응용통계연구 제 **14**권 2호, *2001*년, *pp. 429–437*

층화 이단계 표본추출시 최적 선택율

신민웅 [1] 오상훈 [2]

요 약

단순 이단계 표본 추출의 경우에 최적 선택률은 Hansen과 Hurwitz(1949)에 의하여 구하여 졌다. 그러나 통계청에서 실시하는 표본조사등은 층화 이단계 추출을 한다 따라서 실제적인 필요성에 의하여 층화 2단계 표본 설계를 시도 하였다. 층화 이단계 표본추출시에 주어진 비용아래서 모총계의 추정량의 분산을 최소로 하는 최적의 선택확률(optimum selection probability), 표본추출율과 부차 표본추출율을 Lagrangean 승수법에 의하여 구한다.

주요용어: 최적의 선택확률, 표본추출율, 부차 표본추출율.

1. 서론

표본조사를 할 때에 대규모의 표본일 경우에 층화 이단계 표본 추출법이 주로 사용되고 있다. 많은 사회조사에서는 행정상으로나 조사의 편의를 위하여 원소들을 동질적으로 층화한 후 인근의 조사 단위들을 같은 집락으로 묶는다. Lohr(1999)는 집락 추출을 하는 이유를 관찰치들의 명부(list)를 작성할 수 없거나, 모집단이 지리적으로 넓게 분포되어 자연스럽게 집락을 이루는 경우에 집락추출을 한다고 하였다.

Seheaffer(1990)등은 집락이 너무 많은 조사단위를 포함하고 있어서 모든 측정값을 얻을 수 없거나, 집락내 조사 단위들의 측정값이 거의 비슷하여 단지 몇 개의 조사 단위를 조사해도 전체 집락에 관한 정보를 얻을 수 있는 경우에 먼저 집락을 psu(primary sampling unit)로 확률표본 추출하고, 추출된 집락내에서 조사 단위들을 ssu(secondary sampling unit)로 이차로 추출하는 것을 이단계 표본 추출법이라고 하였다.

Thompson(1992)은 같은 수의 ssu들을 단순랜덤추출하는 것보다 이단계 표본추출의 잇점은 집락안의 ssu들을 관찰하는 것이 모집단에 널리 흩어져 있는 같은 수의 ssu들을 관찰하는 것이 더 쉽고, 비용이 덜 든다고 하였다.

이단계 집락 표본 추출을 하는데 있어서 첫 번째 문제는 집락들을 적절하게 선정하는 것이다. 이 때 다음의 두 개의 조건을 고려하여야 한다. (1) 집락 내에 있는 조사 단위들의 지리적 인접성 (2) 관리하기에 편리성등이다.

적절한 집락의 선정은 집락을 적게 추출하고, 각 집락으로부터 많은 조사 단위를 추출하기를 원하는지, 또는 집락을 많이 추출하고, 각 집락으로 부터는 조사 단위를 적게 추출

1) (449-791) 경기도 용인시 모현면, 한국외국어대학교 통계학과, 교수
E-mail: mwshin@stat.hufs.ac.kr
2) (449-791) 경기도 용인시 모현면, 한국외국어대학교 통계학과 대학원, 석사과정
E-mail: shoh@stat.hufs.ac.kr

하기를 원하는 지에 따라 달라진다. 그러나 결국 비용에 따라 결정된다. 큰 집락들은 이질적인 조사 단위들을 포함하는 경향이 있으므로 모집단 모수에 대한 정확한 추정이 요구될 때는 각 집락으로부터 많은 표본을 추출하는 것이 필요하다. 반대로 작은 집락들은 상대적으로 동질적인 조사 단위들을 포함하므로 이 경우에는 각 집락으로부터 표본을 적게 추출하여도 집락 특성에 관한 정확한 정보를 얻을 수 있다.

이 논문에서는 집락들이 층화되었을 때에 각 층에서 집락을 일차 추출단위로 뽑고, 추출된 집락내에서 다시 부차단위들을 추출하는 층화 이단계 표본추출을 설계한다. 즉,h층의 N_h개의 일차단위(집락)로 부터 n_h개의 일차단위를 추출한다. 그리고, h층의 i번째 집락의 크기가 M_{hi}인 집락에서 m_{hi}개의 이차단위(부차단위)를 추출한다.

우리는 주어진 비용아래서 모총계 Y의 추정량의 분산을 최소로하는 층화 이단계 표본추출을 할 때에, 전체적인 최적 선택 확률 (optimum selection probability), 표본 추출율과 부차 표본추출율을 구하는 문제를 생각한다. 단순 이단계 표본추출의 경우에 최적 추출율을 구하는 것은 Hansen과 Hurwitz(1949)에 의하여 논의 되었다. 2장에서는 주어진 비용아래서 전체적인 최적 선택확률, 표본추출율과 부차표본추출율을 구하는 과정을 설명한다. 3장에서는 모의실험을 통하여 전체적인 최적 선택확률, 표본추출율과 부차표본추출율을 계산하여 구한다. 반복조사를 하는 경우에는 이미 조사 된 센서스의 자료를 이용하던가 또는 예비조사의 자료를 이용하여 표본설계를 한다고 가정한다.

2. 최적의 표본추출율(SAMPLING FRACTION)과 최적 선택확률

예비조사나 이미 조사된 센서스 자료를 이용하여 층화 이단계 집락표본추출(two-stage sampling)할 때에 최적의 표본 추출율과 최적 선택확률을 정하는 문제를 생각한다. 우리는 주어진 비용아래서 모총계 Y의 불편 추정량(unbiased estimate) $\hat{Y}_{ST}$의 분산 $V(\hat{Y}_{ST})$을 최소로 하는 n_h와 최적 선택확률 f_0, 표본 추출율 $\frac{n_h}{N_h}$와 부차 표본추출율 $\frac{m_{hi}}{M_{hi}}$를 구한다.

여기서,$\hat{Y}_{ST} = \sum_h \hat{Y}_h$이다. 그리고, y_h는 h층의 표본총계이고, Y_h는 h층의 층 총계이다. Y_{hi}는 h층의 i번째 집락의 총계이고 y_{hi}는 h층의 i번째 집락의 표본 총계이다. 그리고 $\hat{Y}_h$는 층의 모총계의 추정치이다. Y_h의 ppz추정량은

$$\hat{Y}_h = \frac{1}{n_h}\sum_i^{n_h} \frac{M_{hi}y_{hi}}{m_{hi}z_{hi}} = \frac{1}{n_h}\sum_i^{n_h} \frac{M_{hi}\bar{y}_{hi}}{z_{hi}} = \frac{1}{n_h}\sum_i^{n_h} \frac{\hat{Y}_{hi}}{z_{hi}}$$

이다. 여기서, Cochran(1997)에 의하면 ppz추정량이라 함은 집락을 에 확률비례하여 추출했을때에 추정량을 말한다.

따라서 $\hat{Y}_{ST} = \sum_h^L \frac{1}{n_h}\sum_i^{n_h} \frac{M_{hi}y_{hi}}{m_{hi}z_{hi}}$이다.

그리고 Cochran(1977)과 마찬가지로 z_hi는 h층의 i번째 단위가 추출될 확률로 $\sum_i z_{hi} = 1$이다.

이 논문에서는 일차단위(집락)를 복원으로 z_{hi}에 확률비례하여 추출하는데 특히 $z_{hi} = M_{hi}/M_{h0}$인 경우를 생각한다. 여기서, $M_{h0} = \sum_i M_{hi}$이다. $\hat{Y}_{ST}$를 전체적으로 자기-가중(self-weighting) 으로 만들기 위하여

$$m_{hi} = (f_0 M_{hi})/(n_h z_{hi}) = (f_0 M_{hi})/\pi_{hi} \tag{2.1}$$

이라고 가정한다. f_0는 ssu가 표본으로 추출될 확률로 (2.4)와 같고, π_{hi}는 h층의 i번째 단위가 표본으로 추출될 확률이다. 여기서, 비용함수는

$$C = \sum_h c_{uh} n_h + \sum_h (c_{2h} \sum_i^{n_h} m_{hi}) + \sum_h c_{lh} \sum_i^{n_h} M_{hi}$$

이다. 비용함수에 포함되는 항들은

$C_{uh} = h$층의 일차단위 당 고정비용, $C_{2h} = h$층의 부차단위 당 비용

$C_{lh} = h$층의 추출된 단위 내에서 부차 단위당 리스팅 비용

이다.

$$\begin{aligned} E(\sum_{i=1}^{n_h} m_{hi}) &= E(\sum_{i=1}^{n_h} f_0 M_{hi}/\pi_{hi}) = \sum_{i=1}^{N_h} \pi_{hi}(f_0 M_{hi}/\pi_{hi}) \\ &= \sum_{i=1}^{N_h} f_0 M_{hi} = f_0 M_{h0} \end{aligned}$$

이므로, n_h단위들의 표본추출의 평균비용은

$$E(C) = \sum_{h=1}^{L} c_{uh} n_h + \sum_{h=1}^{L} c_{2h} f_0 M_{hi} + \sum_{h=1}^{L} c_{lh} \sum_i^{N_h} \pi_{hi} M_{hi} \tag{2.2}$$

이다.

$\hat{Y}_{ST}$의 분산은 Cochran(1977)의 (11.53)에 의하여

$$\begin{aligned} V(\hat{Y}_{ST}) &= \sum_h^L V(\hat{Y}_h) = \sum_h^L \frac{1}{n_h} \sum_i^{N_h} \left[z_{hi} (\frac{Y_{hi}}{z_{hi}} - Y_h)^2 + \frac{M_{hi}(M_{hi} - m_{hi})}{z_{hi} m_{hi}} S_{2hi}^2 \right] \\ &= \sum_h^L \frac{1}{n_h} \sum_i^{N_h} \left[\frac{1}{z_{hi}} (Y_{hi} - z_{hi} Y_h)^2 + \frac{M_{hi}(M_{hi} - m_{hi})}{z_{hi} m_{hi}} S_{2hi}^2 \right] \end{aligned}$$

이다. $d_{hij} = y_{hij} - z_{hi}(\sum_i y_{hij})$로 놓으면 $(Y_{hi} - z_{hi} Y_h) = M_{hi} \bar{D}_{hi}$이다. 따라서, $\pi_{hi} = n_h z_{hi}$와 $M_{hi}/n_h z_{hi} m_{hi} = 1/f_0$에서

$$V(\hat{Y}_{ST}) = \sum_h^L \sum_i^{N_h} \left[\frac{M_{hi}^2}{\pi_{hi}} (\bar{D}_{hi}^2 - \frac{S_{2hi}^2}{M_{hi}}) + \frac{M_{hi}}{f_0} S_{2hi}^2 \right] \tag{2.3}$$

이다. 여기서

$$S_{2hi}^2 = \frac{1}{M_{hi} - 1} \sum_j^{M_{hi}} \left[(y_{hij} - \bar{Y}_{hi}) \right]^2$$

이다. 조건

$$z_{hi} n_h \left(\frac{m_{hi}}{M_{hi}} \right) = f_0 \tag{2.4}$$

는 h층의 i번째 집락내의 이차단위가 추출되는 확률이다.

특수한 경우로는 만약 주어진 비용 아래서, f_0가 미리 선택된다면 n_h는 식 (2.2)에서 구할 수 있다. 그리고, m_{hi}는 식 (2.4)에서 구할 수 있다. 일반적인 경우로 우리는 고정된 평균비용 (2.2)와

$$\sum_{i=1}^{N_h} z_{hi} = 1, \sum_i^{N_h} \pi_{hi} = n_h, h = 1, 2, \ldots, L$$

인 조건에서, V를 최소화 하는 n_h, f_0를 정하고자 한다. 그러면, Lagrangian 승수법에 의하여, λ와 μ_h를 Lagrangian 승수로 잡고

$$V+\lambda\left[\sum_{h}^{L} c_{uh}n_h + \sum_{h=1}^{L} c_{2h}f_0 M_{h0} + \sum_{h}^{L} c_{lh}\sum_{i=1}^{N_h}\pi_{hi}M_{hi} - E(C)\right] + \sum_{h=1}\mu_h(n_h - \sum_{i=1}^{N_h}\pi_{hi}) \quad (2.5)$$

를 최소로 한다. 특히, ppz 표본추출이 다음과 같은 경우를 생각하여

$$z_{hi} = \frac{M_{hi}}{M_{h0}}$$

이라 놓자. 여기서 $M_{h0} = \sum_{i=1}^{N_h} M_{hi}$이다. 식 (2.5)를 n_h와 π_{hi}에 관하여 미분하면 n_h : $\lambda c_{uh} + \mu_h = 0$. 즉 $\mu_h = -\lambda c_{uh}$

$$\pi_{hi} : -\frac{M_{hi}^2}{\pi_{hi}^2}(\bar{D}_{hi}^2 - \frac{S_{2hi}^2}{M_{hi}}) + \lambda c_{lh}M_{hi} - \mu_h = 0 \quad (2.6)$$

그리고 표본추출율 f_0에 관하여 (2.5)을 미분하면

$$\sum_i \frac{-M_{hi}}{f_0^2}S_{2hi}^2 + \lambda c_{2h}M_{h0} = 0$$

으로, 이 식을 에 관하여 풀면 최적선택확률은

$$f_0^2 = \frac{\sum_i M_{hi}S_{2hi}^2}{\lambda c_{2h}M_{h0}} \quad (2.7)$$

이다.

그러면, (2.6)에서 λ는

$$\lambda = \frac{\sum_h^L \sum_{i=1}^{N_h}(c_{lh}M_{hi} + c_{uh})}{\sum_h^L \sum_i^{N_h}[\frac{M_{hi}^2}{\pi_{hi}^2}(\bar{D}_{hi}^2 - \frac{S_{2hi}^2}{M_{hi}})]} \quad (2.8)$$

이다. 그러면 식 (2.7)에 $\pi_{hi} = n_h z_{hi}$를 대입하여 식 (2.2)와 연립하여 풀어서 n_h값을 구할 수 있다. 그리고 식 (2.7)에 앞에서 구한 n_h값을 대입하여 f_0값을 구할 수 있다. n_h값과 N_h값으로부터 h층의 표본 추출률은 $\frac{n_{hi}}{N_{hi}}$이다.

m_{hi}값은 (2.4)식에서 구할 수 있고, 부차표본 추출율은 $\frac{m_{hi}}{M_{hi}}$이다.

3. 모의실험

우리는 다음과 같은 표 3.1의 자료로 모의 실험을 통하여 주어진 비용아래서 $\hat{Y}_{ST}$의 분산 $V(\hat{Y}_{ST})$을 최소로하는 n_h와 m_{hi}를 구한다. 일반성을 잃지말고 계산상의 편의를 위하여, 모집단이 2개의 층으로 이루어 졌다고 가정한다. 우리는 $N_h = 30, h = 1, 2$인 집락들로 이루어진 모집단에서 층화 이단계 표본추출을 하고자 한다. 먼저 비용은 모든 h에 대하여 $c_{uh} = 5, c_{2h} = 1$, 리스팅 비용 $c_{lh}M_{hi}$는 무시할 수 있다고 가정한다. 즉, $z_{hi} \propto M_{hi}$라고 가

정한다. 여기서 $\pi_{hi} = n_h z_{hi}$이다. 집락의 크기는 $M_{hi} = 60$, $i = 1, 2, \ldots, 30$으로 잡았는데, y_{hij}는 다음과 같이 이진 자료이다.

$$y_{hij} = \begin{cases} 1 & employed\ men\ of\ over\ 16\ years\ old \\ 0 & men\ of\ over\ 16\ years\ old \end{cases}$$

표 3.1: 16세 이상의 남자인구 중 취업한 남자의 인구

집락	16세 이상의 취업한남자		$\bar{D}_{hi}$		S^2_{2hi}	
(i)	층1(Y_{1i})	층2(Y_{2i})	층1	층2	층1	층2
1	17	14	-0.0222	-0.0722	0.2065	0.1819
2	18	8	-0.0055	-0.1722	0.2136	0.1175
3	12	9	-0.1055	-0.1555	0.1627	0.1297
4	15	11	-0.0555	-0.1222	0.1907	0.1523
5	10	13	-0.1389	-0.0889	0.1412	0.1726
6	14	14	-0.0722	-0.0722	0.1819	0.1819
7	11	13	-0.1222	-0.0889	0.1523	0.1726
8	13	20	-0.0889	0.0278	0.1726	0.2260
9	15	18	-0.0555	-0.0055	0.1907	0.2136
10	19	18	0.0111	-0.0055	0.2201	0.2136
11	16	16	-0.0389	-0.0389	0.1989	0.1989
12	19	15	0.0111	-0.0555	0.2201	0.1907
13	19	18	0.0111	-0.0055	0.2201	0.2136
14	14	19	-0.0722	0.0111	0.1819	0.2201
15	23	20	0.0778	0.0278	0.2404	0.2260
16	24	14	0.0945	-0.0722	0.2441	0.1819
17	23	19	0.0778	0.0111	0.2404	0.2201
18	24	22	0.0945	0.0611	0.2441	0.2362
19	27	12	0.1445	-0.1055	0.2517	0.1627
20	25	17	0.1111	-0.0222	0.2472	0.2065
21	23	20	0.0778	0.0278	0.2404	0.2260
22	20	21	0.0278	0.0445	0.2260	0.2314
23	20	24	0.0278	0.0945	0.2260	0.2441
24	14	21	-0.0722	0.0445	0.1819	0.2314
25	20	21	0.0278	0.0445	0.2260	0.2314
26	23	21	0.0778	0.0445	0.2404	0.2314
27	21	22	0.0445	0.0611	0.2314	0.2362
28	21	21	0.0445	0.0445	0.2314	0.2314
29	19	15	0.0111	-0.0555	0.2201	0.1907
30	11	16	-0.1222	-0.0389	0.1523	0.1989
total	550	512				

표 3.2: 고정된 비용(200)에 대하여 n_1과 n_2가 변할 때 $\sqrt{V(\hat{Y}_{ST})}$

$n_1 \setminus n_2$	15	16	17	18	19	20	21
12							11.319
13						11.254	
14					11.213		
15				11.194			
16			11.196				
17		11.219					
18	11.262						

표 3.3: 고정된 비용(250)에 대하여 n_1과 n_2가 변할 때 $\sqrt{V(\hat{Y}_{ST})}$

$n_1 \setminus n_2$	19	20	21	22	23	24	25
16							10.016
17						9.983	
18					9.964		
19				9.956			
20			9.960				
21		9.976					
22	10.004						

표 3.4: 고정된 비용(300)에 대하여 n_1과 n_2가 변할 때 $\sqrt{V(\hat{Y}_{ST})}$

$n_1 \setminus n_2$	23	24	25	26	27	28	29
20							9.088
21						9.070	
22					9.059		
23				9.056			
24			9.061				
25		9.073					
26	9.093						

표 3.2, 표 3.3, 표 3.4에서는 n_1과 n_2에 변화에 따른 $V(\hat{Y}_{ST})$의 값을 구하여, $V(\hat{Y}_{ST})$가 최소가 되는 n_1과 n_2를 구하였다. 다음 표 3.5에 최적의 선택확률 f_0를 구하였다. 표본 추출율은 $\frac{n_1}{N_1} = \frac{15}{30}$, $\frac{n_2}{N_2} = \frac{18}{30}$이다. 부차표본 추출률은 각 집락 i에 대하여 $\frac{m_{hi}}{60}$이다.
표 3.2, 표 3.3, 표 3.4를 시각적으로 보기에 용이하도록 그림으로 표시하여 $\sqrt{V(\hat{Y}_{ST})}$가 최소가 되는 n_1과 n_2를 나타내었다.

표 3.5: 고정된 비용에 따른 n_h와 표준편차

$E(C)$	f_0	n_1	n_2	$\sqrt{V(\hat{Y}_{ST})}$
200	0.292	15	18	11.194
250	0.375	19	22	9.956
300	0.458	23	26	9.056

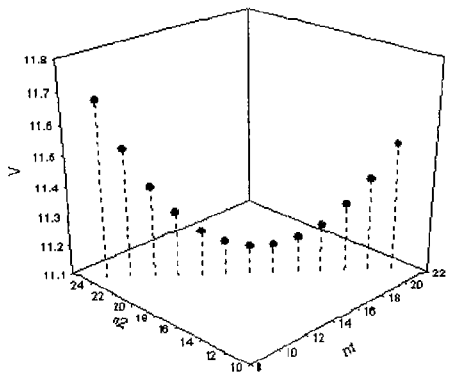

그림 3.1: 고정된 비용(200)에 대하여 n_1과 n_2가 변할 때 $\sqrt{V(\hat{Y}_{ST})}$

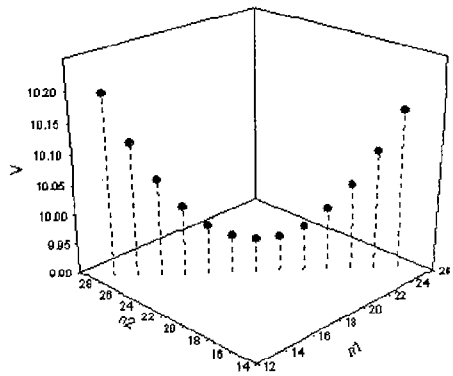

그림 3.2: 고정된 비용(250)에 대하여 n_1과 n_2가 변할 때 $\sqrt{V(\hat{Y}_{ST})}$

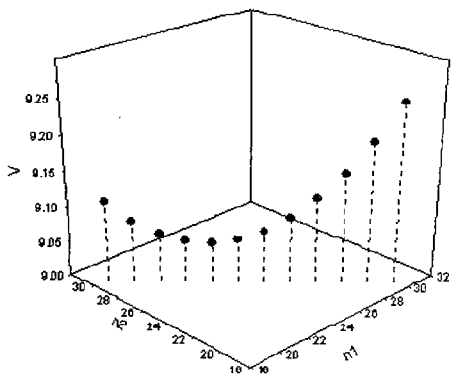

그림 3.3: 고정된 비용(300)에 대하여 n_1과 n_2가 변할 때 $\sqrt{V(\hat{Y}_{ST})}$

4. 결 론

표본조사 비용이 미리 정해졌을 때에, 모총계 추정량의 분산을 최소로 하는 층화 이단계 표본 추출을 하는 문제를 다루었다. 즉, 분산을 최소로 하는 최적의 선택 확률을 구하여 표본 추출율을 정하였다. 우리는 실용성이 높은 방법으로 각 집락의 추출은 ppz로 추출 하였다. 앞으로 연구할 과제는 실제 표본 설계시에 많이 나타나는 문제로 몇 개의 집락에 대하여는 미리 일차단위의 크기와 이차단위의 크기가 정해졌을 때 추정량의 분산을 추정하여 층화 이단계 표본 추출을 하는 것이다.

참고문헌

[1] Cochran (1977). *Sampling Technique*, John Willy and Sons.

[2] Hansen, M.H. and Hurwitz, W.N. (1949). On the determination of the optimum probabilities in sampling. *Ann. Math.*, **20**, 426-432.

[3] Lohr. S. (1999). *Sampling : Design and Analysis.* Duxubury press.

[4] Scheaffer, R.L. Menderhall, w. and ott. L (1990), *Elementary survey sampling.* Duxbury Press.

[5] Thompsom (1992). *Sampling.* John Wiley and Sons, Inc.

[2001년 4월 접수, 2001년 8월 채택]

Optimum Selection Probabilites in Stratified Two-stage Sampling

Min Woong Shin[1] Sang Hoon Oh[2]

ABSTRACT

We determine the optimum selection probabilties, the optimum sampling fractions and the subsampling fractions in stratified two-stage sampling.

Keywords: Optimum selection probability; Sampling fraction; Subsampling fraction.

1) Proffesor, Department of Informetrics and Statistics, Hankuk University of Foreign Studies. E-mail: mwshin@stat.hufs.ac.kr

2) Graduate Student, Department of Informetrics and Statistics, Hankuk University of Foreign Studies. E-mail: shoh@stat.hufs.ac.kr