The Korean Communications
in Statistics Vol. 6, No. 3, 1999
pp. 755-760

A Post-stratified Estimation in Multivariate Stratified Sampling Surveys

Jinwoo Park[1)]

Abstract

In multivariate stratified sampling surveys, it is general to use a few stratification variables which are highly correlated with the important variables at design stage. But, there might be some secondary study variables which are not so highly correlated with those stratification variables. In that case, it is not efficient to use the same type of estimator due to the secondary variables as the one based on the important variables. A post-stratified estimation is proposed to increase the efficiency of the estimator with existence of secondary variables. The proposed method is illustrated with a set of fishery household population survey data.

1. 서 론

하나의 층화표본에서 여러 가지 변수들을 동시에 조사하는 조사를 다변량 층화 표본조사 (multivariate stratified sampling survey)라고 한다 (Bethel J., 1989). 이 경우 표본설계를 위해 모집단을 층화하고자 할 때 여러 가지 변수들을 모두 고려하여 층화한다면 층의 수가 지나치게 많아지게 되어 경우에 따라서는 해당 층에 속하는 단위들이 거의 없는 경우도 생긴다. 따라서 실제 설계시에는 모든 변수들을 다 고려하기 보다는 여러 변수들 중에서 가장 중요한 관심을 갖는 몇 개의 변수들을 선택하여 층화를 하는 것이 보통이다. 이렇게 하면 일부 2차적인 관심변수들은 층화 과정에서 제대로 고려되지 못한다.

다변량 층화 표본조사에 대한 기존의 연구들은 주로 표본설계에 관한 연구 (박홍래, 1987; 이기재 외, 1997; 대한통계협회, 1997; 김규성, 1998)와 일정한 표본의 크기가 주어졌을 때 여러 가지 관심 변수들을 동시에 고려하는 표본의 배분법에 관한 연구 (Kish, L., 1976; Bethel, J. W., 1985, 1989; Rahim, M.A.와 Currie, S., 1993)에 집중되어 왔다. 위의 연구들에 나타난 추정 방법을 관찰해 보면 각각의 변수가 층화 과정에서 어느 정도 고려되었는 지에 상관없이 모든 변수에 대해 동일한 추정식을 적용하도록 하고 있다. 만일 그렇게 한다면 층화 과정에서 충분히 고려된 중요 관심변수들인 경우 추정의 효율이 높을 것으로 예상되지만, 층화시에 제대로 고려되지 않은 2차적인 관심변수들은 층화의 효과가 미미할 것이므로 추정의 효율이 떨어질 우려가 있다.

해양수산부에서 발표하는 어업기본통계 표본조사의 경우가 다변량 층화 표본조사의 전형적인 예이다(대한통계협회, 1997). 이 조사에서 가장 중요한 관심은 전국의 어가수를 정도높게 추정하

1) Assistant Professor, Department of Applied Statistics, Suwon University, Hwasung-Gun, Kyonggi, 445-743, KOREA.

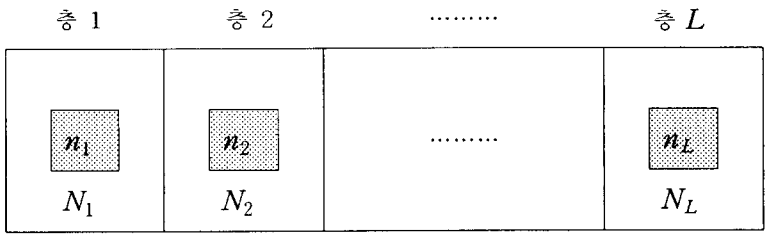
는 것이며, 그 밖에도 전업어가수나 겸업어가수를 파악하는 것도 관심의 대상이 된다. 이 조사를 위한 표본설계는 가장 중요한 관심변수인 어가수를 중심으로 이루어져 있다. 층화 및 표본크기의 결정은 각 조사구의 1995년 어가수를 기준으로 하여 이루어졌다. 어가수가 전업어가수나 겸업어가수의 특성을 잘 반영한다면 별 문제가 없지만 그렇지 않을 경우 전업어가수나 겸업어가수의 추정 효율은 낮을 수 있다. 참고로 2차적인 관심변수인 전업어가수와 1차 관심변수인 어가수 간의 상관계수는 0.4647으로 나타났다.

본 연구는 다변량 층화 표본조사에서 여러 가지 관심변수들의 추정에 있어 사후층화(post-stratification) 방법을 이용한 추정식을 제시하는 것을 목적으로 한다. 표본설계를 위한 층화 과정에서 주로 고려된 변수들을 추정할 때에는 사후층화를 하지 않아도 되지만, 층화 시에 충분히 고려되지 못한 2차적 관심변수들을 추정할 때에는 먼저 이를 각 변수의 특성에 따라 사후층화를 한 후 그에 적합한 추정식을 사용함으로써 추정의 효율을 높일 수 있다.

2. 사후층화를 이용한 추정

2.1 기본사항

일반적으로 다변량 조사를 위한 설계에서 층화는 주 관심변수에 대한 추정의 효율을 높일 수 있도록 하기 위해 주 관심변수와 상관이 높은 보조정보들을 활용하여 이루어지게 된다. 가령 전체 모집단을 L개의 층으로 구분하였다고 가정하자. 각 층 i, $(i=1,2,\cdots,L)$ 에는 N_i 개의 모집단 단위들이 들어 있으며 N 은 모집단 단위들의 합계로서 $N=\sum N_i$ 이다. 각 층 i 로부터 단순임의표집에 의해 n_i 개씩의 표본을 추출하며 전체 표본의 크기는 $n=\sum n_i$ 으로 나타낸다. 이러한 사항을 그림으로 표시한 것이 아래의 <그림 1>이다.



< 그림 1> 층화된 모집단과 표본

위와 같은 층화표본에 대한 모평균의 비편향 추정량과 분산의 식이 아래의 (1), (2) 식에 소개되어 있다.

$$\widehat{\overline{X}}=\sum_{h=1}^{L} W_h\overline{x_h}=\sum_{h=1}^{L} W_h\frac{1}{n_h}\sum_{i=1}^{n_h}x_{hi} \tag{1}$$

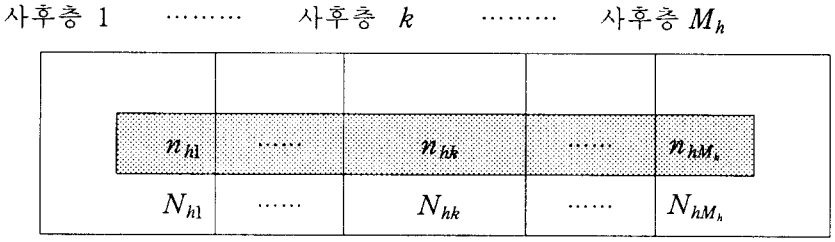
$$Var(\widehat{\overline{X}})=\sum_{h=1}^{L} W_h^2(1-f_h)\frac{S_h^2}{n_h} \tag{2}$$

$W_h=\frac{N_h}{N}$, $f_h=\frac{n_h}{N_h}$, $S_h^2=\frac{1}{N_h-1}\sum_{i=1}^{N_h}(x_{hi}-\overline{X_h})^2$, $\overline{X_h}=\frac{1}{N_h}\sum_{i=1}^{N_h}x_{hi}$ 이다.

2.2 사후층화

모든 관심변수들의 특성을 충분히 반영하여 층화가 이루어졌다면 각 관심변수에 대한 모평균의 추정량과 그 분산의 식으로 위에서 소개한 추정식 (1), (2)를 사용하면 된다. 그러나 만일 주 관심변수들만 고려해서 층화가 이루어진 경우, 2차 관심변수들 중에는 층화의 기준이 된 보조변수들과 상관의 정도가 낮은 것들이 있을 수 있다. 따라서 이러한 변수들의 측면에서 본다면 효과적인 층화가 이루어지지 않은 것이므로 추정식 (1), (2)를 사용하는 것은 비효율적일 수 있다. 층화 과정에 제대로 반영되지 못한 관심변수들에 관한 추정의 효율을 높이기 위해 사후층화(post stratification)를 이용할 수 있다.

2차 관심변수인 X를 잘 반영할 수 있는 보조변수를 이용하여 기존의 h층을 다시 $M_h, (h=1,2,\cdots,L)$ 개의 사후층으로 구분한 것을 나타내는 그림이 아래의 <그림 2>이다. 두 번째 첨자 k 는 h 층 내에서 k번째 사후층을 의미하는 첨자로서 1에서 M_h 사이의 값을 가지게 된다. 각각의 사후층 k 에는 N_{hk} 개의 모집단 단위, n_{hk}개의 표본단위들이 포함되어 있다고 하자. 아래 그림에서 짙은 색으로 표시된 부분은 표본을 의미한다.



<그림 2> h 번째 층에서의 사후층화

아래의 [정리]는 사후층화를 이용할 때의 모평균의 추정량과 추정량의 분산의 식을 나타내고 있다.

[정리] L 개의 층 각각에 대해 다시 $M_h\,(h=1,2,\cdots,L)$ 개씩의 층으로 사후층화하는 경우, 각 사후층에 포함되는 모집단 단위들의 수를 N_{hk} , 표본 단위들의 수를 n_{hk} ($k=1,2,\cdots,M_h; h=1,2,\cdots,L$)로 나타낸다면 식 (3)의 추정량은 모평균 $\overline{X}$ 의 불편 추정량이다.

$$\begin{aligned}\widehat{\overline{X}^*} &= \sum_{h=1}^{L} W_h\,\overline{x_h^*} \\ &= \sum_{h=1}^{L} W_h \sum_{k=1}^{M_h} \frac{Q_{hk}}{n_{hk}} \sum_{j=1}^{n_{hk}} x_{hkj}\ .\end{aligned} \tag{3}$$

또한 위 추정량의 분산의 식은 다음의 (4)식과 같다.

$$\begin{aligned}Var(\widehat{\overline{X}^*}) &= \sum_h W_h^2\, Var(\overline{x_h^*}) \\ &= \sum_h W_h^2 \left[\frac{1-f_h}{n_h}\left(\sum_k Q_{hk}\, S_{hk}^2 + \frac{1}{n_h}\sum_k S_{hk}^2(1-Q_{hk})\right)\right]\ .\end{aligned} \tag{4}$$

여기서 $f_h = n_h/N_h$, $S_{hk}^2 = \frac{1}{N_{hk}} \sum_{j=1}^{N_{hk}} (x_{hkj} - \overline{X_{hk}})^2$, $\overline{X_{hk}} = \frac{1}{N_{hk}} \sum_{j=1}^{N_{hk}} x_{hkj}$, $Q_{hk} = N_{hk}/N_h$

를 나타낸다.

<증명> 먼저 (3)식의 추정량이 불편 추정량임은 다음과 같이 증명할 수 있다.

$$\begin{aligned} E(\widehat{\overline{X}^*}) &= E(\sum_{h=1}^{L} W_h \sum_{k=1}^{M_h} \frac{Q_{hk}}{n_{hk}} \sum_{j=1}^{n_{hk}} x_{hkj}) \\ &= E[E_{n_{hk}}(\sum_{h=1}^{L} W_h \sum_{k=1}^{M_h} \frac{Q_{hk}}{n_{hk}} \sum_{j=1}^{n_{hk}} x_{hkj} \mid n_{hk})] \\ &= E[\sum_{h=1}^{L} W_h \sum_{k=1}^{M_h} Q_{hk} \overline{X_{hk}}] = \overline{X}. \end{aligned}$$

다음으로 분산의 식을 구하면

$$\begin{aligned} Var(\widehat{\overline{X}^*}) &= Var(\sum_{h=1}^{L} W_h \overline{x_h^*}) \\ &= \sum_h W_h^2 \, Var(\overline{x_h^*}) \end{aligned} \tag{5}$$

이 되므로 $Var(\overline{x_h^*})$의 식을 유도하는 것이 필요하다. 그런데 이 식은 단순임의표본에서 사후추정을 이용한 모평균 추정량의 분산의 식이므로 아래와 같이 구해질 수 있다 (Hansen, M. H. 외, 1953).

$$Var(\overline{x_h^*}) = \frac{1-f_h}{n_h}[\sum_k Q_{hk} \, S_{hk}^2 + \frac{1}{n_h} \sum_k S_{hk}^2 (1-Q_{hk})]$$

이 식을 위의 (5) 식에 대입하게 되면 (4)식과 같은 분산 추정량이 구해진다. ∎

참고로 사후층화를 않는다면 $M_h = 1$이 되어 (3)식의 추정량은 기존의 추정량인 (1)식과 같게 된다. 따라서 사후층화를 이용한 추정량은 사후층화를 사용않는 추정량을 포함하는 보다 일반적인 형태의 식이라고 볼 수 있다. 한편 여러 가지 변수들에 대해 추정을 하고자 할 때에는 변수에 따라 각각에 맞는 사후층화를 하여야 하며 사후층화의 기준이나 층의 크기 등도 변수에 따라 달리 결정되어야 한다. 뿐만 아니라 사후층화를 했을 때 각 층별 모집단의 크기인 N_{hk}의 값도 알려져 있어야 한다.

3. 예 제

위에서 소개한 추정방법의 효율성을 살펴보기 위해 어업기본통계조사의 예를 들어 보자. 어업기본통계조사는 매년 1회씩 이루어지는 조사로서 전국의 인구주택 조사구들 중 1995년 어업총조사 당시 1가구 이상의 어가를 포함하는 조사구들을 모집단으로 하고 있으며 이는 층화 표본조사로서 어가수가 주 관심변수, 전입어가수와 겸업어가수가 2차 관심변수인 조사로 볼 수 있다.

어업기본통계조사 표본설계에서 층화는 1995년 어업총조사 당시의 조사구 당 어가수의 크기에 의해 이루어졌는데, 어가수가 10가구 미만인 층 1, 10-29 가구 사이인 층 2, 30가구 이상인 층 3으로 나누어졌다. 한편 전국의 표본의 크기는 405개인데, 네이만 배분법에 의해 층별로 배분되어 층 1에는 106개, 층 2에는 164개, 층 3에는 135 개의 표본이 배분되었다. 따라서 이 조사는 전적으로 주 관심변수인 어가수를 중심으로 하여 표본설계가 이루어졌음을 알 수 있다. 그런데 1995년 어업총조사 자료를 기초로 하여 2차 관심변수인 전업어가수와 주 관심변수인 어가수간의 상관계수를 구한 결과 상관계수가 0.4647인 것으로 나타났다. 따라서 기존의 층화가 전업어가수의 특성을 충분히 반영하고 있다고 보기는 어렵다.

1995년 어업총조사 자료를 이용하여 2절에서 소개한 두 추정량의 효율을 비교하기 위해서는 먼저 사후층화가 필요하므로 어가수의 크기에 따라 구분된 3개의 층 내에서 각각 전업어가수의 규모에 따라 전업어가수 2가구 이하인 층과 3가구 이상인 2개의 층으로 사후층화를 하였다. 각 층별 모집단 조사구수를 나타낸 표가 다음의 <표 1>이다. 이 표를 통해 볼 때 원래의 층화가 전업어가수의 특성을 제대로 반영하지 못하였다는 것을 확인할 수 있다.

다음으로 <표 2>에는 앞 절의 (2)와 (4) 식에 나타난 두 추정량의 분산의 식을 이용하여 각 추정량의 분산의 식을 계산한 결과가 나와 있다. 이 표를 보면 2차 관심변수인 전업어가수의 추정을 위해 주 관심변수에 대한 추정식인 (1)식을 그대로 사용하는 경우 추정량의 분산의 값이 0.0481인 반면, 사후층화를 이용한 추정식 (3)을 사용하였을 때에는 그 분산이 0.0293이 된다는 사실을 알 수 있다. 이를 상대효율로 나타내자면 1.64가 되어 사후층화를 이용한 (3)식의 추정값이 이전에 사용하던 (1)식의 추정값에 비해 64%나 더 효율적임을 알 수 있다.

<표 1> 층별 모집단 조사구수

원래의 층 \ 사후층	전업어가수 2가구 이하	전업어가수 3가구 이상	합 계
어가수 10가구 미만	2160	1437	3597
10 - 29 가구	877	1377	2254
30 가구 이상	442	721	1163
합 계	3479	3535	7014

<표 2> 전업어가수 추정식의 분산의 계산

	분산값	상대효율
주 관심변수와 동일한 추정방식 사용 ((2)식)	0.0481	RE = 1.64
사후층화를 사용한 추정방식 사용 ((4)식)	0.0293	

4. 결 론

대규모의 표본조사에서는 다변량 층화 표본조사를 이용하는 경우가 허다하다. 이 경우 일반적으로 가장 중요한 관심변수들을 중심으로 층화가 이루어지게 되므로 층화과정에서 제대로 고려되지 못하는 2차적인 관심변수들도 생겨나게 된다. 설계시 층화과정에서는 관심변수의 중요성에 따라 고려되는 정도가 다른 반면, 추정 과정에서는 모든 변수들에 대해 동일한 추정식을 적용하고 있다. 이 경우 층화 과정에서 제대로 고려되지 못한 2차 관심변수들인 경우 층화가 추정의 효율을 높이는데 크게 도움이 되지 않을 수도 있다.

본 논문에서는 다변량 층화 표본조사의 추정시에 각 변수에 따라 사후층화를 함으로써 각 변수들의 추정의 효율을 높일 수 있는 방안을 제시하였으며, 사후층화를 이용한 추정식이 기존의 추정식을 포함하는 보다 일반적인 개념임을 지적하였다. 또한 어업기본통계조사의 예를 들어 전업어가 수와 같이 2차적인 관심변수에 대한 추정에 있어서는 사후층화를 사용하는 것이 기존의 추정법에 비해 60% 이상 효율을 높인다는 결과를 보였다.

참 고 문 헌

[1] 김규성 (1998). 농가경제조사의 현황과 개선방향, 「응용통계연구」, 제11권 1호, 29-40.

[2] 대한통계협회 (1997). 「어업기본통계조사 표본설계」.

[3] 박홍래 (1987). 농업기본통계 및 가축통계조사의 표본설계연구, 「응용통계연구」, 제1권 2호, 12-20.

[4] 이기재, 전종우 (1997). 노동통계조사를 위한 표본설계 연구, 「응용통계연구」, 제10권 2호, 215-227.

[5] Bethel, J. W. (1985). An optimal allocation algorithm for multivariate surveys. *Proceeding of the social statistics section, American Statistical Association*, 209-212.

[6] Bethel, J. W. (1989). Sampling allocation in multivariate surveys, *Survey Methodology*, 15,47-57.

[7] Hansen, M. H., Hurwitz, W. N. and Madow, W. G. (1953). *Sample Survey Methods and Theory*, Vol. 2, John Wiley & Sons, New York.

[8] Kish, L. (1976). Optima and Proxima in Linear Sample Designs, *Journal of Royal Statistical Society Series A*, 139, 80-95.

[9] Rahim, M. A. and Currie, S. (1993). Optimizing sample allocation for multiple response variables, *Proceeding of the social statistics section, American Statistical Association*, 209-212.