信賴性應用硏究
제3권, 제2호, pp. 117-125, 2003

사전정보가 있는 경우 다중층화를 이용한 모수추정연구*

이해용

성신여자대학교 통계학과

A Study of Parameter Estimation with the Prior-Information by Using the Multiple Stratification

Hae-Yong Lee

Department of Statistics Sungshin Women's University

Abstract

In sampling survey, prior-information about population has been generally ignored to estimate parameters. But if there is some believable prior-information about population, it is very useful to get more efficiency estimators by using the prior-information. This paper shows how to estimate the parameter, to evaluate the variance of the estimator, and to un-biasness of the estimator by using multiple stratification with prior-information about survey population. The proposed method is illustrated with a set of hypothetical data. The results show that the proposed estimator is very efficiency and strongly recommendable.

* 이 논문은 2002년도 성신여자대학교 학술연구조성비 지원에 의하여 연구되었음.

1. 서 론

표본조사의 실시 과정을 간단히 기술하면 먼저 관심의 대상이 되는 모집단을 구성하고 구성된 모집단으로부터 추출을 할 수 있는 목록(list)을 작성하여 목록으로부터 원하는 크기의 표본을 추출한 다음 해당되는 표본으로부터 연구목적에 맞는 자료를 조사한다. 이 경우 모집단에 대한 사전지식은 없는 것으로 가정하는 것이 일반적이다. 그러나 어떤 조사이든지 조사대상이 되는 모집단에 대한 정보가 전무한 경우는 그리 흔하지 않다. 특히 정보화시대에 접어들면서 정보의 양은 그만큼 더 많이 공개되고 보존되고 있다. 예로 실업률이나. 여론조사나 소득추계와 같은 경우에도 기존의 조사 자료나 유사연구에서 조사된 자료가 있어 모집단에 대한 일부를 사전에 알 수 있는 경우가 있다. 또한 일부 조사된 자료나 알려진 자료도 얼마든지 있을 수 있다. 이와 같이 알려진 정보는 경우에 따라서 얼마든지 활용될 수 있다. 특히 남여 성별과 같이 한번 조사된 자료의 성격의 변화가 불변인 경우에는 더욱 확실한 사전정보로 활용할 수 있다. 또한 대학생의 체중을 추정하는 문제의 경우에도 신입생에 대한 신체검사의 자료가 존재하는 경우 이 자료는 사전정보로서 사용할 수 있는 가치가 충분한 것이다. 이처럼 조사대상의 집단에 대한 사전정보가 있는 경우 이 정보를 활용하지 않는 것은 주어진 자원을 낭비하는 것과 같다. 따라서 사전정보가 있는 경우 이를 이용하여 표본조사를 실시한다면 조사비용의 절감은 물론 추정치에 대한 정도를 높일 수 있을 것이다.

따라서 본 연구에서는 모집단에 관한 사전정보가 있는 경우 다중층화(multiple stratification) 중에서 보조변수가 두개 있는 경우를 이용하여 정도 높은 추정량을 제시하고 그 신뢰도를 확인하고자 한다. 다중층화란 모집단을 층화하기위하여 두개 이상의 층화변수를 이용하여 모집단을 층화하는 방법이다. 이 방법은 본래 조사하고자 하는 여러 개의 관심변수가 두개 이상의 보조변수와 상관이 있을 경우 그 중 하나의 보조변수만으로 모집단을 층화하는 것은 좋은 추정치를 얻기에 바람직하지 못하기 때문에 고안되었다. 예를 들어 우리나라의 공식통계 생산에 많이 활용되고 있다. 대표적인 예로 노동부에서 작성하고 있는 매월노동통계조사는 산업분류와 규모를 층화변수로 활용하고 있으며, 통계청에서 작성하고 있는 어가경제조사에서는 지역과 어선규모를 층화변수로 사용하고 있다. 또한 농업기본통계조사에서는 지역과 농가 수를 층화변수로 활용하고 있다.

그 동안 층화의 문제는 많은 사람들에 의해서 연구되었으며 그 결과는 Cochran(1977), Kish(1965), DesRaj(1968), Sukhatma et al(1984), Govidarajulu(1999) 및 박 홍래(2000)에서도 이에 관한 내용이 수록되어있다. 또한 다중층화에서 층의 수가 표본의 크기보다 큰 경우 표본추출에 대한 연구는 Bryant et al.(1960)에 연구된 예가 있다. 그러나 이들의 연구는 주로 주어진 표본의 크기의 배분 및 추정에 관한 연구가 주였다. 사전정보를 층화요인으로 한 다중층화와 추정에 관한 연구는 되어있지 않았다. 특히 사전정보를 활용함으로서 추정치의 정도를 높이고 비용을 절감할 수 있다면 금상첨화가 아닐 수 없다. 따라서 본 연구에서는 다중층화에서 사정정보를 어떻게 이용할 수 있으며, 그 결과 추정식과 추정치에 대한 기대값 및 분산을 구하고 일반 층화추출과 분산의 상대효율을 보이고자 한다.

2. 사전정보가 있는 경우 층화추출에서 모수추정

일반적으로 층화추출의 경우 모집단을 층으로 나누는 작업이 중요하다. 층을 나누는 기본 원칙은 층 내에는 동질적이고, 층간에는 이질적으로 층을 나누는 것이 바람직한 것으로 알려져 있다. 그러나 원칙에 맞도록 실제로 층화를 하는 것은 쉬운 작업이 아니다. 따라서 일반적으로 표본설계 시에는 모집단을 분석한 후에 가능한 한 층화원칙에 가깝게 층을 나누거나 층화변수를 이용하고 있다. 예로 대학생의 의식조사를 층화추출을 이용하는 경우 학년 혹은 단과대학별로 층을 나누었다면 학년이나 단과대학은 층화변수이다. 또한 서울지역 가계소득조사를 층화추출을 이용하여 실시하려는 경우에는 동이나 구와 같은 행정구역별로 층을 나누었다면 동이나 구가 층화변수가 된다.

2.1 다중층화의 표본추출

모집단을 정보가 있는 경우와 없는 경우의 2개의 층으로 나누고, 또 다른 층화변수를 정하여 그 변수의 특성에 따라 L개의 층으로 나누면 층화변수가 2개이고 각 변수에 따른 층화요인이 각각 2개와 L개로 이루어진 Two-way 다중층화가 된다. 이 관계를 표로 정리하며 다음 <표 2-1>과 같다.

<표 2-1> 모집단의 층별 구성

층 별 / 정보 유무	1	2	..... h	L	합
정보 없는 부분	N_{11}	N_{21}	N_{h1}	N_{L1}	$N_{.1}$
정보 있는 부분	N_{12}	N_{22}	N_{h2}	N_{L2}	$N_{.2}$
합	N_1	N_2	N_h	N_L	N

일반적으로 이상과 같이 모집단이 층화되면 다중층화(multi-variate stratifying) 중에서 Two-way 층화라고 한다. 따라서 <표 2-1>과 같이 구성된 모집단으로부터 표본크기 n을 추출하면 표본의 구성은 다음 <표 2-2>와 같게 된다.

<표 2-2> 사전정보 유무에 따른 층화추출의 표본구성

층 별 / 정보 유무	1	2	..... h	L	합
정보 없는 부분	n_{11}	n_{21}	n_{h1}	n_{L1}	$n_{.1}$
정보 있는 부분	n_{21}	n_{22}	n_{h2}	n_{L2}	$n_{.2}$
합	n_1	n_2	n_h	n_L	n

그러나 위 <표 2-2>에서 정보가 있는 부분에 해당되는 표본은 별도로 추출할 필요가 없다. 사전정보가 있는 부분에 대해서는 이미 층 전체에 대한 값들이 알려져 있으므로 일부의 표본을 추출하는 대신 전체를 표본으로 활용해도 추가적인 비용이나 시간 및 작업량이 발생하지 않는다. 따라서 이런 경우에는 사전정보가 없는 부분에서 만 일반적인 층화추출방법과 같이 표본을 추출하면 된다. 그 결과 다음 <표 2-3>과 같다.

<표 2-3> 사전정보를 층화추출에서 표본으로 모두 활용한 경우 표본 구성

층 별 / 정보 유무	1	2	..... h	L	합
정보 없는 부분	n_{11}	n_{21}	n_{h1}	n_{L1}	$n_{.1}$
정보 있는 부분	N_{21}	N_{22}	N_{h2}	N_{L2}	$N_{.2}$
합	n'_1	n'_2	n'_h	n'_L	n'

2.2 모수의 추정

모수 중에는 여러 가지가 있으나 본 논문에서는 모집단 평균을 대상으로 설명하고자한다. 모집단평균을 $\overline{Y}$ 라하고 층화추출에서 사전정보를 고려하지 않은 모집단평균의 추정량을 $\bar{y}_{st}$ 라 하면 추정량 $\bar{y}_{st}$와 추정량의 분산 $Var(\overline{y_{st}})$는 다음 식과 같다. Cochran(1977).

$$\overline{y_{st}} = \frac{1}{N}\sum_{i=1}^{L} N_h \overline{y_h} \tag{2-1}$$

$$Var(\overline{y_{st}}) = \sum_{h=1}^{L} W_h^2 (1-f_h)\frac{S_h^2}{n_h} \tag{2-2}$$

단, $S_h^2 = \frac{1}{N_h}\sum_{i=1}^{N_h} (y_{hi} - \overline{Y_h})^2$, $\overline{y_h} = \frac{1}{n_h}\sum_{i=1}^{n_h} y_{hi}$ 로 층 h에서 구한 표본의 평균이고, $W_h = N_h/N$, $f_h = n_h/N_h$, 이다.

그러나 사전정보를 고려한 층화추출에서 모집단평균의 추정량을 $\overline{y_I}$라 하면 추정량 $\overline{y_I}$와, 그 분산 $Var(\overline{y_I})$은 다음 [정리 1]과 같다.

[정리 1]

층화추출에서 모집단에 대한 정보를 이용하여 모집단의 평균 $\overline{Y}$의 추정량 $\overline{y_I}$ 를 다음 식(2-3)과 같이 표현하면 이는 모집단평균 $\overline{Y}$ 의 불편추정량이다.

$$\overline{y_I} = \frac{1}{N}\left[\sum_{h=1}^{L} N_{h1}\overline{y_{h1}} + \sum_{h=1}^{L} N_{h2}\overline{Y_{h2}}\right] \tag{2-3}$$

또한 추정량 $\overline{y_I}$ 의 분산은 다음 식(2-4)와 같다.

$$Var(\overline{y_I}) = \sum_{h=1}^{L} W_{h1}^2 (1-f_{h1})\frac{S_{h1}^2}{n_{h1}} \tag{2-4}$$

여기서 $W_h = N_h/N$, $W_{h1} = N_{h1/N}$, $f_{h1} = n_{h1}/N_{h1}$, $S_{h1}^2 = \frac{1}{N_{h1}-1}\sum_{j=1}^{N_{h1}} (y_{h1j} - \overline{Y_{h1}})^2$, 이다.

[증명]

(i) 추정량의 기대 값

식 (2-3)이 모집단평균 $\overline{Y}$의 불편추정량임을 증명하기 위하여 $\overline{y_I}$의 기대 값을 $E(\overline{y_I})$라 하면 이는 다음과 같다.

$$E(\overline{y_I}) = E[\frac{1}{N}[\sum_{h=1}^{L} N_{h1}\overline{y_{h1}} + \sum_{h=1}^{L} N_{h2}\overline{Y_{h2}}]]$$

$$= E[\frac{1}{N}[\sum_{h=1}^{L} N_{h1}\overline{y_{h1}} + T_2]]$$

$$= \frac{1}{N}[\sum_{h=1}^{L} N_{h1}E(\overline{y_{h1}}) + T_2]$$

$$= \frac{1}{N}(\sum_{h=1}^{L} N_{h1}\overline{Y_{h1}} + T_2)$$

$$= \frac{1}{N}(T_1 + T_2) = \frac{T}{N} = \overline{Y}$$

단, $\overline{y_{h1}} = \frac{1}{n_{h1}}\sum_{i=1}^{n_{h1}} y_{h1}$, $T_1 = \sum_{h=1}^{L} N_{h1}\overline{Y_{h1}}$ 및 $T_2 = \sum_{h=1}^{L} N_{h2}\overline{Y_{h2}}$ 로 각각 모집단의 정보가 알려지지 않은 총계 및 알려진 총계를 의미하고, $T = T_1 + T_2$로 모집단총계를 의미한다.

(ii) 추정량의 분산

추정량 $\overline{y_I}$ 의 분산의 식을 구하면 다음과 같다.

$$Var(\overline{y_I}) = Var[\frac{1}{N}(\sum_{h=1}^{L} N_{h1}\overline{y_{h1}} + \sum_{h=1}^{L} N_{h2}\overline{Y_{h2}})]$$

$$= \frac{1}{N^2}[(\sum_{h=1}^{L} N_{h1}^2 Var(\overline{y_{h1}}) + Var(T_2)]$$

$$= \frac{1}{N^2}\sum_{h=1}^{L} N_{h1}^2 Var(\overline{y_{h1}})$$

$$= \sum_{h=1}^{L} W_{h1}^2(1-f_{h1})\frac{S_{h1}^2}{n_{h1}}$$

3. 추정량 분산의 비교

본 장에서는 새로운 추정량의 효율성을 확인하기 위하여 일반적인 층화추출에 의한 추정량의 분산 $Var(\overline{y_{st}})$와 사전정보를 이용한 추정량의 분산 $Var(\overline{y_I})$을 비교하면 다음 [정의 2]와 같다.

[정리 2]

일반적인 층화추출에서 모집단평균의 추정치의 분산 $Var(\overline{y_{st}})$는 사전정보를 이용한 모집단평균의 추정량의 분산 $Var(\overline{y_I})$보다 크거나 같다.

[증명]

$$Var(\overline{y_{st}}) = \sum_{h=1}^{L} W_h^2(1-f_h)\frac{S_h^2}{n_h}$$

$$Var(\overline{y_I}) = \sum_{h=1}^{L} W_{h1}^2(1-f_{h1})\frac{S_{h1}^2}{n_{h1}}$$

위 두 식으로부터 어느 한 층 h에 대해서 사전정보를 사용하지 않은 경우의 층화추출에서 추정량의 분산을 $Var(\overline{y_{st,h}})$라하고 사전정보를 활용한 층화추출에서 추정량의 분산을 $Var(\overline{y_{I,h}})$라 하면 이들은 다음 식 (3-1) 및 식 (3-2)와 같다.

$$Var(\overline{y_{st,h}}) = W_{h1}^2(1-f_{h1})\frac{S_{h1}^2}{n_{h1}} + W_{h2}^2(1-f_{h2})\frac{S_{h2}^2}{n_{h2}} \quad (3\text{-}1)$$

$$Var(\overline{y_{I,h}}) = W_{h1}^2(1-f_{h1})\frac{S_{h1}^2}{n_{h1}} \quad (3\text{-}2)$$

두 식으로부터 $Var(\overline{y_{st,h}}) \geq Var(\overline{y_{I,h}})$임을 알 수 있다. 따라서 $Var(\overline{y_{st}}) = \sum_{h=1}^{L} Var(\overline{y_{st,}h})$ 이고 $Var(\overline{y_I}) = \sum_{h=1}^{L} Var(\overline{y_{i,h}})$ 이므로 $Var(\overline{y_I}) \leq Var(\overline{y_{st}})$이다.

4. 예 제

위에서 설명한 추정량에 대한 효율성을 살펴보기위하여 중학교의 체중조사를 예로 들어 보기로 한다. 중학교에서는 1년에 한번 정도의 체중조사를 실시한다. 그런데 1학년의 경우 입학당시 신체검사를 통하여 체중을 조사하고 그 자료는 잘 정리되어 있다. 따라서 학년을 제 1 층화변수로 하여 정보가 알려진 1학년을 층1로 하고 알려지지 않은 2학년과 3학년을 다른 층2로 구분하고, 거주지를 제 2의 층화변수로 하여 거주지역을 4개의 층으로 나누어

각 층으로부터 일부를 추출하여 학생들의 평균체중을 추정하는 문제를 생각하면 이는 앞에서 언급한 추정량을 이용할 수 있는 좋은 예라 할 수 있다.

다음 <표 4-1>은 1학년을 층1로 2학년과 3학년을 층 2로 하고 또한 각 전 학생을 4개의 층으로 구분하여 정리한 자료로 1학년 학생들의 체중은 알려있다고 가정 한다. 그리고 총 60명을 각 층으로부터 비례배분으로 추출하는 것으로 하고 사전정보를 이용하는 경우와 이용하지 않는 경우의 모집단평균의 추정량과 그 분산을 구해 보기로 한다. 단, 각 학년별로는 5명씩 배당하여 추출하기로 한다.

<표 4-1> A중학교의 체중조사 표

학년 \ 거주지	A	B	C	D
층 1(1학년) (정보 유)	53, 55, 66, 72, 64, 52, 49, 75, 60, 83, 64, 50, 54, 58, 69, 47, 55, 63, 58, 46,	50, 48, 62, 60, 72, 45, 55, 48, 50, 52, 64, 68, 61, 60, 70, 55, 58, 69, 61, 69	69, 48, 62, 64, 53, 55, 47, 58, 60, 70, 75, 62, 61, 51, 43, 44, 58, 55, 62, 54,	48, 57, 73, 47, 65, 60, 71, 83, 46, 73, 63, 54, 45, 69, 57, 70, 64, 69, 55, 47,
층 2(2,3학년) (정보 무)	63, 71, 56, 58, 53, 59, 66, 73, 79, 85, 51, 72, 81, 50, 48, 67, 70, 54, 70, 49,	45, 62, 53, 55, 58, 60, 50, 46, 71, 63, 59, 63, 46, 72, 50, 45, 63, 60, 50, 49,	70, 63, 54, 63, 48, 46, 56, 55, 61, 60, 52, 50, 48, 79, 68, 59, 67, 74, 47, 58	55, 81, 72, 42, 55, 58, 62, 62, 57, 51, 60, 63, 52, 44, 58, 65, 62, 50, 52, 56,
	70, 67, 48, 63, 60, 52, 88, 90, 55, 72, 64, 71, 65, 64, 61, 63, 74, 47, 65, 60,	72, 66, 60, 52, 55, 61, 60, 78, 46, 44, 59, 61, 48, 58, 50, 51, 55, 66, 74, 53,	54, 55, 65, 63, 50, 58, 68, 76, 62, 82, 56, 63, 47, 67, 44, 77, 63, 59, 50, 65,	72, 56, 56, 62, 48, 86, 72, 48, 55, 69, 51, 64, 66, 70, 77, 51, 53, 50, 61, 63,

위 자료로부터 각 $Var(\overline{y_{st}})$와 $Var(\overline{y_I})$의 값을 구하면 다음과 같다.

$$Var(\overline{y_{st}}) = \sum_{h=1}^{L} W_h^2(1-f_h)\frac{S_h^2}{n_h} = 1.18495$$

$$Var(\overline{y_I}) = \sum_{h=1}^{L} W_{h1}^2(1-f_{h1})\frac{S_{h1}^2}{n_{h1}} = 0.80837$$

따라서 $Var(\overline{y_{st}}) > Var(\overline{y_I})$ 이고, 이의 상대효율 RE = 1.4658 이다.

5. 결론 및 제언

자원의 재활용이 적극 권장되고 있는 요즈음 비록 무형의 재화라 할지라도 알려져 있는 자료를 버리지 않고 재활용하는 것은 경제성을 떠나서 큰 의미를 가지고 있다. 더불어 사전정보를 활용해서 추정량의 효율성까지 높일 수 있다면 이는 금상첨화가 아닐 수 없다. 특히 이방법의 가장 큰 장점은 동일한 비용으로 더 많은 표본을 사용할 수 있어 그 만큼 추정량의 효율을 높일 수 있다는 것이다. 예제에서 살펴본바와 같이 일반적인 층화추출의 경우에 비해 사전정보를 이용한 추정량의 상대효율이 1.4658로 나타나고 있다. 이는 한 예이지만 사전정보를 이용함으로써 기존의 추정방법에 비하여 46.58% 이상의 효율을 높일 수 있는 것을 보여주고 있다. 이는 사전에 알려진 정보는 모두 표본으로 활용하여 추정량을 구하기 때문에 추정량의 효율이 높아지는 것이다. 물론 이 방법을 사용하기 위해서는 사전에 정보가 있어야 한다는 전제조건과 사전정보가 믿을 수 있는 것이어야 한다. 정보가 불확실한 경우에는 결과를 믿을 수 없는 불신이 있을 수 있다. 따라서 사전정보가 신뢰성이 있는지를 사전에 확인하는 것은 매우 바람직할 것이다. 사전정보의 신뢰성을 확인하는 방안의 하나로 사전정보와 사전정보에 해당하는 일부분을 조사를 통하여 두 정보의 상관관계를 구해봄으로써 가능할 것이다. 이에 대한 구체적인 방안은 차후 연구가 더 요구된다. 부수적인 것으로는 사전정보의 유용성이 확인됨으로서 자료의 보관과 관리에도 획기적인 전기가 마련될 것으로 기대된다.

참고문헌

[1] 박 홍래(2000), 통계조사론, 영지문화사.

[2] 통계청(2002), 한국통계조사현황(상, 하), 통계청

[3] Bryant, E. C., Hartley, H. O. and Jessen, R. J.(1968), Design and Estimation in Two-Way Stratification, J. *American Statistical Association*, 55, 105-124.

[4] DesRaj(1968), *Sampling Theory*, New York : McGraw-Hill Book Co..

[5] Sukhatme, P. V., Sukhatme, S., and Ashok, C.(1984), *Sampling Theory of Survey with Applications*, 3rd ed. Ames, Iowa: Iowa State University Press.

[6] Cochran, W. G.(1977), *Sampling Techniques*, 3rd ed., John Wiley & Sons, New York.