J. of Basic Sciences, Cheju Nat'l Univ
13(1), 99~109, 2000

기초과학연구 제주대학교
13(1), 99~109, 2000

연속조사에서의 Maximum entropy

김익찬[1] · 강형창[2]
1) 전산통계학과 교수 , 2) 제주대학교 전산통계학과 대학원

Maximum entropy on the successive occasions sampling

요 약

모집단들은 시간에 따라 변동한다. 시간에 따라 변동하는 모집단의 특성을 파악하기 위해서는 표본조사를 연속적으로 수행해야한다. 그리고, 모집단을 구성하고 있는 단위들은 크기가 서로 같지 않다. 크기가 서로 같지 않은 모집단을 추출하는 것을 불균등확률추출이라 한다. 불균등확률 비복원추출을 할 때 표본으로 선택될 확률을 가장 크게 하게 위해 maximum entropy를 이용한다. 그리고, 연속조사에서 과거 모집단 단위들을 현재 모집단에 부분중복 하여 모집단 특성을 추정하게 되면, 추정량의 효율이 높아진다. 본 논문에서는 연속조사를 2회로 가정하여 maximum entropy를 이용한 모집단을 연속조사 하고, 표본을 부분 교체하여 관심 있는 모집단의 특성으로 두 번째 시기의 모집단 총합을 추정하는 선형추정량을 제시한다.

1. 서 론

시간에 따라 변동하는 모집단의 특성을 파악하기 위한 표본조사는 시간에 따라 연속적으로 수행되어져야 한다. 왜냐하면, 표본조사에 의한 조사 값들은 시간에 따라 변하는 모집단 단위와 더불어 같이 바뀌기 때문이다. 따라서 시간경과에 따라 모집단의 단위들이 변하는 표본조사는 연속적으로 수행해야한다.

또한 연속적으로 표본을 추출하는 경우, 모집단 단위들이 시간이 지남에 따라 단위들이 전부 변하는 것이 아니기 때문에 시간에 따라 변동하는 모집단은 과거의 단위들을 포함하고 있을 것이다. 따라서, 현재의 모집단 특성을 파악하는데 있어서 과거의 표본을 이용한다면, 현재의 모집단 특성을 파악하는데 효율을 높일 수 있다[Hansen 등, 1953].

그리고, 모집단이 포함하고 있는 단위들의 크기가 서로 같지 않은 표본의 추출은 각 단위들을 같은 확률로서 추출하는 것이 아니라, 각 단위들이 가지고 있는 가치나 크기를 이용하여, 서로 다른 확률로 추출해야 모집단의 특성을 파악하는데 효율적이다. 즉, 모집단을 구성하는 단위들의 크기가 서로 같지 않은 모집단으로부터 표본을 추출하는데 있어서 단위들의 크기가 크거나 단위가 포함하고 있는 가치가 클수록 표본에 포함되는 확률이 커진다는 것이다. 따라서, 본 논문에서는 시간에 따라 변하는 모집단에 대해 연속조사

를 하는데 있어서 모집단 단위들의 크기가 서로 같지 않은 경우의 표본추출 즉, 불균등확률 비복원표본추출에서 모집단 단위가 표본에 포함되는 확률을 maximum entropy를 이용하여 모집단으로부터 표본을 추출한다. 그리고, 현재 모집단 특성을 파악하기 위해 과거의 표본들을 이용하는 부분교체의 문제는 maximum entropy의 표본교체를 이용하여 다루게된다.

따라서, 현재 관심 있는 모집단 특성을 추정하기 위해 연속조사를 2회로 가정하여 처음시기의 표본은 과거의 표본으로 하고, 두 번째 시기의 표본은 현재시점의 표본으로 가정하여 현재시점의 모집단 총합 즉, 두 번째 시기의 모집단 총합의 추정량을 제시한다. 그리고, 연속조사에서 과거의 표본을 이용하는 표본의 부분교체를 하지 않고, 두 번째 시기의 표본만을 가지고 모집단 총합을 추정하는 추정량과 비교한다.

2. 관 련 연 구

2-1. 불균등확률추출과 Maximum entropy

모집단들이 포함하고 있는 단위들의 크기는 같은 경우보다 서로 다른 경우가 많이 존재한다. N개의 단위들로 구성된 모집단으로부터 n개의 서로 다른 단위들을 확률표본추출 할 때, ${}_NC_n$개의 선택 가능한 표본들로부터 표본선택확률이 모두 같지 않은 표본추출을 불균등확률추출이라 한다.

불균등확률추출에서 i번째 모집단 단위를 표본으로 포함하는 포함확률 π_i는 다음과 같다[Hannif and Brewer, 1980 ; Chaudhuri and Vos, 1988].

$$0<\pi_i<1\ ,\ i=(1,2,\cdots,N)\ ,\ \sum_{i=1}^{N}\pi_i=n\ . \tag{2.1}$$

$D^n=\{\boldsymbol{x}=(x_1,\cdots,x_N):\ x_i=1\ \text{또는}\ 0\ ,\ \text{그리고}\ x_1+\cdots+x_N=n\}$. 이라 하고, 포함확률 π_i가 (2.1)을 만족하면, i번째 모집단 단위가 표본에 포함될 확률은 다음과 같다.

$$\pi_i=E(X_i)=\sum_{\boldsymbol{x}\in D^n}x_i\,p(\boldsymbol{x})\ . \tag{2.2}$$

여기서, $p(\boldsymbol{x})$는 maximum entropy이고[Stern and Cover, 1989], maximum entropy 모형은 다음과 같다.

$$p(\boldsymbol{x})=\frac{\prod_{i=1}^{N}w_i^{x_i}}{\sum_{\boldsymbol{x}\in D^n}\left(\prod_{i=1}^{N}w_i^{x_i}\right)}\ . \tag{2.3}$$

다음의 기호를 이용하여 (2.4)와 같은 관계를 정의하자. $S=\{1,\cdots,N\}$ 이고, A , B , C는 S의 부분집합, $A^c=S\backslash A$, $|A|$는 부분집합 A의 원소의 크기라 하면,

$$R(k,C)=\sum_{B\subset C,|B|=k}\left(\prod_{i\in B}w_i\right). \tag{2.4}$$

이 관계식에서, $C(\neq\phi)\subset S$, $1\leq k\leq|C|$ 이면 $R(0,C)=1$ 이고, 상수 $k>|C|$ 이면, $R(k,C)=0$ 이다.

(2.4)식의 정의로부터 다음의 관계식을 제안한다. $C(\neq\phi)\subset S$, $1\leq k\leq|C|$ 일 때,

(a) $\sum_{j\in C}w_jR(k-1,C\setminus\{j\})=kR(k,C)$.

포함확률 π_i와 w_i의 관계는 (2.1)식과 (a) 관계식을 이용하면, 다음과 같이 성립한다.

$$\pi_i=\frac{w_iR(n-1,\{i\}^c)}{R(n,S)}\ (i=1,\cdots,N). \tag{2.5}$$

w_i를 계산하기 위해서 일반성에 위배되지 않게 $\pi_1\leq\pi_2\leq\cdots\leq\pi_N$, $w_N=\pi_N$라 가정하여, 처음의 $N-1$개 방정식을 N개의 방정식으로 양변을 나누면

$$w_i=\frac{\pi_iR(n-1,\{N\}^c)}{R(n-1,\{i\}^c)}\ (i=1,\cdots,N-1),\ w_N=\pi_N. \tag{2.6}$$

여기서, w_i를 이용하여 표본을 추출하면 포함확률비례추출과 같게 된다.

w_i는 다음의 반복 절차를 이용하여 구할 수 있다[Deming and Stephan, 1940].

$$w_i^{(k+1)}=\left.\frac{\pi_iR(n-1,\{N\}^c)}{R(n-1,\{i\}^c)}\right|_{w=w^{(k)}}$$
$$(i=1,\cdots,N-1),\ w_N^{(k+1)}=w_N^{(k)}=\pi_N,\ (w_1^{(k)},\cdots,w_N^{(k)}). \tag{2.7}$$

2-2. 표본확률선택절차

몇몇 표본조사에서는 π_i가 미리 정해져 있어서 n과 w_i는 미리 정해지게 되는 경우가 있고, 어떤 표본조사에서는 w_i는 미리 정해져 있으나 표본의 크기가 달라지는 표본조사가 있게된다. 이런 경우 표본확률선택절차는 표본크기 n에 영향을 받지 않는다. 따라서, 선택절차를 사용하는데 있어 표본크기 n이 미리 고정이 되었는지 여부에 따라 선택절차가 달라진다[Xiang-Hui Chen 등, 1994].

표본선택확률을 계산하는 방식으로는 *forward*와 *backward*가 있다. *forward*는 모집단으로부터 n개의 단위를 표본으로 선택하는 방식이고, *backward*는 모집단으로부터 $N-n$단위를 제거하여 남아있는 n개를 표본으로 선택하는 방식이다.

모집단 단위가 표본으로 선택될 확률을 구하는 절차들을 설명하기 위해, k회 표본을 추출한 후의 결과

를 $A_0, A_1, \cdots, A_n$ 이라 하자. 그리고, $A_0 = \phi$, $A_k \subset S$ 이다.

◎절차 1. (*forward* , n 이 고정)

표본을 $k(k=1,2,\cdots,n)$ 회 추출할 때, 한 단위 $j \in A_{k-1}^c$ 를 추출할 확률은 다음과 같다.

$$P_1(j, A_{k-1}^c) = \frac{w_j R(n-k, A_{k-1}^c \setminus \{j\})}{(n-k+1) R(n-k+1, A_{k-1}^c)} \quad . \tag{2.8}$$

maximum entropy로부터 $A_k = \{i_1, \cdots, i_n\}$, $k=1, \cdots, n$ 의 표본을 순서를 고려하여 $i_1, i_2, \cdots, i_n$ 순서로 표본이 추출된다면, 절차 1을 이용하여 추출하는 경우, 표본으로 추출될 확률은 다음과 같다.

$$\begin{aligned}
\prod_{k=1}^{n} P_1 = \ & (i_k, A_{k-1}^c) = \prod_{k=1}^{n} \frac{w_{i_k} R(n-k, A_k^c)}{(n-k+1) R(n-k+1, A_{k-1}^c)} \\
= \ & \frac{w_{i_1} R(n-1, A_1^c)}{n R(n, A_0^c)} \cdot \frac{w_{i_2} R(n-2, A_2^c)}{(n-1) R(n-1, A_1^c)} \cdots \frac{w_{i_n} R(0, A_n^c)}{(n-n+1) R(1, A_n^c)} \\
= \ & \prod_{k=1}^{n} w_{i_k} \frac{1}{n \cdot (n-1) \cdots 2 \cdot 1} \cdot \frac{R(0, A_n^c)}{R(n, S)} = \frac{1}{n!} \cdot \frac{w_{i_1} \cdot w_{i_2} \cdots w_{i_n}}{R(n, S)} \\
= \ & \frac{1}{n!} \Pr(x_t = 1, t \in A_n) .
\end{aligned}$$

따라서, 순서를 고려하지 않는 경우 표본 $A_k = \{i_1, \cdots, i_n\}$, $k=1, \cdots, n$ 를 포함하는 포함확률은 $\Pr(x_t = 1, t \in A_n)$ 이다.

이 결과로부터 $i_1, \cdots, i_k$ 단위를 포함하는 포함확률 $\pi_{i_1 \cdots i_k}$ 를 maximum entropy로부터 절차 1을 이용하여 나타내면,

$$\Pr(A_k = \{i_1, \cdots, i_k\}) \propto \left(\prod_{l=1}^{k} w_{i_l} \right) \frac{R(n-k, \{i_1, \cdots, i_k\}^c)}{R(n, S)} = \pi_{i_1, \cdots, i_k} . \tag{2.9}$$

이다.

$1 \le k \le n-1$ 이고, $j \in A_k^c$ 이면, (2.9)는 다음과 같이 나타낼 수 있다.

$$P_1(j, A_k^c) = \frac{w_{i_k} P_1(j, A_{k-1}^c) - w_j P_1(i_k, A_{k-1}^c)}{(n-k)(w_{i_k} - w_j) P_1(i_k, A_{k-1}^c)} \quad . \tag{2.10}$$

(2.10)을 이용하여 단위들이 표본으로 선택될 확률을 계산하는 절차는 다음과 같다.

◎표본선택확률의 계산절차

단계 1. π_j/n 를 이용하여 $j=1,\cdots,N$까지 $P_1(j,S)$를 계산한다. 그 다음, 단위 i_1 의 추출확률은

$P_1(i_1,S)$ 에 의해 구한다.

단계 2. 만약 $n>1$ 이면 $A_0\leftarrow\phi$, $A_1\leftarrow\{i_1\}$, $k\leftarrow 2$ 이고, 단계 3으로 간다. 그렇지 않으면 수행을 멈춘다.

단계 3. $P_1(j,A_{k-2}^c)$ 와 $P_1(i_{k-1},A_{k-2}^c)$ 로부터 (2.10)를 이용하여 모든 j에 대해서 $P_1(j,A_{k-1}^c)$ 을 계산한다. 그 다음, 단위 i_k 의 추출확률은 $P_1(i_k,A_{k-1}^c)$ 에 의해 구한다.

단계 4. 만약 $k<n$ 이면, $A_k\leftarrow A_{k-1}\cup\{i_k\}$, $k\leftarrow k+1$ 이고, 단계 3으로 간다. 그렇지 않으면 수행을 멈춘다.

◎절차 2. (*forward* , n 이 고정되어 있지 않음)

$k(k=1,2,\cdots,n)$ 회 추출할 때, 한 단위 $j\in A_{k-1}^c$ 를 추출할 확률은 다음과 같다.

$$P_2(j,A_{k-1}^c)=\sum_{i=0}^{k-1}\frac{w_jR(k-i-1,A_{k-1}^c\setminus\{j\})R(i,A_{k-1})}{(k-i)R(k,S)} \quad . \tag{2.11}$$

절차 1과 절차 2는 다르게 사용된다. 절차 1은 (2.10)를 이용하면, 절차 2보다 계산과정이 빠르지만, n이 고정되지 않은 경우에는 사용할 수 없다. 그리고 절차 2는 표본조사에서 표본이 변하는 경우에 유용하게 사용할 수 있다[Xiang-Hui Chen 등, 1994].

2-3. 표본의 대체 문제

표본조사가 연속적으로 이루어지는 경우, 표본 단위들은 새로운 단위들로 교체되어야 한다. 그리고, 교체되는 새로운 단위들은 기존의 표본 단위들과 같은 확률분포를 가진다. 연속조사에서 모집단의 특성을 추정하는 추정량의 효율을 높이기 위해 과거의 표본을 이용하는 표본교체문제를 maximum entropy를 이용한 표본교체문제를 다룬다.

표본의 교체는 다음과 같은 과정에 의해 이루어진다.

◎표본교체절차

단계 1.

한 단위 $i\in A^c$ 를 절차 2를 이용하여 선택한 다음, 한 단위 $j\in A\cup\{i\}$를 절차 2′을 이용하여 선택한다. (절차 2′은 절차 2에서 *backward* 인 경우)

단계 2.

만약 $i\neq j$ 이면 $A\cup\{i\}\setminus\{j\}$ 를 새로운 표본으로 선택하고, $i=j$ 이면 단계 1을 다시 수행한다.

이 표본교체절차는 연속조사에서 제거된 표본을 추가하는데 이용한다.

3. Maximum entropy를 이용한 연속조사

3-1 Maximum entropy를 이용한 표본추출설계

모집단이 N개의 단위 $U_1, \cdots, U_N$로 구성되어 있고, 2회 연속조사를 한다고 하자.

처음시기의 모집단을 $X_1, \cdots, X_N$, 두 번째 시기의 모집단을 $Y_1, \cdots, Y_N$이라 하자.

모집단으로부터 표본을 선택하기 위해 maximum entropy에 의한 확률로서 표본을 추출한다. 처음시기에 모집단으로부터 크기 n인 표본을 maximum entropy 절차 1(*forward*, n이 고정)를 이용하여 추출한 다음, 처음시기에서 추출된 표본으로부터 크기 u인 표본을 s_1, 크기 m인 표본을 s_2라 하자. 두 번째 시기에서 과거의 표본을 이용하기 위해, 처음시기의 표본 중 s_1은 maximum entropy 절차 2′(*backward*, u는 고정되지 않음)를 이용하여 제거하거나, 또는 s_2를 maximum entropy 절차 2(*forward*, m은 고정되지 않음)를 이용하여 선택한 후 두 번째 시기의 표본에 포함한다. 그 다음, 두 번째 시기에서 제거된 표본크기 u인 표본을 모집단 $U - s_1$로부터 maximum entropy 표본교체절차를 이용하여 추출하고, 두 번째 시기의 표본에 포함하면, 처음시기와 두 번째 시기의 표본크기는 n으로 항상 같게 된다. 그리고, 이들 표본으로부터 현재 관심 있는 모집단의 특성인 두 번째 시기의 모집단 총합을 추정한다.

3-2 Horvitz-Thompson 추정량

모집단 총합을 추정하는데 있어서 관측치 y_i는 i번째 단위에 영향을 받는다. 일반적으로 y_i가 단위에 영향을 받는 경우에, 모집단 총합 $Y(=\sum y_i)$의 추정량은 다음과 같은 Horvitz-Thompson 추정량을 사용한다[Horvitz, and Thompson, 1952].

$$\hat{Y}_{HT} = \sum_{i \in s} \frac{y_i}{\pi_i} .$$

π_i는 maximum entropy 모형으로부터 i번째 단위가 표본에 포함될 포함확률이고, π_{ij}는 maximum entropy 모형으로부터, i번째 단위와 j번째 단위가 동시에 표본에 포함될 포함확률이다 $(i \neq j = 1, 2, \cdots, N)$. 그리고, $\hat{Y}_{HT}$의 분산은 다음과 같다.

$$V(\hat{Y}_{HT}) = \sum_{i=1}^{N} \sum_{j>i}^{N} (\pi_i \pi_j - \pi_{ij}) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 .$$

처음시기와 두 번째 시기의 모집단 총합의 추정량을 Horvitz-Thompson 추정량을 이용하여 정의하면 다음과 같다.

$$\pi_i(s_k)>0,\ i=1,\cdots,N \text{ 이면, } \hat{X}_k=\sum_{i\in s_k}\frac{X_i}{\pi_i(s_k)}\cdot t_i,\ k=1,2, \tag{3.1}$$

$$\pi_i(s_k)>0,\ i=1,\cdots,N \text{ 이면, } \hat{Y}_k=\sum_{i\in s_k}\frac{Y_i}{\pi_i(s_k)}\cdot t_i,\ k=2,3. \tag{3.2}$$

$\pi_i(s_k)=\Pr\{i\in s_k\}$, $\pi_{ij}(s_k)=\Pr\{(i,j)\in s_k$, $i\neq j=1,\cdots,N$, $k=1,2,3$.

$t_i=\begin{cases}1: i\text{번째 모집단 단위가 표본에 포함}\\ 0: i\text{번째 단위가 표본에 포함되지 않음}\end{cases}$, $i=1,\cdots,N$이다.

$\hat{X}_k$, $\hat{Y}_k$는 각각 처음시기의 모집단 총합 X와 두 번째 시기의 모집단 총합 Y의 추정량이고, 각각은 불편추정량이다.

그리고, 각 시기의 모집단 총합의 분산은 다음과 같다.

$$V(\hat{X}_k)=\sum_{i=1}^{N}\sum_{j>i}^{N}(\pi_i(s_k)\pi_j(s_k)-\pi_{ij}(s_k))\left(\frac{X_i}{\pi_i(s_k)}-\frac{X_j}{\pi_j(s_k)}\right)^2. \tag{3.3}$$

$$V(\hat{Y}_k)=\sum_{i=1}^{N}\sum_{j>i}^{N}(\pi_i(s_k)\pi_j(s_k)-\pi_{ij}(s_k))\left(\frac{Y_i}{\pi_i(s_k)}-\frac{Y_j}{\pi_j(s_k)}\right)^2. \tag{3.4}$$

그리고, 각 시기의 모집단들간에는 공분산이 존재하는데, 각 공분산들은 다음과 같다.

$\pi_i(s_2)>0$, $i=1,\cdots,N$ 이면, $\hat{X}_2$ 와 $\hat{Y}_2$ 의 공분산은

$$Cov(\hat{X}_2,\hat{Y}_2)$$
$$=\sum_{i=1}^{N}\sum_{j>i}^{N}(\pi_i(s_2)\pi_j(s_2)-\pi_{ij}(s_2))\left(\frac{X_i}{\pi_i(s_2)}-\frac{X_j}{\pi_j(s_2)}\right)\left(\frac{Y_i}{\pi_i(s_2)}-\frac{Y_j}{\pi_j(s_2)}\right). \tag{3.5}$$

$\pi_i(s_k)>0$, $i=1,\cdots,N$, $k=1,2$ 이면, $\hat{X}_1$ 와 $\hat{X}_2$ 의 공분산은

$$Cov(\hat{X}_1,\hat{X}_2)$$
$$=\frac{1}{2}\sum_{i=1}^{N}\sum_{j\neq i}^{N}(\pi_i(s_1)\pi_j(s_2)-\pi_{ji}(s_2|s_1))\left(\frac{X_i}{\pi_i(s_1)}-\frac{X_j}{\pi_j(s_1)}\right)\left(\frac{X_i}{\pi_i(s_2)}-\frac{X_j}{\pi_j(s_2)}\right). \tag{3.6}$$

$\pi_i(s_k)>0$, $i=1,\cdots,N$, $k=2,3$ 이면, $\hat{X}_k$ 와 $\hat{X}_2$, $\hat{Y}_2$ 와 $\hat{Y}_3$ 의 공분산은

$Cov(\hat{X}_k,\hat{Y}_3)$

$$= \frac{1}{2}\sum_{i=1}^{N}\sum_{j\neq i}^{N}(\pi_i(s_k)\pi_j(s_3) - \pi_{ji}(s_3|s_k))\left(\frac{X_i}{\pi_i(s_k)} - \frac{X_j}{\pi_j(s_k)}\right)\left(\frac{X_i}{\pi_i(s_3)} - \frac{X_j}{\pi_j(s_3)}\right). \tag{3.7}$$

$$Cov(\hat{Y}_2, \hat{Y}_3)$$

$$= \frac{1}{2}\sum_{i=1}^{N}\sum_{j\neq i}^{N}(\pi_i(s_2)\pi_j(s_3) - \pi_{ji}(s_3|s_2))\left(\frac{X_i}{\pi_i(s_2)} - \frac{X_j}{\pi_j(s_2)}\right)\left(\frac{X_i}{\pi_i(s_3)} - \frac{X_j}{\pi_j(s_3)}\right). \tag{3.8}$$

4. Maximum entropy를 이용한 연속조사에서의 선형추정량

3장에서 모집단으로부터 추출된 표본을 이용하여 두 번째 시기의 모집단 총합 Y를 추정하기 위해 다음의 선형추정량을 이용한다[Hansen 등,1953].

$$\hat{Y} = a\hat{X}_1 + b\hat{X}_2 + c\hat{Y}_2 + d\hat{Y}_3 .$$

$\hat{Y}$가 두 번째 시기의 모집단의 총합 Y의 불편추정량이 되기 위한 조건은 다음과 같다.

$$a+b=0, \quad c+d=1 .$$

그러므로, 모집단 총합 $\hat{Y}$는 다음과 같이 다시 쓸 수 있다.

$$\hat{Y} = a(\hat{X}_1 - \hat{X}_2) + c(\hat{Y}_2 - \hat{Y}_3) + \hat{Y}_3 . \tag{4.1}$$

$\hat{Y}$의 분산

$$\begin{aligned} V(\hat{Y}) = {} & a^2 V(\hat{X}_1 - \hat{X}_2) + c^2 V(\hat{Y}_2 - \hat{Y}_3) + 2ac\,Cov(\hat{X}_1 - \hat{X}_2, \hat{Y}_2 - \hat{Y}_3) \\ & + 2ac\,Cov(\hat{X}_1 - \hat{X}_2, \hat{Y}_3) + 2c\,Cov(\hat{Y}_2 - \hat{Y}_3, \hat{Y}_3) + V(\hat{Y}_3) \end{aligned} \tag{4.2}$$

이다.

여기서,

$$V_X = V(\hat{X}_1 - \hat{X}_2), \; V_Y = V(\hat{Y}_2 - \hat{Y}_3), \; V_{XY} = Cov(\hat{X}_1 - \hat{X}_2, \hat{Y}_2 - \hat{Y}_3),$$

$$C_{XY} = Cov(\hat{X}_1 - \hat{X}_2, \hat{Y}_3), \; C_{YY} = Cov(\hat{Y}_2 - \hat{Y}_3, \hat{Y}_3) .$$

이라 하면, $\hat{Y}$의 분산은 다음과 같이 다시 쓸 수 있다.

$$V(\hat{Y}) = a^2 V_X + c^2 V_Y + V(\hat{Y}_3) + 2ac\,V_{XY} + 2aC_{XY} + 2cC_{YY} . \tag{4.3}$$

총합 Y의 최적 추정량을 구하기 위해 이 분산을 최소로 하는 a, b를 미분법에 의하여 구하면 다음과 같다.

$$a_0 = \frac{V_Y C_{XY} - V_{XY} C_{YY}}{V_{XY}^2 - V_X V_Y} \ , \ c_0 = \frac{V_X C_{YY} - V_{XY} C_{XY}}{V_{XY}^2 - V_X V_Y} \ .$$

그리고, 처음시기와 두 번째 시기의 상관계수

$$\rho = \frac{Cov(\hat{X}_1 - \hat{X}_2, \hat{Y}_2 - \hat{Y}_3)}{\sqrt{V(\hat{X}_1 - \hat{X}_2)\, V(\hat{Y}_2 - \hat{Y}_3)}} = \frac{V_{XY}}{\sqrt{V_X V_Y}} \tag{4.4}$$

이다.

최적값 a_0, c_0를 상관계수를 이용하여 다시 쓰면,

$$a_0 = \frac{\rho\sqrt{\frac{V_X}{V_Y}}\, C_{YY} - C_{XY}}{(1-\rho^2) V_X} \ , \ c_0 = \frac{\rho\sqrt{\frac{V_Y}{V_X}}\, C_{XY} - C_{YY}}{(1-\rho^2) V_Y}$$

이다.

최적값 a_0, c_0를 (4.1)식에 대입하면 다음의 최적 추정량 $\hat{Y}_0$ 얻을 수 있다.

$$\hat{Y}_0 = \frac{\rho\sqrt{\frac{V_X}{V_Y}}\, C_{YY} - C_{XY}}{(1-\rho^2) V_X} (\hat{X}_1 - \hat{X}_2) + \frac{\rho\sqrt{\frac{V_Y}{V_X}}\, C_{XY} - C_{YY}}{(1-\rho^2) V_Y} (\hat{Y}_2 - \hat{Y}_3) + \hat{Y}_3 \ . \tag{4.5}$$

따라서, $\hat{Y}_0$ 의 분산은 다음과 같다.

$$V(\hat{Y}_0) = \frac{1}{(1-\rho^2) V_X V_Y} \{2 C_{XY} C_{YY} \sqrt{V_X V_Y}\, \rho - C_{XY}^2 V_Y - C_{YY}^2 V_X\} + V(\hat{Y}_3) \ . \tag{4.6}$$

모집단이 시간에 따라 변동하는 경우에 모집단의 특성을 추정하기 위해 표본조사를 연속적으로 수행하는 경우, 관심 있는 모집단의 특성인 현재시점의 모집단 총합 즉, 두 번째 시기의 모집단 총합 Y를 추정하기 위해 두 번째 시기의 표본만을 이용하는 경우의 선형추정량 다음과 같다.

$$\hat{Z} = a\hat{Y}_2 + b\hat{Y}_3 \ .$$

$\hat{Z}$가 두 번째 시기의 모집단 총합 Y의 불편추정량이 되기 위한 조건은 다음과 같다.

$$a + b = 1 \ .$$

그러므로, 모집단 총합의 추정량 $\hat{Z}$는 다음과 같이 다시 쓸 수 있다.

$$\hat{Z} = a(\hat{Y}_2 - \hat{Y}_3) + \hat{Y}_3 \ . \tag{4.7}$$

따라서, $\hat{Z}$ 의 분산은 다음과 같다.

$$V(\hat{Z}) = a^2 V(\hat{Y}_2 - \hat{Y}_3) + 2aCov(\hat{Y}_2 - \hat{Y}_3, \hat{Y}_3) + V(\hat{Y}_3) \tag{4.8}$$

여기서, $V_Y = V(\hat{Y}_2 - \hat{Y}_3)$, $C_{YY} = Cov(\hat{Y}_2 - \hat{Y}_3, \hat{Y}_3)$ 이라 하면, $\hat{Z}$ 의 분산은 다음과 같이 쓸 수 있다.

$$V(\hat{Z}) = a^2 V_Y + V(\hat{Y}_3) + 2aC_{YY} . \tag{4.9}$$

모집단의 총합 Y의 최적 추정량을 구하기 위해 이 분산을 최소로 하는 a를 미분법에 의하여 구하면 다음과 같다.

$$a_0 = -\frac{C_{YY}}{V_Y} .$$

최적값 a_0를 이용하여 최적추정량 $\hat{Z}_0$을 구하면

$$\hat{Z}_0 = -\frac{C_{YY}}{V_Y}(\hat{Y}_2 - \hat{Y}_3) + \hat{Y}_3 . \tag{4.10}$$

이고, 분산은 다음과 같다.

$$V(\hat{Z}_0) = -\frac{C_{YY}^2}{V_Y} + V(\hat{Y}_3) . \tag{4.11}$$

모집단이 시간에 따라 변동하는 경우에 표본을 연속적으로 조사하는 경우, 관심 있는 모집단의 특성인 현재시점의 모집단 총합 즉, 두 번째 시기의 모집단 총합을 추정하기 위해 처음시기의 표본과 두 번째 시기의 표본을 이용한 선형추정량과 처음시기의 표본을 이용하지 않고 두 번째 시기의 표본만을 이용하여 두 번째 시기의 모집단의 총합을 추정하는 선형추정량과 분산차이를 비교해 보면, 다음과 같다.

$$\begin{aligned}
&V(\hat{Z}_0) - V(\hat{Y}_0) \\
&= \frac{1}{(1-\rho^2)V_X V_Y}\{C_{XY}^2 V_Y + C_{YY}^2 V_X - 2C_{XY}C_{YY}\sqrt{V_X V_Y}\,\rho\} - \frac{C_{XY}^2}{V_Y} \\
&= \frac{1}{(1-\rho^2)V_X V_Y}\{C_{XY}^2 V_Y - 2C_{XY}C_{YY}\sqrt{V_X V_Y}\,\rho + V_X C_{XY}^2 \rho^2\} \\
&= \frac{1}{(1-\rho^2)V_X V_Y}\{C_{XY}\sqrt{V_Y} - C_{YY}\sqrt{V_X}\,\rho\}^2 \\
&\geq 0
\end{aligned}$$

따라서, 모집단의 단위가 시간에 따라 변동하는 모집단의 특성을 추정하기 위해 표본을 연속적으로 조사하는 연속조사에서 두 번째 시기의 모집단의 총합을 추정하는데 있어서, 처음시기의 표본과 두 번째 시기의 표본을 이용하는 추정량이 두 번째 시기의 표본만을 이용하는 추정량보다 효율이 높다는 것을 알 수 있다.

5. 결 론

시간에 따라 모집단 단위들이 변동하는 연속조사에서 표본단위들은 새로운 단위들로 교체된다. 그리고, maximum entropy 모형에서 새로운 단위들은 모집단과 같은 확률분포를 따르게된다. 따라서, maximum entropy 모형으로부터 표본을 선택하여 모집단의 특성을 추정하게 되면 효율적인 추정량을 얻을 수 있다. 모집단의 단위가 시간에 따라 변동하는 표본추출에서 연속조사를 2회로 가정하고, maximum entropy 모형을 이용하여 표본을 추출한 다음, 두 번째 시기의 모집단 총합을 추정하는 선형추정량 $\hat{Y}_0$ (4.5)을 제시하였다. 그리고, 추정량에 대한 분산 (4.6)를 제시하였다.

$\hat{Y}_0$는 처음시기와 두 번째 시기의 표본으로 구성되어있고, $\hat{Y}_0$는 두 번째 시기의 표본만을 이용한 선형추정량 $\hat{Z}_0$ (4.10)보다 분산이 작아서 효율이 더 높게나왔다.

6. 참 고 문 헌

Chaudhuri, A. and Vos, J. W. E.(1988). Unified Theory Strategies of Survey Sampling. New York: Elsevier Science pp.143-346.

Deming, W. E. and Stephan, D.(1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. Ann. Math. Statist. 11, pp.427-444.

Hanif, M. and Brewer, K. R. W.(1980). Sampling with unequal probabilities without replacement: A review. int. Statist. Rev. 48. pp.317-335.

Hansen, M. H., Hurwitz, W. N and Madow, W. G.(1953). Sampling Survey Methods and Theory, Vol. I, John Wiley and Sons, New York.

Horvitz, D. G., and Thompson, D. J.(1952). A generalization of sampling without replacement from a finite universe, Journal of the American Statistical Association, Vol. 47, pp.663-685.

Stern, H. and Cover, T. M.(1989). Maximum entropy and the lottery. J. Am. Statist. Assoc. 84, pp980-985.

Xiang-Hui Chen, Arthur P. Dempster and Jun S. Liu.(1994). Weighted finite population sampling to maximum entropy. Biometrika 81, 3, pp. 457-469.