\bar{y}_{cbi}$는 항상 $\bar{y}_{cc}$보다 효율적이라는 사실을 알 수 있다.

이와 같은 방식으로 모든 경우에 대하여 여러 방법들을 비교한 결과를 다음과 같이 정리할 수 있다. 간결한 표현을 위하여 $EMSE(\bar{y}_{sy})$, $EMSE(\bar{y}_{cbi})$ 등을 각각 sy, cbi 등으로 나타내기로 한다. 따라서, 예컨대 "$cbi < sy$"라는 표현은 $\bar{y}_{cbi}$가 $\bar{y}_{sy}$보다 효율적이라는 것을 의미한다. 그리고 k가 짝수이고 n이 홀수인 모든 경우에 $cm = cb = cmFP = cbFP = tes$이므로 간편성을 위하여 cb 한 가지만 대표로 표시하겠다.

1) $k=2$이고 n이 5 이상의 홀수인 경우

i) $\sigma^2 < 9b^2/2$이면 $cbi < cb = mod = bal < cen = sy$

ii) $9b^2/2 \le \sigma^2 < 9b^2n^2/2$이면 $cb = mod = bal \le cbi < cen = sy$

iii) $9b^2n^2/2 \le \sigma^2$ 이면 $cb = mod = bal < cen = sy \le cbi$

2) k가 4 이상의 짝수, n이 5 이상의 홀수이고 $n < \sqrt{(k^2-1)/3}$인 경우

i) $\sigma^2 < b^2(k+1)^2/2$이면 $cbi < cb < cen < mod = bal < sy$

ii) $b^2(k+1)^2/2 \le \sigma^2 < b^2n^2(k+1)^2/2$이면 $cb \le cbi < cen < mod = bal < sy$

iii) $b^2n^2(k+1)^2/2 \le \sigma^2 < b^2(k+1)^3(k-1)/6$ 이면 $cb < cen \le cbi < mod = bal < sy$

iv) $b^2(k+1)^3(k-1)/6 \le \sigma^2 < b^2n^2(k+1)^3(k-1)/6$이면 $cb < cen < mod = bal \le cbi < sy$

v) $b^2n^2(k+1)^3(k-1)/6 \le \sigma^2$이면 $cb < cen < mod = bal < sy \le cbi$

3) k가 4 이상의 짝수, n이 5 이상의 홀수이고 $n = \sqrt{(k^2-1)/3}$인 경우 (예를 들면 $k = 26$, $n = 15$)

i) $\sigma^2 < b^2(k+1)^2/2$이면 $cbi < cb < cen = mod = bal < sy$

ii) $b^2(k+1)^2/2 \le \sigma^2 < b^2n^2(k+1)^2/2$이면 $cb \le cbi < cen = mod = bal < sy$

iii) $b^2n^2(k+1)^2/2 \le \sigma^2 < b^2n^2(k+1)^3(k-1)/6$ 이면 $cb < cen = mod = bal \le cbi < sy$

iv) $b^2n^2(k+1)^3(k-1)/6 \le \sigma^2$ 이면 $cb < cen = mod = bal < sy \le cbi$

4) k가 4 이상의 짝수, n이 5 이상의 홀수이고 $n > \sqrt{(k^2-1)/3}$인 경우

i) $\sigma^2 < b^2(k+1)^2/2$이면 $cbi < cb < mod = bal < cen < sy$

ii) $b^2(k+1)^2/2 \le \sigma^2 < b^2(k+1)^3(k-1)/6$이면 $cb \le cbi < mod = bal < cen < sy$

iii) $b^2(k+1)^3(k-1)/6 \le \sigma^2 < b^2n^2(k+1)^2/2$이면 $cb < mod = bal \le cbi < cen < sy$

iv) $b^2n^2(k+1)^2/2 \le \sigma^2 < b^2n^2(k+1)^3(k-1)/6$ 이면 $cb < mod = bal < cen \le cbi < sy$

v) $b^2n^2(k+1)^3(k-1)/6 \le \sigma^2$이면 $cb < mod = bal < cen < sy \le cbi$

예제 6.1: $N = 150$개의 단위 중에서 $n = 15$개의 단위를 뽑는 경우를 생각해 보자. 이 때 $k = 10$이 된다. 선형추세의 기울기는 $b = 0.5$라고 하자. 이 경우 여러 방법의 효율성을 비교해 보면 다음과 같다.

i) $\sigma^2 < 15.125$이면 $cbi < cb < mod = bal < cen < sy$

ii) $15.125 \le \sigma^2 < 499.125$이면 $cb \le cbi < mod = bal < cen < sy$

iii) $499.125 \le \sigma^2 < 3403.125$이면 $cb < mod = bal \le cbi < cen < sy$

iv) $3403.125 \le \sigma^2 < 112303.125$이면 $cb < mod = bal < cen \le cbi < sy$

v) $112303.125 \le \sigma^2$이면 $cb < mod = bal < cen < sy \le cbi$

여기서 볼 수 있듯이, σ^2이 비정상적으로 큰 값만 갖지 않는다면 $\bar{y}_{cbi}$는 아주 효율적인 추정량이 된다(*(i)* 과 *(ii)*의 경우). 사실 σ^2이 너무 큰 값을 갖는 경우는 선형추세의 의미가 거의 없어지기 때문에 논의할 가치조차 없게 된다.

예제 6.2: (모의실험을 이용한 설명)
모형

$$y_i = 2 + 0.6i + e_i \ (i = 1, 2, \cdots, 36) \tag{6.5}$$

를 설정하자(즉 $a = 2$, $b = 0.6$). 오차항 e_i의 값들을 실제로 발생시킴으로써 무한초모집단으로부터 크기 $N = 36$인 모집단을 생성한 다음, 다시 이 모집단으로부터 크기 $n = 9$인 표본을 뽑아 모평균을 추정하는 문제를 생각해 보기로 하자. $k = 4$이며, 오차항 e_i는 5.1절에서와 같은 조건을 만족시킨다. e_i의 분산 σ^2의 값은 4로 하였고, e_i의 분포의 형태는 정규분포로 정하였으며, 미니탭(MINITAB)의 RANDOM 명령문을 이용하여 e_i의 값들을 발생시켰다. 생성된 모집단은 다음과 같다.

1.0263	4.3456	5.5614	0.4551	1.6530	3.8664	2.9773	8.5975
8.1807	10.2294	7.3890	9.3456	10.3689	8.3357	10.4477	13.7419
13.7887	13.3193	12.5806	11.5972	15.8172	13.2916	13.7551	11.4084
17.6483	18.6418	20.8810	18.1690	18.8417	20.4046	20.2823	22.8446
21.7381	21.2438	26.4664	23.8162				

이 모집단의 평균은 $\bar{Y} = 12.8627$이며, 이 모집단은 대체적으로 증가하는 선형추세를 가지 고 있음을 볼 수 있다. 3절에서 소개된 기존의 열 가지 방법에 의한 $\bar{Y}$의 추정량들의 평균제곱오차(비편향추정량의 경우에는 분산이 평균제곱오차임)는 아래와 같다.

$$V(\bar{y}_{sy}) = 0.2714,\ MSE(\bar{y}_{cc}) = 0.1197,\ V(\bar{y}_{mod}) = 0.2083,$$
$$V(\bar{y}_{bal}) = 0.1609,\ MSE(\bar{y}_{ccn}) = 0.1562,\ MSE(\bar{y}_{cm}) = 0.2660,$$
$$MSE(\bar{y}_{cb}) = 0.0577,\ MSE(\bar{y}_{cmFP}) = 0.2385,\ MSE(\bar{y}_{cbFP}) = 0.2847,$$
$$MSE(\bar{y}_{tcs}) = 0.0434$$

한편 본 논문에서 제안된 방법인 CBSSI를 사용하면, 모평균 $\bar{Y}$를 다음과 같은 여섯 개의 값 중 하나로 추정하게 된다 (확률은 각각 1/6).

$$\bar{y}_2''^*(3) = 12.8071,\ \bar{y}_2''^*(5) = 12.8095,\ \bar{y}_2''^*(7) = 12.8229,$$
$$\bar{y}_3''^*(3) = 13.1583,\ \bar{y}_3''^*(5) = 13.1502,\ \bar{y}_3''^*(7) = 13.1131$$

따라서 CBSSI에 의한 $\bar{Y}$의 추정량 $\bar{y}_{cbi}$의 평균제곱오차는

$$MSE(\bar{y}_{cbi}) = 0.0400$$

이다. 이 값은 위의 열 가지를 포함한 열한 가지의 방법에 의한 평균제곱오차 중 가장 작은 값이므로, CBSSI가 열한 가지의 방법 중 가장 효율적이라는 것을 보여 준다.

7. 결론

본 논문에서는, 선형추세를 갖는 모집단에서 표본크기 n이 5 이상의 홀수이고 추출률의 역수 k가 짝수인 경우 모평균 $\bar{Y}$에 대한 새로운 추정방법을 제안하였다. 이 방법은 Kim (1985)이 제시한 CBSS(중심균형계통추출)를 써서 표본을 뽑은 뒤 표본평균 $\bar{y}_{cb}$보다 수정된 추정량을 써서 $\bar{Y}$를 추정하는 것이다. 수정된 추정량을 정하는 데에 보간법의 개념을 사용하여 두 집락 $S''_{k/2}$와 $S''_{k/2+1}$ 사이의 균형을 맞추었다.

제안된 추정량 $\bar{y}_{cbi}$와 기존의 방법에 의한 추정량들과의 효율 비교는 Cochran(1946)의 무한초모집단모형에 근거를 둔 기대평균제곱오차를 기준으로 하여 이루어졌다. 그 결과 $\bar{y}_{cbi}$는 오차항의 분산 σ^2이 작을수록 (즉 선형추세가 강할수록) 효율적이라는 것이 밝혀졌으며, σ^2이 아주 크지 않은 값을 갖는 대부분의 현실적인 경우에 기존의 추정량들에 비해서 효율적인 것으로 나타났다. 특히 식 (6.4)에서 볼 수 있듯이 σ^2의 값에 관계없이 $\bar{y}_{cbi}$가 Yates(1948)의 끝값수정법보다 효율적이다. 선형추세를 갖는 모집단의 경우 끝값수정법이 아주 좋은 방법으로 인정받아 온 점을 감안할 때, $\bar{y}_{cbi}$가 이보다 더 효율적이라는 사실은 대단히 주목할 만한 것이라고 할 수 있다.

이러한 비교 결과들은 $\bar{y}_{cbi}$가 선형추세를 갖는 모집단의 특성을 잘 고려한 좋은 추정량이라는 것을 말해 준다. $\bar{y}_{cbi}$는 계산하기도 용이하므로 실무에 사용하는 데에 어려움이 없을 것이다.

감사의 글

본 논문에 대하여 귀중한 조언을 해 주신 심사위원님들과 편집위원님들께 감사드립니다.

참고문헌

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

[2] Buckland, W.R. (1951). A review of the literature of systematic sampling. *Journal of the Royal Statistical Society*, B, Vol. 13, 208-215.

[3] 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.

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

[5] Iachan, R. (1982). Systematic sampling : a critical review. *International Statistical Review*, Vol. 50, 293-303.

[6] 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.

[7] 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.

[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*, B, Vol. 27, 315-320.

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

[11] Yates, F. (1948). Systematic sampling. *Philosophical Transactions of the Royal Society of London*, A, Vol. 241, 345-377.

[2000년 1월 접수, 2000년 7월 채택]

부록 : 5절의 정리 5.1의 증명

본 논문에서 제안된 방법인 CBSSI에 관한 식, 즉 식 (5.9)에 대한 증명만 보이기로 하자. 나머지 식들도 유사한 방법으로 증명할 수 있다.

식 (5.1)에 의해

$$\begin{aligned}
\bar{y}''^{*}_{k/2}(m) &= \bar{\mu}''^{*}_{k/2}(m) + \bar{e}''^{*}_{k/2}(m), \\
\bar{y}''^{*}_{k/2+1}(m) &= \bar{\mu}''^{*}_{k/2+1}(m) + \bar{e}''^{*}_{k/2+1}(m), \\
\bar{Y} &= \bar{\mu} + \bar{e}
\end{aligned}$$

임을 쉽게 알 수 있다. 따라서

$$\begin{aligned}
&E\left[\left\{\bar{y}''^{*}_{k/2}(m) - \bar{Y}\right\}^2 + \left\{\bar{y}''^{*}_{k/2+1}(m) - \bar{Y}\right\}^2\right] \\
&= \left\{\bar{\mu}''^{*}_{k/2}(m) - \bar{\mu}\right\}^2 + \left\{\bar{\mu}''^{*}_{k/2+1}(m) - \bar{\mu}\right\}^2 \\
&\quad + E\left[\left\{\bar{e}''^{*}_{k/2}(m) - \bar{e}\right\}^2\right] + E\left[\left\{\bar{e}''^{*}_{k/2+1}(m) - \bar{e}\right\}^2\right] \qquad (A.1)
\end{aligned}$$

인데, 여기서 셋째 항과 넷째 항은 다음과 같이 쓸 수 있다.

$$\begin{aligned}
&E\left[\left\{\bar{e}''^{*}_{k/2}(m) - \bar{e}\right\}^2\right] = E\left[\left\{\bar{e}''_{k/2} - \bar{e} + P_1(m)\right\}^2\right] \\
&= E\left\{(\bar{e}''_{k/2} - \bar{e})^2\right\} + 2E\left\{(\bar{e}''_{k/2} - \bar{e})P_1(m)\right\} + E\left[\{P_1(m)\}^2\right] \qquad (A.2) \\
&E\left[\left\{\bar{e}''^{*}_{k/2+1}(m) - \bar{e}\right\}^2\right] = E\left[\left\{\bar{e}''_{k/2+1} - \bar{e} - P_2(m)\right\}^2\right] \\
&= E\left\{(\bar{e}''_{k/2+1} - \bar{e})^2\right\} - 2E\left\{(\bar{e}''_{k/2+1} - \bar{e})P_2(m)\right\} + E\left[\{P_2(m)\}^2\right] \qquad (A.3)
\end{aligned}$$

단, 여기서

$$\begin{aligned}
P_1(m) &= \frac{1}{2n(k+1)}(e''_{k/2,m+1} - e''_{k/2,m}), \\
P_2(m) &= \frac{1}{2n(k+1)}(e''_{k/2+1,m} - e''_{k/2+1,m-1})
\end{aligned}$$

이다.

식 (A.2)의 값을 구해 보자. 우선

$$E\left\{(\bar{e}''_{k/2} - \bar{e})^2\right\} = E\left\{(\bar{e}''_{k/2})^2\right\} - 2E\left\{(\bar{e}''_{k/2})(\bar{e})\right\} + E\left\{(\bar{e})^2\right\}$$

인데, $\bar{e}''_{k/2}$ 와 $\bar{e}$는 각각 n개와 N개의 독립인 오차항들의 평균이므로 $E\left\{(\bar{e}''_{k/2})^2\right\} = \sigma^2/n$,

$E\left\{(\bar{e})^2\right\} = \sigma^2/N$임을 쉽게 알 수 있다. 한편

$$\begin{aligned} E\left\{(\bar{e}''_{k/2})(\bar{e})\right\} &= E\left\{\left(\frac{1}{n}\sum_{j=1}^{n} e''_{k/2,j}\right)\left(\frac{1}{N}\sum_{u=1}^{N} e_u\right)\right\} \\ &= \frac{1}{nN} E\left(\sum_{j=1}^{n}\sum_{u=1}^{N} e''_{k/2,j}\cdot e_u\right) \\ &= \frac{1}{nN}\sum_{j=1}^{n}\sum_{u=1}^{N} E\left(e''_{k/2,j}\cdot e_u\right) \end{aligned}$$

인데, 각각의 j에 대하여 $e''_{k/2,j}$는 $e_u(u = 1, 2, \cdots, N)$ 중 정확히 한 개와 동일한 것이므로

$$E\left\{(\bar{e}''_{k/2})(\bar{e})\right\} = \frac{1}{nN}\cdot n\sigma^2 = \frac{\sigma^2}{N}$$

이 된다. 따라서

$$E\left\{(\bar{e}''_{k/2} - \bar{e})^2\right\} = \sigma^2\left(\frac{1}{n} - \frac{1}{N}\right)$$

을 얻는다. 같은 방법으로 식 (A.2)의 둘째 항은 0이 됨을 보일 수 있으며 셋째 항의 값은 $E\left[\{P_1(m)\}^2\right] = \sigma^2/2n^2(k+1)^2$으로 얻어진다. 그러므로 식 (A.2)의 값은

$$E\left[\{\bar{e}''^{*}_{k/2}(m) - \bar{e}\}^2\right] = \sigma^2\left(\frac{1}{n} - \frac{1}{N}\right) + \frac{\sigma^2}{2n^2(k+1)^2} \qquad (A.4)$$

으로 얻어진다. 식 (A.3)의 값도 식 (A.2)와 동일한 방법에 의해 다음과 같이 얻어진다.

$$E\left[\{\bar{e}''^{*}_{k/2+1}(m) - \bar{e}\}^2\right] = \sigma^2\left(\frac{1}{n} - \frac{1}{N}\right) + \frac{\sigma^2}{2n^2(k+1)^2} \qquad (A.5)$$

식 (4.13), (A.1), (A.4), (A.5)를 종합하면 식 (5.9)를 얻게 된다.

Efficient Estimation of the Mean for Populations with a Linear Trend : An Extension of Systematic Sampling *

Hyuk Joo Kim[1] Eun-Yang Seok[2]

ABSTRACT

In this study, we have proposed a sampling method and an estimation method for efficiently estimating the mean of a population which has a linear trend. These methods involve drawing a sample by the so-called "centered balanced systematic sampling", which is an extension of systematic sampling, and then estimating the population mean with an adjusted estimator, not with the sample mean itself. We used the concept of interpolation in determining the adjusted estimator.

We compared the efficiency of the proposed estimator with those of the estimators from existing methods, under the expected mean square error criterion based on the infinite superpopulation model introduced by Cochran(1946). The proposed method is for use in the case when the sample size $n(\geq 5)$ is an odd number and k(the reciprocal of the sampling fraction) is an even number. A good result was also obtained in an example using computer simulation.

Keywords: Linear trend; Centered balanced systematic sampling; Interpolation; Infinite superpopulation model.

* This work was supported by grant No. 981-0105-034-1 from the Basic Research program of the KOSEF.

1) Associate Professor, Division of Mathematical Science, Wonkwang University.
E-mail: hjkim@wonnms.wonkwang.ac.kr

2) Lecturer, Division of Mathematical Science, Wonkwang University.
E-mail: seok1221@hanmail.net