碩士學位論文

연속조사에서의 Maximum entropy

濟州大學校 大學院
電算統計學科

姜 亨 吕

2000年 12月

연속조사에서의 Maximum entropy

指導敎授 金 益 贊

姜 亨 昌

이 論文을 理學 碩士學位 論文으로 提出함

2000年 12月

姜 亨 昌의 理學 碩士學位 論文을 認准함

審査委員長 _____ (인)

委　　員 _____ (인)

委　　員 _____ (인)

濟州大學校 大學院

2000年 12月

Maximum entropy on the successive occasions sampling

Hyung-Chang Kang
(Supervised by Professor Ik Chan Kim)

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE

DEPARTMENT OF OCEANOGRAPHY
GRADUATE SCHOOL
CHEJU NATIONAL UNIVERSITY

DECEMBER 2000

목차

Abstract

We consider occasion sampling carried over every occasion. The survey values for the same population unit have a trend to change in accordance with time varying. Hence we use the information contained the previous samples to improve the current estimator for the population quantities. We estimate the current total with partial replacement of units.

In this thesis, Assume that two successive occasion, when a sample is select from the population by the maximum entropy model, partial replacement scheme is that a part of the sample consisting of up units on the first occasions is retained in the sample on the second occasion, and the remaining part of the sample is replaced on second occasion by new units distinct from any previous ones to improve the precision of an estimator.

we suggest a linear estimator for the second population total which has the smallest variance.

I. 서론

시간경과에 따라 변동하는 모집단의 특성을 파악하기 위한 표본조사는 시간이 경과함에 따라 연속적으로 수행되어져야 한다. 왜냐하면, 표본조사에 의한 조사 값들은 시간의 변동에 따라 변하는 모집단 단위와 더불어 같이 바뀌기 때문이다. 즉, 모집단으로부터 표본을 추출하여 모집단 특성을 추정하면, 그 시점에서는 모집단의 특성을 파악할 수 있으나, 시간이 경과한 후 모집단의 단위들이 변하여 있는 경우에, 지난 표본을 가지고 현재 모집단의 특성을 추정한다는 것은 시간이 경과된 현재시점에서는 별로 도움을 주지 못한다. 따라서, 모집단의 단위들이 시간에 따라 변동하는 경우에는 표본조사를 연속적으로 수행하지 않으면 모집단의 특성을 제대로 파악할 수 없게 된다. 따라서 시간경과에 따라 모집단의 단위들이 변하는 표본조사는 연속적으로 수행해야한다.

또한 연속적으로 표본을 추출하는 경우, 모집단 단위들이 시간이 지남에 따라 완전히 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 있을 것이다. 따라서, 연속적으로 표본을 추출하여 현재 모집단의 특성을 파악하고자 할 때, 과거의 표본을 이용할 필요가 있다.

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 연속조사를 하는데 있어서, 모집단 단위들의 크기가 서로 같지 않은 경우의 표본추출, 즉 불균등확률표본추출에서 모집단 단위가 표본에 포함되는 확률을 maximum entropy를 이용하여 모집단으로부터

표본을 추출한다. 그리고, 현재 모집단 특성을 파악하기 위해 과거의 표본들을 이용하는 부분복원의 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.

본 논문의 순서와 내용은 다음과 같다.

I장에서는 본 논문의 제기부분과 본 논문의 전반적인 내용에 대해 다루고 있다.

II장에서는 문헌연구를 통하여 이론적 고찰을 하였는데, 모집단이 시간 경과에 따라 변동하는 표본추출을 연속적으로 조사하는 연속조사와 과거의 표본을 이용하는 표본의 부분교체문제에 대한 내용을 고찰한다.

III장에서는 문헌연구를 통하여 불균등확률추출과 maximum entropy의 관계와 성질, maximum entropy를 이용하여 모집단 단위가 표본으로 추출될 확률과 표본의 대체문제에 대한 내용을 다룬다.

IV장에서는 연속조사를 2회로 가정하고, maximum entropy를 이용하여 표본을 추출한다. 그리고, 추출된 표본들로부터 각 시기의 모집단 총합을 추정한다.

V장에서는 IV장에서 추정한 추정량을 이용하여 현재시점 즉, 두 번째 시기의 모집단 총합을 선형추정량을 이용하여 추정한다.

마지막으로 VI장에서는 본 논문에 대한 결론과 요약을 다룬다.

II. 연속조사와 표본부분교체

모집단 단위들이 시간에 따라 변동하는 모집단의 표본추출은 시간에 따라 연속적으로 수행되어져야 한다. 왜냐하면, 시간경과에 따라 모집단 단위들이 변동하게되면, 표본의 값들도 변동되기 때문이다. 따라서, 모집단의 특성을 제대로 파악하려면 표본조사를 연속적으로 수행해야 한다. 또한, 연속조사에 있어서 표본의 효율을 높이기 위해 과거의 표본을 이용한 표본의 교체를 고려해야 한다.

연속조사에서 표본의 부분교체 문제는 Jessen에 의해 처음 연구되었다. 그 후 Hansen, Hurwitz 와 Madow는 연속조사에서 두 번째 시기의 모집단 평균에 대한 선형추정량을 제시하였다(Hansen 등 1953). 표본의 추출방법은 처음시기에서 크기 n인 표본을 단순임의 복원추출하고, 추출된 단순임의 표본에서 Pn단위는 두 번째 시기의 표본에 포함하고, 나머지 Qn단위는 두 번째 시기에서 모집단으로부터 추출하였다. 따라서, 표본은 처음시기와 두 번째 시기의 두 집합 단위로 구성된다. 즉, 처음시기의 표본으로부터 Pn단위를 선택하고, 선택된 단위는 두 번째 시기의 표본에 포함이 된다. 그리고, 포함되지 않은 Qn단위는 두 번째 시기에서 모집단으로부터 독립적으로 선택하였다 ($P+Q=1$). 그러므로, 처음시기와 두 번째 시기의 표본크기는 항상 n이 된다.
그리고, 두 번째 시기의 모집단 평균을 추정하기 위해 두 시기의 표본들을 이용하여 다음과 같은 선형추정량을 구하였다.

$$\bar{y}_2' = a(\bar{y}_{11}' - \bar{y}_{12}') - c(\bar{y}_{21}' - \bar{y}_{22}') + \bar{y}_{22}' . \qquad (2.1)$$

여기서, $\bar{y}'_{12}$와 $\bar{y}_{21}'$는 처음시기의 Pn단위의 표본평균들이고, $\bar{y}_{11}'$와 $\bar{y}_{22}'$는 각각 처음시기와 두 번째 시기의 Qn단위의 표본평균이다.
따라서, $\bar{y}_2'$의 분산은 다음과 같이 표시할 수 있다.

$$V(\bar{y}_2') = a^2 V(\bar{y}_{11}' - \bar{y}_{12}') + c^2 V(\bar{y}_{21}' - \bar{y}_{22}') + V(\bar{y}_{22}') + 2ac\,Cov(\bar{y}_{11}' - \bar{y}_{12}', \bar{y}_{21}' - \bar{y}_{22}') + 2c\,Cov(\bar{y}_{11}' - \bar{y}_{12}', \bar{y}_{22}') . \quad (2.2)$$

이 분산을 최소로 하는 최적값 a와 c값을 구하면, 다음과 같다.

$$a_0 = \frac{\rho PQ}{1-Q^2\rho^2}\frac{\sigma_2}{\sigma_1}\ ,\ c_0 = \frac{P}{1-Q^2\rho^2}\ . \quad (2.3)$$

여기서, ρ는 $\bar{y}_{12}'$과 $\bar{y}_{21}'$의 상관계수 이다.

최적값 a_0, c_0를 (2.1)에 대입하면 다음의 최적 추정량을 얻을 수 있다.

$$opt.\ \bar{y}_2' = \frac{\rho PQ}{1-Q^2\rho^2}\frac{\sigma_2}{\sigma_1}(\bar{y}_{11}' - \bar{y}_{12}') + \frac{P}{1-Q^2\rho^2}(\bar{y}_{21}' - \bar{y}_{22}') + \bar{y}_{22}' . \quad (2.4)$$

$opt.\ \bar{y}_2'$의 분산은 최적값 a_0, c_0를 (2.2)식에 대입하여 정리하면,

$$V(opt.\ \bar{y}_2') = \frac{\sigma_2^2}{n}\frac{1-Q\rho^2}{1-Q^2\rho^2}\ . \quad (2.5)$$

이다. 그리고, 이 분산을 최소로 하는 최적 표본교체비율은

$$Q = \frac{1}{1+\sqrt{1-\rho^2}}\left(= \frac{1-\sqrt{1-\rho^2}}{\rho^2}\right). \quad (2.6)$$

이므로, $opt.\ \bar{y}_2'$ 의 분산을 (2.6)을 이용하여 다시 쓰면, 다음의 최소분산을 얻게 된다.

$$V(opt.\ \bar{y}_2') = \frac{\sigma_2^2}{n}\,\frac{1+\sqrt{1-\rho^2}}{2}\ . \tag{2.7}$$

이 결과로부터 연속조사에서 두 번째 모집단의 평균을 추정하는 경우, 부분교체를 사용하는 경우와 그렇지 않고 두 번째 시기의 표본만을 이용하여 모집단을 추정하는 경우를 비교해보면, 두 번째 시기의 모집단 평균을 두 번째 시기의 표본만을 이용하여 모집단 평균을 추정하는 추정량은 다음과 같이 쓸 수 있다.

$$\bar{y}' = P\,\bar{y}_{21}' + Q\,\bar{y}_{22}'\ ,\ P+Q=1\ . \tag{2.8}$$

여기서, 모집단의 분산을 $\sigma_2^2=\sigma^2$ 을 가정하여, (2.7)과 (2.8)의 두 추정량의 분산을 비교하면, ρ 가 어떤 값을 가질 때,

$$V(opt.\ \bar{y}_2') \le V(\bar{y}')\left(=\frac{\sigma^2}{n}\right). \tag{2.9}$$

이다.

이 결과로부터 표본의 부분교체를 이용하는 경우의 추정량은 두 번째 시기의 표본만을 이용한 추정량보다 효율이 높다는 것을 알 수 있다. 그리고, 이 결과는 복원추출방법이나 비복원추출방법 모두 같은 결과를 나타낸다(김규성, 1990).

따라서, 연속적인 조사에서 현재시점의 관심 있는 모집단의 특성을 추정하는데 있어 과거 처음시기의 표본을 두 번째 시기의 표본에 일부 포함하여 모집단의 특성을 추정하게 되면, 추정량의 효율이 높아짐을 알 수 있다.

III. 불균등확률추출과 Maximum entropy

1. 불균등확률추출과 Maximum entropy

모집단들이 포함하고 있는 단위들의 크기는 같은 경우보다 서로 다른 경우가 많이 존재한다. 따라서, 본 논문에서는 모집단 단위들의 크기가 서로 다른 표본추출인 불균등확률표본추출을 다룬다.

N개의 단위들로 구성된 모집단으로부터 n개의 서로 다른 단위들을 확률표본추출할 때, ${}_NC_n$개의 선택 가능한 표본들로부터 표본을 선택할 때 표본선택확률이 모두 같지 않은 표본추출을 불균등확률추출이라 한다.

표본추출설계에 있어서 조사항목에 관하여 큰 값을 갖는 단위가 작은 값을 갖는 단위보다 표본에 포함되는 확률을 더 크게 하는 것이 합리적이고 또한 효율적이다. 왜냐하면, 등확률추출은 단위와 무관하게 표본으로 추출된 확률이 같지만, 불균등확률추출은 큰 값을 갖는 단위는 작은 값을 갖는 단위보다 표본으로 추출될 확률을 크게 하여 좀 더 효과적으로 모집단을 대표하는 표본을 추출할 수 있기 때문이다. 따라서, 이번 장에서는 단위의 크기가 클수록 표본에 포함되어지는 확률을 크게 하는 maximum entropy에 대한 내용을 다룬다. 그런데, 어떤 표본추출에서는 단위의 크기를 알 수 있는 경우가 있고, 다른 어떤 경우에서는 단위의 크기를 알 수 없는 경우가 있을 수 있다. 그러나, 단위의 크기를 알지 못하는 경우에 단위의 크기는 조사항목과 높은 상관을 보이는 것으로써 단위의 크기를 측정할 수 있다. 예를 들면, 어떤 한 병원의 크기를 측정할 때 병원의 크기는 병원에 있는 침대 개수의 총합으로 측정할 수도 있고, 또는 어떤 기간동안 환자들이 침대를 사용한 평균으로 측정할 수도 있다(박혜경, 1989).

따라서, 본 논문에서는 단위의 크기를 알고 있는 경우로 가정한다.

불균등확률추출설계에서 i번째 모집단 단위를 표본으로 포함하는 포함확률 π_i는

$$0 < \pi_i < 1 \ , \ i = (1,2,\cdots,N) \ , \ \sum_{i=1}^{N} \pi_i = n \ . \tag{3.1}$$

이다(Hanif and Brewer, 1980 ; Chaudhuri and Vos, 1988).

모집단의 단위들이 표본으로 선택되는 확률표본을 지시변수 $\boldsymbol{X}$를 이용하여 표현하면 다음과 같다.

$$\boldsymbol{X} = (X_1, X_2, \cdots, X_N) \ .$$

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

$D^n = \{ \boldsymbol{x} = (x_1, \cdots, x_N) : x_i = 1 \text{ 또는 } 0, \text{ 그리고 } x_1 + x_2 + \cdots + x_N = n \}$. 이라 하면, 확률벡터 $\boldsymbol{X}$는 D^n의 값을 갖는다.

어떤 $\boldsymbol{x} \in D^n$이고, $p(\boldsymbol{x}) > 0$, $\sum_{\boldsymbol{x} \in D^n} p(\boldsymbol{x}) = 1$.을 만족하면, 불균등확률표본추출설계에서 $p(\boldsymbol{x})$는 확률밀도함수(Probability Density Function)로 표시한다.

따라서, 포함확률 π_i가 (3.1)을 만족하면, i번째 모집단 단위가 표본에 포함될 확률은 다음과 같다.

$$\pi_i = E(X_i) = \sum_{\boldsymbol{x} \in D^n} x_i \, p(\boldsymbol{x}) \ . \tag{3.2}$$

불균등확률추출설계에서 표본집합을 구하기 위한 다음의 세 가지 방법을 제안하고, 제안한 세 가지 방법들은 서로 같음을 보인다.

방법 1. 모집단 단위들의 가중치들 $w = (w_1, \cdots, w_N)$, $w_i > 0$, $i = 1, \cdots, N$에서 어떤

가중치 w_i를 선택하면, 즉 i번째 단위가 표본으로 선택된다면, $p(x)$와 가중치 w_i는 다음과 같은 관계로 정의된다.

$$p(x) \propto \prod_{i=1}^{N} w_i^{x_i} . \tag{3.3}$$

(3.3)에서 x_i는 표본으로 선택되면 1, 표본으로 선택되지 않으면 0 값을 갖는 양의 상수이므로, i번째 단위가 표본으로 선택되면, $p(x)$는 x_i에 의해 계산되는 $w_i^{x_i}$와 같거나 비례하게 된다.

방법 2. 포함확률 $\pi=(\pi_1,\cdots,\pi_N)$, $\pi_i>0$, $i=1,\cdots,N$에서 i번째 단위가 표본에 포함될 포함확률 π_i가 (3.1)을 만족하면, (3.2)에서의 $p(x)$는 entropy를 최대로 한다. 즉, $-\sum p(x)\log p(x)$를 가장 크게 하는 $p(x)$는 i번째 단위가 표본에 포함되는 확률을 가장 크게 한다(Stern and Cover, 1989).

여기서, 만약 가중치 벡터 w가 결정되어진다면, 즉 방법 1에서 정의된 $p(x)$와 방법 2에서 주어진 π가 서로 만족하면, 방법 2에 의해서 방법 1은 유일한 maximum entropy 표본추출설계가 된다. 이 경우 불균등확률추출은 포함확률비례추출과 같은 방법이 된다.

방법 3. 확률 $p=(p_1,\cdots,p_N)$, $0<p_i<1$, $i=1,\cdots,N$이고, 확률 $p=(p_1,\cdots,p_N)$는 독립 베르누이 시행에 의해 표본으로 나타나는 경우로서 $Z=(Z_1,\cdots,Z_N)$라고 정의하면, X의 표본분포는 Z의 조건부분포로 정의된다 ($\sum Z_i=n$).

만약, $w_i \propto p_i/(1-p_i)$ 이면, 방법 3은 방법 1과 같은 표본추출설계를 의미한다.

즉, 모집단에서 확률 p_i를 갖는 i번째 단위가 표본으로 선택되면, i번째 단위의 가중치 w_i가 $p_i/(1-p_i)$에 비례하게되고, 방법 1과 방법 3은 서로 같은 표본추출설계

임을 의미한다. 이 경우 불균등확률추출은 확률비례추출과 같은 방법이 된다.

따라서, 불균등확률추출에서 방법 1과 방법 2가 서로 관계가 있으면, 포함확률비례추출이 되고, 방법 1과 방법 3이 관계가 있으면, 확률비례추출과 같게된다.

앞의 세 가지 방법들에 의해서, w가 (3.2)에서 π에 의해 결정된다면 maximum entropy 모형은 다음과 같다.

$$p(\boldsymbol{x}) = \frac{\prod_{i=1}^{N} w_i^{x_i}}{\sum_{\boldsymbol{x} \in D^*} \left(\prod_{i=1}^{N} w_i^{x_i} \right)} \quad . \tag{3.4}$$

다음의 2절에서는 포함확률과 가중치의 관계에 대한 내용으로, 포함확률 π가 (3.1)를 만족하고, 대응하는 w가 존재한다면 반복절차를 이용하여 w를 계산한다.

3절에서는 maximum entropy 성질로서 2차이상의 포함확률과 w의 관계 및 포함확률의 성질에 대해서 알아본다. 4절에서는 maximum entropy 모형으로부터 모집단 단위가 표본으로 선택되는 확률에 대한 내용을 다룬다. 그리고, 5절에서는 maximum entropy를 이용한 표본의 대체문제에 대해 다루게 된다.

2. w와 포함확률의 관계

포함확률 π와 w의 관계는 정리 3-1을 이용하여 보일 수 있다.

정리 3-1. 포함확률 π가 (3.1)을 만족하고, (3.2)에서 $p(\boldsymbol{x})$가 maximum entropy 모형을 만족하여 w가 존재하면, w는 다시 계산되어 정해진다(Brown, 1986).

π로 부터 w를 계산하기 위해 다음과 같은 기호를 사용한다.

$S=\{1,\cdots,N\}$, $A, B, C \subset S$, $A^c = S \backslash A (A \subset S)$, $|A|$는 부분집합 A의 원소의 크기.

이 기호를 이용하여, 포함확률 π와 w의 관계를 관계함수 R로 표현하면, 다음과 같이 정의된다.

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

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.

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

(a) $\sum_{j \in C} w_j R(k-1, C \backslash \{j\}) = k R(k, C)$,

(b) $\sum_{j \in C} R(k, C \backslash \{j\}) = (|C| - k) R(k, C)$,

(c) $\sum_{i=0}^{k} R(i, C) R(k-i, C^c) = R(k, C)$.

포함확률 π_i와 가중치 w_i의 관계식은 (3.2)를 앞에서 제안한 (a) 관계식을 이용하면, 다음과 같이 성립한다.

$$\pi_i = \frac{w_i R(n-1, \{i\}^c)}{R(n, S)} \quad (i = 1, \cdots, N). \tag{3.5}$$

(3.5)는 (3.1)에서 $\sum \pi_i = n$ 이므로, 제안된 관계식 (a)를 변형하면, (3.5)가 성립한다.

w_i를 계산하기 위해 다음의 과정을 거친다.

일반성에 위배되지 않게 $\pi_i \leq \pi_2 \leq \cdots \leq \pi_N$, $w_N = \pi_N$라 가정하여, 처음의 $N-1$ 개의 방정식을 N개의 방정식으로 양변을 나누면,

$$\frac{\pi_i}{\pi_N} = \left(\frac{w_i R(n-1, \{i\}^c)}{R(n, S)} \right) \Big/ \left(\frac{w_N R(n-1, \{N\}^c)}{R(n, S)} \right).$$

이고, 다음의 관계를 얻을 수 있다.

$$w_i = \frac{\pi_i R(n-1, \{N\}^c)}{R(n-1, \{i\}^c)} \quad (i = 1, \cdots, N-1), \quad w_N = \pi_N . \tag{3.6}$$

여기서, w_i는 π_i와 연관되어 나타나게 된다. 따라서, w_i는 포함확률 π_i에 비례하게 되므로, 이는 포함확률비례추출과 같은 방법이 된다.

(3.6)에서 w_i는 다음의 반복 절차를 이용하여 구할 수 있다.

$$w_i^{(k+1)} = \frac{\pi_i R(n-1, \{N\}^c)}{R(n-1, \{i\}^c)} \Bigg|_{w = w^{(k)}}$$

$$(i = 1, \cdots, N-1), \quad w_N^{(k+1)} = w_N^{(k)} = \pi_N, \quad (w_1^{(k)}, \cdots, w_N^{(k)}) . \tag{3.7}$$

$W = \{w : 0 < w_i \leq \pi_i, \ i = 1, \cdots, N-1, \ w_N = \pi_N\}$라 하면, (3.6)에서 방정식의 집합은 W에서 유일한 해 w^*를 갖는다. $w^{(0)} = \pi$로 시작하여, (3.7)을 반복하면, 벡터 $w^{(k)}$, $k = 1, 2, \cdots$는 단조적으로 수렴한다(Deming and Stephan, 1940 ; Darroch and Ratcliff, 1972).

3. Maximum entropy의 성질

성질 1.

π_{ij}는 i번째 모집단 단위와 j번째 모집단 단위가 모두 표본에 포함되는 2차 포함확률이고, 2차 포함확률은 다음과 같이 쓸 수 있다.

$$\pi_{ij} = w_i w_j R(n-2, \{i,j\}^c) / R(n, S) . \tag{3.8}$$

(3.8)은 (3.5)로부터 성립함을 알 수 있다.

따라서, 일반적으로 k개의 모집단 단위들 $A_k = \{i_1, \cdots, i_k\}$을 표본으로 포함하는 k차 포함확률은 다음과 같다.

$$\pi_{i_1,\cdots,i_k} = \left(\prod_{l \in A} w_l\right) \cdot R(n-k, A^c)/R(n, S) . \tag{3.9}$$

성질 2.

포함확률 π_i, π_{ij}는 다음이 성립한다.

$$\sum_{i=1}^{N} \pi_i = n, \quad \sum_{j \neq i}^{N} \pi_{ij} = (n-1)\pi_i , \quad \sum_{i}^{N}\sum_{j>i}^{N} \pi_{ij} = \frac{1}{2} n(n-1) . \tag{3.10}$$

성질 3.

maximum entropy 모형에서 다음의 조건을 만족한다.

$$0 < \pi_{ij} < \pi_i \pi_j , \qquad i \neq j . \tag{3.11}$$

4. 표본선택절차

이번 절에서는 maximum entropy 모형으로부터 표본을 선택하는 절차에 대해 논의한다. 표본의 선택은 $k(k=1,\cdots,n)$ 회 선택한다. 그리고, 표본의 선택절차는 '*forward*' 방식과 '*backward*'방식이 있다. '*forward*'방식은 모집단으로부터 n 개의 단위를 표본으로 선택하는 방식이고, '*backward*'방식은 $N-n$ 단위를 모집단으로부터 제거하여 남아있는 나머지 n 개를 표본으로 선택하는 방식이다.

즉, 어떤 $\boldsymbol{x}\in D^n$ 일 때 '*forward*' 절차는 다음과 같이 표시할 수 있다.

$$p(\boldsymbol{x}) \propto \prod_{i\in A_x} w_i \propto \prod_{i\in A_x} w_i / \prod_{i\in S} w_i , \qquad A_x=\{i: x_i=1\} .$$

그리고, 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 표본선택절차는 표본크기 n에 영향을 받지 않는다. 따라서, 선택절차를 사용하는데 있어 표본크기 n이 미리 고정이 되었는지 여부에 따라 선택절차가 달라진다(Xiang-Hui Chen 등, 1994).

따라서, 표본크기 n의 고정 여부에 따라 '*forward*', '*backward*' 두 방법에 대해 각각 표본크기 n이 고정된 경우와 n이 고정되어 있지 않은 경우로 구분된다.

모집단 단위가 표본으로 선택될 확률을 구하는 절차들을 설명하기 위해, k회 표본을 추출한 후의 결과를 $A_0, A_1, \cdots, A_n$이라 하자. 그리고, $A_0=\phi$, $A_k \subset S$이다. 다음의 두 절차는 '*forward*'절차에서 n이 고정된 경우와 n이 고정되지 않은 경우이다. 그리고, '*backward*'절차는 '*forward*'의 두 절차와 마찬가지로 정의되는데, 표본이 아닌 단위들을 선택할 확률을 구한다.

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.

$$P_1(j,A_{k-1}^c)=\frac{w_jR(n-k,A_{k-1}^c\setminus\{j\})}{(n-k+1)R(n-k+1,A_{k-1}^c)}\ . \tag{3.12}$$

여기서, $w_i\propto p_i/(1-p_i)$ 일 때 관계함수 R 은

$$R(k,C)=\Pr\left(\sum_{t\in C}Z_t=k\right)\prod_{i\in C}(1+w_i)\ .$$

이고, $C(\neq\phi)\subset S$, $0\le k\le|C|$ 이면, 한 단위를 추출할 확률은

$$P_1(j,A_{k-1}^c)=\frac{1}{n-k+1}\Pr\left(Z_j=1\,|\sum_{t\in A_{k-1}^c}Z_t=n-k+1\right)\ .$$

이다. $Z_1,\cdots,Z_N$ 은 앞의 방법 3에서 정의된 독립 베르누이 시행이다.

따라서, 확률비례추출인 경우에도 maximum entropy 모형에 의해 표본으로 추출할 수 있다.

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

$$\prod_{k=1}^{n}P_1=\ (i_k,A_{k-1}^c)=\prod_{k=1}^{n}\frac{w_{i_k}R(n-k,A_k^c)}{(n-k+1)R(n-k+1,A_{k-1}^c)}$$

$$= \frac{w_{i_1} R(n-1, A_1^c)}{nR(n, A_0^c)} \cdot \frac{w_{i_2} R(n-2, A_2^c)}{(n-1)R(n-1, A_1^c)} \cdots \frac{w_{i_n} R(0, A_n^c)}{(n-n+1)R(1, A_n^c)}$$

$$= \prod_{k=1}^{n} w_{i_k} \frac{1}{n \cdot (n-1) \cdots 2 \cdot 1} \cdot \frac{R(0, A_n^c)}{R(n, S)} = \frac{1}{n!} \cdot \frac{w_{i_1} \cdot w_{i_2} \cdots w_{i_n}}{R(n, S)}$$

$$= \frac{1}{n!} \Pr(x_t = 1, t \in A_n).$$

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

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

$$\Pr(A_k = \{i_1, \cdots, i_k\}) \propto \left(\prod_{l=1}^{k} w_{i_l} \right) \frac{R(n-k, \{i_1, \cdots, i_k\}^c)}{R(n, S)} = \pi_{i_1, \cdots, i_k}. \qquad (3.13)$$

이다. 여기서, $P_1(j, A_0^c) = \pi_j / n$ 을 이용하여 (3.13)식을 (3.14)와 같은 순환적인 형식으로 표현할 수 있다.

$1 \le k \le n-1$ 이고, $j \in A_k^c$ 이면, (3.13)은 다음과 같이 나타낼 수 있다.

$$P_1(j, A_k^c) = \frac{w_{i_k} P_1(j, A_{k-1}^c) - w_j P_1(i_k, A_{k-1}^c)}{(n-k)(w_{i_k} - w_j) P_1(i_k, A_{k-1}^c)} . \qquad (3.14)$$

(3.14)를 이용하면 연속적으로 표본이 선택될 확률을 계산할 수 있다.

(3.14)를 이용하여 단위들이 표본으로 선택될 확률을 계산하는 절차는 다음과 같다.

표본선택확률의 계산절차

단계 1. π_j/n를 이용하여 $j=1,\cdots,N$까지 $P_1(j,S)$를 계산한다. 그 다음, 단위 i_1의 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. $P_1(j,A^c_{k-2})$와 $P_1(i_{k-1},A^c_{k-2})$로부터 (3.14)를 이용하여 모든 j에 대해서 $P_1(j,A^c_{k-1})$을 계산한다.

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.

표본선택확률절차를 이용하게되면, 모집단 단위가 표본으로 선택될 확률이 maximum entropy를 만족하게된다.

절차 1-1. (*backward* , n이 고정)

표본을 $m(m=1,2,\cdots,N-n)$회 추출할 때, $w_i\propto p_i/(1-p_i)$라고 하면, 한 단위 $j\in B^c_{m-1}$를 추출할 확률은,

$$P_1(j,B^c_{m-1})=\frac{1}{N-n-m+1}\Pr\left(Z_j=0\mid\sum_{t\in B^c_{m-1}}Z_t=N-n-m+1\right).$$

이다.

절차 2. (*forward* , n이 고정되어 있지 않음)

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

$$P_2(j, A_{k-1}^c) = \sum_{i=0}^{k-1} \frac{w_j R(k-i-1, A_{k-1}^c \setminus \{j\}) R(i, A_{k-1})}{(k-i) R(k, S)} \quad . \tag{3.15}$$

절차 1과 절차 2는 다르게 사용된다. 절차 1은 (3.14)를 이용하면, 절차 2보다 계산과정이 빠르지만, n이 고정되지 않은 경우에는 사용할 수 없다. 그리고 절차 2는 표본조사에서 표본이 변하는 경우에 유용하게 사용할 수 있다.

5. 표본의 대체 문제

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

IV. Maximum entropy 모형에 의한 추정량

1. Maximum entropy를 이용한 표본추출설계

모집단이 N개의 단위 $U_1, \cdots, U_N$로 구성되어 있고, 2회 연속조사를 한다고 가정하여, 처음시기의 단위들을 $X_1, \cdots, X_N$이라 하고, 두 번째 시기의 단위들을 $Y_1, \cdots, Y_N$이라 하자.

모집단으로부터 표본을 선택하기 위해 III장에서 정의된 절차들을 이용한다. 처음시기에 크기 n인 표본을 절차 1(*forward* , n이 고정)를 이용하여 추출한 다음, 처음시기의 표본으로부터 크기 u인 표본을 s_1, 크기 m인 표본을 s_2라 두자. 두 번째 시기에서 과거의 표본을 이용하기 위해, 처음시기의 표본 중 s_1은 절차 2'(*backward* , u는 고정되지 않음)를 이용하여 제거하거나, 또는 s_2를 절차 2(*forward* , m은 고정되지 않음)를 이용하여 선택한 후 두 번째 시기의 표본에 포함한다. 그 다음, 제거된 표본 크기 u인 표본을 $U-s_1$로부터 maximum entropy 표본교체절차를 이용하여 추출하고, 두 번째 시기의 표본에 포함하고, 이들 표본으로부터 두 번째 시기의 모집단 총합을 추정한다. 따라서, 처음시기와 두 번째 시기의 표본 크기는 n으로 항상 같게 된다.

2. Horvitz-Thompson 추정량

모집단의 총합 ($\sum y_i$)을 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 ($i \neq j = 1, 2, \cdots, N$).

모집단 총합 Y의 추정량 $\hat{Y}_{HT}$의 분산은 다음과 같다.

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

Horvitz-Thompson 추정량을 이용하여 1절에서 정의된 표본추출설계방법에 의해 선택된 표본들로부터 2회 연속조사에서 모집단 특성인 현재시점의 모집단 총합 즉, 두 번째 시기의 모집단 총합을 추정하기 위한 추정량을 구하기 위해 처음시기의 모집단 총합과 두 번째 시기의 모집단 총합의 추정량을 Horvitz-Thompson 추정량을 이용하여 정의하면 다음과 같다.

$$\hat{X}_k = \sum_{i \in s_k} \frac{X_i}{\pi_i(s_k)} \cdot t_i , \quad k = 1, 2 , \tag{4.1}$$

$$\hat{Y}_k = \sum_{i \in s_k} \frac{Y_i}{\pi_i(s_k)} \cdot t_i , \quad k = 2, 3 . \tag{4.2}$$

$$\pi_i(s_k) = \Pr\{ i \in s_k \} , \quad i = 1, \cdots, N ,$$

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

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

$\pi_i(s_k)$는 maximum entropy 모형으로부터 i번째 단위가 표본 s_k에 포함되는 포함확률이고, $\pi_{ij}(s_k)$는 maximum entropy 모형으로부터 i번째 단위와 j번째 단위가 표본 s_k에 동시에 포함되는 포함확률이다.

$\hat{X}_k$, $\hat{Y}_k$는 각각 처음시기의 모집단 총합 X와 두 번째 시기의 모집단 총합 Y의 추정량이고, 각각은 불편추정량임을 다음의 정리들로부터 알 수 있다.

정리 4-1. $\pi_i(s_2) > 0$, $i = 1, \cdots, N$이면,

$$\hat{X}_2 = \sum_{i \in s_2} \frac{X_i}{\pi_i(s_2)} \cdot t_i .$$

이고, X의 불편추정량이다. 그리고 분산은 다음과 같다.

$$V(\hat{X}_2) = \sum_{i=1}^{N} \sum_{j>i}^{N} (\pi_i(s_2)\pi_j(s_2) - \pi_{ij}(s_2)) \left(\frac{X_i}{\pi_i(s_2)} - \frac{X_j}{\pi_j(s_2)} \right)^2 . \qquad (4.3)$$

증명. t_i, $(i = 1, \cdots, N)$는 i단위가 s_2표본에 포함되면 $t_i = 1$, 표본에 포함되지 않으면 $t_i = 0$을 취하는 확률변수이다. 따라서 t_i는 다음의 성질을 갖는다.

$$E(t_i) = 1 \cdot \pi_i(s_2) + 0 \cdot (1 - \pi_i(s_2)) = \pi_i(s_2) ,$$
$$V(t_i) = \pi_i(s_2) \cdot (1 - \pi_i(s_2)) ,$$
$$Cov(t_i, t_j) = E(t_i t_j) - E(t_i) \cdot E(t_j) = \pi_{ij}(s_2) - \pi_i(s_2)\pi_j(s_2) .$$

여기서, i단위와 j단위가 동시에 표본으로 선택되면 $t_i t_j = 1$, 동시에 표본으로 선택되지 않으면 $t_i t_j = 0$ 이므로, $E(t_i t_j) = \pi_{ij}(s_2)$ 이다.

따라서, X_i를 계수로 보면, t_i는 확률변수이므로,

$$E(\hat{X}_2) = E\left\{\sum_{i=1}^{N} \frac{X_i}{\pi_i(s_2)} t_i\right\} = \sum_{i=1}^{N} \frac{X_i}{\pi_i(s_2)} E(t_i)$$

$$= \sum_{i=1}^{N} \frac{X_i}{\pi_i(s_2)} \pi_i(s_2) = X$$

그러므로, $\hat{X}_2$는 모집단 총합 X의 불편추정량이다.

X의 불편추정량 $\hat{X}_2$의 분산은 다음과 같다.

$$V(\hat{X}_2) = V\left\{\sum_{i=1}^{N} \frac{X_i}{\pi_i(s_2)} t_i\right\}$$

$$= \sum_{i=1}^{N}\left(\frac{X_i}{\pi_i(s_2)}\right)^2 V(t_i) + 2\sum_{i=1}^{N}\sum_{j>i}^{N} \frac{X_i}{\pi_i(s_2)} \frac{X_j}{\pi_j(s_2)} Cov(t_i, t_j)$$

$$= \sum_{i=1}^{N} \frac{(1-\pi_i(s_2))}{\pi_i(s_2)} X_i^2 + 2\sum_{i=1}^{N}\sum_{j>i}^{N} \frac{\pi_{ij}(s_2) - \pi_i(s_2)\pi_j(s_2)}{\pi_i(s_2)\pi_j(s_2)} X_i X_j$$

여기서, maximum entropy의 성질 2. 중에서 $\sum_{j\neq i}^{N} \pi_{ij} = (n-1)\pi_i$를 이용하면,

$$\sum_{j\neq i}^{N} \{\pi_{ij}(s_2) - \pi_i(s_2)\pi_j(s_2)\} = \sum_{j\neq i}^{N} \pi_{ij}(s_2) - \sum_{j\neq i}^{N} \pi_i(s_2)\pi_j(s_2)$$

$$= \{(m-1)\pi_i(s_2)\} - \pi_i(s_2)\sum_{j\neq i}^{N} \pi_j(s_2)$$

$$= \{(m-1)\pi_i(s_2)\} - \pi_i(s_2)(m-\pi_i(s_2))$$

$$= -\pi_i(s_2)(1-\pi_i(s_2)) .$$

이므로, 앞의 분산식 첫째항에 $1-\pi_i(s_2)$ 대신에 $\frac{\sum_{j\neq i}(\pi_i(s_2)\pi_j(s_2)-\pi_{ij}(s_2))}{\pi_i(s_2)}$ 를 대입하면,

$$\sum_{i=1}^{N}\frac{(1-\pi_i(s_2))}{\pi_i(s_2)}X_i^2 = \sum_{i=1}^{N}\sum_{j\neq i}^{N}(\pi_i(s_2)\pi_j(s_2)-\pi_{ij}(s_2))\left(\frac{X_i}{\pi_i(s_2)}\right)^2$$

$$= \sum_{i=1}^{N}\sum_{j>i}^{N}(\pi_i(s_2)\pi_j(s_2)-\pi_{ij}(s_2))\left(\frac{X_i}{\pi_i(s_2)}+\frac{X_j}{\pi_j(s_2)}\right)^2 .$$

여기서, $\sum_{i=1}^{N}\sum_{j\neq i}^{N}a_i^2 = \sum_{i=1}^{N}\sum_{j>i}^{N}(a_i^2+a_j^2)$ 이다.

따라서, 위 분산식은 다음과 같다.

$$V(\hat{X}_2) = \sum_{i=1}^{N}\sum_{j>i}^{N}(\pi_i(s_2)\pi_j(s_2)-\pi_{ij}(s_2))\left(\frac{X_i}{\pi_i(s_2)}+\frac{X_j}{\pi_j(s_2)}-2\frac{X_iX_j}{\pi_i(s_2)\pi_j(s_2)}\right)^2$$

$$= \sum_{i=1}^{N}\sum_{j>i}^{N}(\pi_i(s_2)\pi_j(s_2)-\pi_{ij}(s_2))\left(\frac{X_i}{\pi_i(s_2)}-\frac{X_j}{\pi_j(s_2)}\right)^2 .$$

정리 4-1 의 방법을 이용하면, 다음의 정리들 4-2, 4-3, 4-4 가 성립함을 알 수 있다.

정리 4-2. $\pi_i(s_2)>0$, $i=1,\cdots,N$ 이면,

$$\hat{Y}_2 = \sum_{i \in s_2} \frac{Y_i}{\pi_i(s_2)} \cdot t_i$$

이고, Y의 불편추정량이다. 그리고 분산은 다음과 같다.

$$V(\hat{Y}_2) = \sum_{i=1}^{N} \sum_{j>i}^{N} (\pi_i(s_2)\pi_j(s_2) - \pi_{ij}(s_2)) \left(\frac{Y_i}{\pi_i(s_2)} - \frac{Y_j}{\pi_j(s_2)} \right)^2 . \qquad (4.4)$$

정리 4-3. $\pi_i(s_3) > 0$, $i = 1, \cdots, N$ 이면,

$$\hat{Y}_3 = \sum_{i \in s_3} \frac{Y_i}{\pi_i(s_3)} \cdot t_i$$

이고, Y의 불편추정량이다. 그리고 분산은 다음과 같다.

$$V(\hat{Y}_3) = \sum_{i=1}^{N} \sum_{j>i}^{N} (\pi_i(s_3)\pi_j(s_3) - \pi_{ij}(s_3)) \left(\frac{Y_i}{\pi_i(s_3)} - \frac{Y_j}{\pi_j(s_3)} \right)^2 . \qquad (4.5)$$

정리 4-4. $\pi_i(s_2) > 0$, $i = 1, \cdots, N$ 이면, $\hat{X}_2$ 와 $\hat{Y}_2$ 의 공분산 $Cov(\hat{X}_2, \hat{Y}_2)$ 는 다음과 같이 주어진다.

$$Cov(\hat{X}_2, \hat{Y}_2)$$
$$= \sum_{i=1}^{N} \sum_{j>i}^{N} (\pi_i(s_2)\pi_j(s_2) - \pi_{ij}(s_2)) \left(\frac{X_i}{\pi_i(s_2)} - \frac{X_j}{\pi_j(s_2)} \right) \left(\frac{Y_i}{\pi_i(s_2)} - \frac{Y_j}{\pi_j(s_2)} \right) . \qquad (4.6)$$

정리 4-5. $\pi_i(s_k) > 0$, $i = 1, \cdots, N$, $k = 1, 2$ 이면,

$$Cov(\hat{X}_1, \hat{X}_2)$$

$$= \frac{1}{2} \sum_{i=1}^{N} \sum_{j \neq i}^{N} (\pi_i(s_1)\pi_j(s_2) - \pi_{ji}(s_2|s_1)) \left(\frac{X_i}{\pi_i(s_1)} - \frac{X_j}{\pi_j(s_1)} \right) \left(\frac{X_i}{\pi_i(s_2)} - \frac{X_j}{\pi_j(s_2)} \right). \tag{4.7}$$

증명. $\pi_{ji}(s_2|s_1) = \Pr\{i \in s_1, j \in s_2\}$ 는 i번째 단위는 s_1 표본으로, j번째 단위는 s_2 표본으로 각각 포함될 포함확률이다. 따라서, $E(t_i t_j) = \pi_{ji}(s_2|s_1)$ 이다.

$$Cov(\hat{X}_1, \hat{X}_2) = Cov\left(\sum_{i \in s_1} \frac{X_i}{\pi_i(s_1)} t_i, \sum_{j \in s_2} \frac{X_j}{\pi_j(s_2)} t_j \right)$$

$$= E\left(\sum_{i=1}^{N} \frac{X_i}{\pi_i(s_1)} t_i \cdot \sum_{j=1}^{N} \frac{X_j}{\pi_j(s_2)} t_j \right) - X^2$$

$$= \sum_{i=1}^{N} \sum_{j \neq i}^{N} \frac{X_i}{\pi_i(s_1)} \frac{X_j}{\pi_j(s_2)} \pi_{ji}(s_2|s_1)$$

$$- \sum_{i=1}^{N} \frac{X_i}{\pi_i(s_1)} \frac{X_i}{\pi_i(s_2)} \pi_i(s_1)\pi_i(s_2)$$

$$- \sum_{i=1}^{N} \sum_{j \neq i}^{N} \frac{X_i}{\pi_i(s_1)} \frac{X_j}{\pi_j(s_2)} \pi_i(s_1)\pi_j(s_2)$$

$$= \sum_{i=1}^{N} \sum_{j \neq i}^{N} \frac{X_i}{\pi_i(s_1)} \frac{X_j}{\pi_j(s_2)} \{\pi_{ji}(s_2|s_1) - \pi_i(s_1)\pi_j(s_2))\}$$

$$- \sum_{i=1}^{N} \frac{X_i}{\pi_i(s_1)} \frac{X_i}{\pi_i(s_2)} \pi_i(s_1)\pi_i(s_2)$$

여기서,

$$\sum_{j \neq i}^{N} \{\pi_{ji}(s_2|s_1) - \pi_i(s_1)\pi_j(s_2)\}$$

$$= \{(n-u)\pi_i(s_1)\} - \pi_i(s_1)\{(n-u)\pi_i(s_1)\} = \pi_i(s_1)\pi_i(s_2)\,.$$

이므로, 앞의 공분산식 두 번째 항에 $\pi_i(s_1)\pi_i(s_2)$ 대신 대입하면, $Cov(\hat{X}_1, \hat{X}_2)$는

$$\sum_{i=1}^{N}\sum_{j\neq i}^{N}\frac{X_i}{\pi_i(s_1)}\frac{X_j}{\pi_j(s_2)}\{\pi_{ji}(s_2|s_1) - \pi_i(s_1)\pi_j(s_2)\}$$

$$-\sum_{i=1}^{N}\sum_{j\neq i}^{N}\frac{X_i}{\pi_i(s_1)}\frac{X_i}{\pi_i(s_2)}\{\pi_{ji}(s_2|s_1) - \pi_i(s_1)\pi_j(s_2)\}$$

$$= \frac{1}{2}\left\{\sum_{i=1}^{N}\sum_{j\neq i}^{N}(\pi_i(s_1)\pi_j(s_2) - \pi_{ji}(s_2|s_1))\left(\frac{X_i}{\pi_i(s_1)}\frac{X_i}{\pi_i(s_2)} + \frac{X_j}{\pi_j(s_1)}\frac{X_j}{\pi_j(s_2)}\right)\right.$$

$$\left.-2\sum_{i=1}^{N}\sum_{j\neq i}^{N}(\pi_i(s_1)\pi_j(s_2) - \pi_{ji}(s_2|s_1))\frac{X_i}{\pi_i(s_1)}\frac{X_j}{\pi_j(s_2)}\right\}$$

$$= \frac{1}{2}\sum_{i=1}^{N}\sum_{j\neq i}^{N}(\pi_i(s_1)\pi_j(s_2) - \pi_{ji}(s_2|s_1))\left(\frac{X_i}{\pi_i(s_1)} - \frac{X_j}{\pi_j(s_1)}\right)\left(\frac{X_i}{\pi_i(s_2)} - \frac{X_j}{\pi_j(s_2)}\right).$$

이다.

정리 4−5의 방법을 이용하면, 다음의 정리 4−6, 4−7이 성립함을 알 수 있다.

정리 4−6. $\pi_i(s_k) > 0$, $i=1,\cdots,N$, $k=2,3$ 이면,

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

$$= \frac{1}{2}\sum_{i=1}^{N}\sum_{j\neq i}^{N}(\pi_i(s_k)\pi_j(s_3) - \pi_{ji}(s_3|s_k))\left(\frac{X_i}{\pi_i(s_k)} - \frac{X_j}{\pi_j(s_k)}\right)\left(\frac{X_i}{\pi_i(s_3)} - \frac{X_j}{\pi_j(s_3)}\right). \tag{4.8}$$

정리 4−7. $\pi_i(s_k) > 0$, $i=1,\cdots,N$, $k=2,3$ 이면

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

$$= \frac{1}{2}\sum_{i=1}^{N}\sum_{j\neq i}^{N}(\pi_i(s_2)\pi_j(s_3) - \pi_{ji}(s_3|s_2))\left(\frac{X_i}{\pi_i(s_2)} - \frac{X_j}{\pi_j(s_2)}\right)\left(\frac{X_i}{\pi_i(s_3)} - \frac{X_j}{\pi_j(s_3)}\right). \tag{4.9}$$

V. Maximum entropy를 이용한 연속조사에서의 선형추정량

모집단의 단위들이 시간에 따라 변동하는 경우에 모집단의 특성을 추정하기 위해 표본조사를 2회 연속적으로 수행하는 경우, 관심 있는 모집단의 특성인 현재시점의 모집단 총합 즉, 두 번째 시기의 모집단 총합 Y를 추정하기 위해 처음시기의 표본과 두 번째 시기의 표본을 이용하여 다음의 선형추정량을 이용한다.

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

$\hat{X}_1, \hat{X}_2$는 처음시기의 모집단 총합 추정량의 불편추정량이고, $\hat{Y}_2, \hat{Y}_3$는 두 번째 시기의 모집단 총합 추정량의 불편추정량임을 앞에서 이미 증명하였다.

여기서, $\hat{Y}$가 두 번째 시기의 모집단의 총합 Y의 불편추정량이 되기 위한 조건은

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

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

$$\hat{Y} = a(\hat{X}_1 - \hat{X}_2) + c(\hat{Y}_2 - \hat{Y}_3) + \hat{Y}_3 . \tag{5.1}$$

따라서, $\hat{Y}$의 분산은

$$\begin{aligned} V(\hat{Y}) = {} & a^2 V(\hat{X}_1 - \hat{X}_2) + c^2 V(\hat{Y}_2 - \hat{Y}_3) + 2ac\,Cov(\hat{X}_1 - \hat{X}_2, \hat{Y}_2 - \hat{Y}_3) \\ & + 2ac\,Cov(\hat{X}_1 - \hat{X}_2, \hat{Y}_3) + 2c\,Cov(\hat{Y}_2 - \hat{Y}_3, \hat{Y}_3) + V(\hat{Y}_3) \end{aligned} \tag{5.2}$$

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.

$$V(\hat{Y}) = a^2 V_X + c^2 V_Y + V(\hat{Y}_3) + 2ac V_{XY} + 2aC_{XY} + 2cC_{YY}. \qquad (5.3)$$

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

$$\frac{\partial V(\hat{Y})}{\partial a} = 2\{aV_X + cV_{XY} + C_{XY}\} = 0,$$
$$\frac{\partial V(\hat{Y})}{\partial c} = 2\{cV_Y + aV_{XY} + C_{YY}\} = 0.$$

따라서, 최적값 a_0, c_0는

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

이다. 그리고, 처음시기와 두 번째 시기의 상관계수는

$$\rho = \frac{Cov(\hat{X}_1 - \hat{X}_2,\ \hat{Y}_2 - \hat{Y}_3)}{\sqrt{V(\hat{X}_1 - \hat{X}_2)\, V(\hat{Y}_2 - \hat{Y}_3)}} = \frac{V_{XY}}{\sqrt{V_X V_Y}} \qquad (5.4)$$

이다.

따라서, 최적값 a_0 , c_0 를 상관계수를 이용하여 다시 쓰면,

$$a_0 = \frac{\rho\sqrt{\frac{V_X}{V_Y}}\, C_{YY} - C_{XY}}{(1-\rho^2)V_X} \ , \ c_0 = \frac{\rho\sqrt{\frac{V_Y}{V_X}}\, C_{XY} - C_{YY}}{(1-\rho^2)V_Y} \tag{5.5}$$

이다. 최적값 a_0 , c_0 를 (5.1)에 대입하면 다음의 최적 추정량 $\hat{Y}_0$ 얻을 수 있다.

$$\hat{Y}_0 = \frac{\rho\sqrt{\frac{V_X}{V_Y}}\, C_{YY} - C_{XY}}{(1-\rho^2)V_X}(\hat{X}_1 - \hat{X}_2) + \frac{\rho\sqrt{\frac{V_Y}{V_X}}\, C_{XY} - C_{YY}}{(1-\rho^2)V_Y}(\hat{Y}_2 - \hat{Y}_3) + \hat{Y}_3 . \tag{5.6}$$

따라서, $\hat{Y}_0$ 의 분산은 다음과 같다.

$$V(\hat{Y}_0) = \frac{1}{(1-\rho^2)V_X V_Y}\{2C_{XY}C_{YY}\sqrt{V_X V_Y}\,\rho - C_{XY}^2 V_Y - C_{YY}^2 V_X\} + V(\hat{Y}_3) . \tag{5.7}$$

여기서, 모집단의 단위가 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.

$$\hat{Z} = a\hat{Y}_2 + b\hat{Y}_3 . \tag{5.8}$$

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

$$a+b=1\,.$$

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

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

따라서, $\hat{Z}$의 분산은

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

이다. 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.

$$V(\hat{Z})=a^2V_Y+V(\hat{Y}_3)+2aC_{YY}\,. \tag{5.9}$$

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

$$\frac{\partial V(\hat{Z})}{\partial a}=2aV_Y+2C_{YY}=0\,.$$

따라서, 최적값 a_0는 다음과 같다.

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

최적값 a_0를 이용하여 최적추정량 $\hat{Z}_0$을 구하면,

$$\hat{Z}_0 = -\frac{C_{YY}}{V_Y}(\hat{Y}_2 - \hat{Y}_3) + \hat{Y}_3 . \tag{5.10}$$

이고, 분산은 다음과 같다.

$$V(\hat{Z}_0) = -\frac{C_{YY}^2}{V_Y} + V(\hat{Y}_3) . \tag{5.11}$$

모집단의 단위가 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,

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

이다.

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.

VI. 결론 및 요약

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. 본 논문에서는 모집단의 단위가 시간에 따라 변동하는 표본추출에서 연속조사를 2회로 가정하고, maximum entropy 모형을 이용하여 표본을 추출한 다음, 두 번째 시기의 모집단 총합을 추정하는 선형추정량 $\hat{Y}_0$ (5.6)을 제시하였다. 그리고, 추정량에 대한 분산 (5.7)을 제시하였다.

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

시간에 따라 변하는 모집단을 연속적으로 표본조사 하는 경우, maximum entropy 모형을 이용한 표본추출에서 새로운 단위들로 표본을 교체할 때, 표본교체비율을 고려하면, 최적의 표본교체비율을 찾을 수 있고, 최적의 표본교체비율에 의해 $\hat{Y}_0$ 의 분산은 더 작아질 것이다.

VII. 참고문헌

Brown. L. D.(1986). Fundamentals of Statistical Exponential Families(with Applications in Statistical Decision Theory). Hayward, CA: Institute of Mathematical Statistics. p.74.

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

Darroch, J. N. and Ratcliff, D.(1972). Generalized iterative scaling for log-linear models. Ann. Math. Statist. 43, pp1470-1480.

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.

김규성.(1990). SAMPLING ON TWO SUCCESSIVE OCCASIONS WITH UNEQUAL PROBABILITIES. 서울대 이학석사 학위논문.

김규성.(1994). Comparison of estimators with Unequaled Selection Probability on Two Successive Occasions. 서울대 이학박사 학위논문.

박혜경.(1989). A STUDY ON ESTIMATION WITH UNEQUAL PROBABILITY FOR TWO SUCCESSIVE OCCASIONS. 서울대 이학석사 학위논문.

감사의 글

본 논문이 완성되기까지 바쁘신 와중에도 시간을 쪼개가며 지도를 해주신 김익찬 교수님에 감사를 드립니다. 그리고, 대학원생활에 많은 도움을 주신 김철수 교수님을 비롯하여 대학원 수업을 하면서 열성으로 가르침을 주셨던 이봉규 교수님, 박경린 교수님, 이정훈 교수님 그리고 바쁜 시간에도 많은 말씀을 아끼지 않으셨던 김진효 교수님께도 감사드립니다.

항상 친형처럼 도움을 주신 진용문 조교선생님과 바쁜 시간 쪼개가며 같이 공부하고, 또는 친형처럼 질타를 아끼시지 않으셨던 김봉모 선생님께도 감사 드리며, 밤샘을 할 때 항상 곁에 있어준 길남에게 고맙다는 말을 전해주고 싶습니다.