응용통계연구 제 **13**권 2호, 2000년, pp. 207 224

층화 2-단 표본 추출시 최적 집락의 크기 결정

신민웅 [1] 신기일 [2]

요 약

모집단을 집락화하여 층화 2-단 표본 추출을 할 때에 일반적으로 집락의 크기는 정해져 있다. 그러나 집락이 아파트 단지 등과 같을 경우에 집락의 크기는 큰 차이를 보인다. 이 경우 집락을 합치거나 또는 분할할 필요가 생긴다. 대 표본조사(large sample survey)에서 행정상 또는 조사 편의상 동질의 원소들이 집락화 되어 있고 집락의 크기를 결정할 필요가 있을 경우가 고려되었으며 본 논문에서는 집락의 최적 크기를 결정하는 문제를 다루었다. 또한 주어진 비용 하에서 최적의 일차 추출 단위 수와 최적의 이차 추출 단위 수를 구하였다.

주요용어: 층화 2-단 추출법, 결합비 추정, 분리 비 추정, 최적 집락 크기.

1. 서론

전국적 규모의 표본 설계를 할 때 층화 2-단 표본 추출법이 주로 사용되고 있다. 예를 들어 주택은행에서는 '전국도시 주택가격 동향조사'를 실시하는데 층화 2-단 표본 추출법을 사용하고 있다. 즉 전국의 도시를 하나의 층으로, 층내의 동 또는 아파트단지를 하나의 집락으로 생각하고 집락안의 가구를 추출하는 방법을 사용하고 있다.

따라서, 많은 사회조사에서처럼 행정상으로나 조사의 편의를 위하여 원소들을 동질적으로 층화한 후 인근의 조사 단위들을 같은 집락으로 묶는다. 특히, 아파트 가격 조사시에는 같은 평형별로 층화한 후에 집락화를 하므로 동질의 원소들이 같은 집락에 묶인다. 즉, 일반적으로 이질적인 원소가 집락화되어 추정의 효율을 높이는 문제와는 다르다.

모집단이 L 개의 층으로 층화되었을 때에 h 층의 N_h개의 일차단위(집락)로부터 n_h개의 일차 단위를 추출한다. 그리고 h층 i번째 집락이 (크기가 M_{hi}) 추출되었을 때에 이 집락에서 m_{hi}개의 이차단위(부차단위)를 추출한다. 본 논문은 이와 같이 층화 2-단 표본 추출을 할 때에 주어진 비용아래서 모총계 Y의 비 추정량의 분산을 최소로 하는 최적 집락의 크기 M_{hi} , 최적 일차 표본 수 n_h, 그리고 최적 부차 표본 수 m_{hi}을 구하는 문제를 다룬다.

Scheaffer(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 집락에 대해 확률 표본을 추출하고, 추출된 각 집단내에서 조사 단위들을 이차로 추출하는 것을 2-단 표본 추출법이라고 하

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

였다. 그리고 집락 표본 추출에서는 적절한 집락을 가장 먼저 정하여야 하고, 적절한 집락의 크기를 결정하는데 바람직한 규칙이 항상 있는 것이 아니고 각각의 문제에 따라 연구되어야 한다고 하였다. Efron(1986)은 집락 표본 추출에서 집락의 크기가 적절하지 않으면 과대 산포의 원인이 될 수 있음을 지적하였고 Ten Have와 Chinchilli (1998)는 랜덤 집락 크기를 갖는 집락화된 이항반응 자료에 대한 포아송 모형에 대하여 논의하였다. 이와 같이 최근에는 집락의 크기에 대한 많은 연구가 진행되고 있다.

Hansen 과 Hurwitz(1949)는 단순 2-단 표본 추출시에 표본 추출률과 부차표본 추출률을 구하는 문제를 다루었다. 또한 보조 자료를 이용한 비 추정에 있어서 추정량의 분산에 대해서는 Rao(1988)와 Royall(1988)에 의하여 논의되었다. Rao와 Rao(1971)는 비 추정량에 대하여 작은 표본의 모의실험 결과를 논의하였으며 신, 최와 이(1997)는 모총계의 결합 비 추정량의 분산을 최소로 하는 최적의 선택률을 구하였다.

2절에서는 주어진 비용아래서 최적 집락의 크기, 최적 표본 추출율 그리고 최적 부차 표본 추출율을 구하는 과정을 설명한다. 이때 추정량은 비 추정량을 사용하였으며 일반적으로 비 추정량은 결합 비 추정량과 분리 비 추정량으로 나누어지므로 본 논문에서는 이들 두 추정량 모두를 분석하였다. 3절에서는 서울 지역의 아파트 단지에서 아파트 가격을 조사한 후 이를 이용하여 예상되는 모의 자료를 만들어내고 이를 이용하여 최적 집락 크기를 결정하는 방법을 제시하였다. 끝으로 결론은 4절에 나와있다.

2. 최적 집락 크기의 결정

이미 조사된 센서스 자료를 이용하여 층화 2-단 집락 표본추출(two-stage sampling)을 할 때에 주어진 비용 아래서 모총계 Y의 분산을 최소로 하는 최적의 일차 추출 단위 수, 최적 이차 추출 단위 수 그리고 집락의 크기를 결정하는 문제를 생각한다. 주택은행의 '전국 도시 주택가격 동향조사'와 같이 최근에는 반복적인 조사가 이루어지는 경우가 많이 있다. 따라서 본 논문에서는 모총계 Y의 추정을 위하여 비 추정을 이용하였다. 비 추정은 분리 비 추정과 결합 비 추정으로 나누어지므로 이 절에서는 주어진 비용 하에서 모총계 Y의 두 추정량의 분산을 이용하여 이를 최소로 하는 최적의 일차 추출 단위 수, 최적의 이차 추출 단위 수 그리고 집락의 크기를 결정하였다. 참고로 본 논문에서는 Cochran (1977)에서 사용한 기호를 사용하였다.

2.1. 분리 비 추정량을 사용하였을 경우

주어진 비용아래서 모총계 Y의 분리 비 추정량(separate ratio estimate) $\hat{Y}_{RS}$의 분산 $V(\hat{Y}_{RS})$을 최소로 하는 n_h와 m_{hi}, 그리고 집락의 크기인 M_{hi}를 구한다. 분리 비 추정량은 다음과 같다.

$$\hat{Y}_{RS} = \sum_{h=1}^{L} \left(\frac{y_h}{x_h} \right) X_h$$

여기서, y_h, x_h는 h 층의 표본총계이고, X_h는 h층의 총계이다.

먼저 h층의 i번째 단위에 할당된 확률을 z_{hi}라 하자. 여기서 $\sum_{i=1}^{N_h} z_{hi} = 1$이다. 이 논문에서는 일차단위(집락)를 복원으로, 측도 z_{hi}인 확률로 추출하는데 특히 $z_{hi} = M_{hi}/M_{h0}$ (pps)인 경우를 생각한다. 여기서 $M_{h0} = \sum_{i=1}^{N_h} M_{hi}$이다. f_{0hi}를 h층의 i번째 집락내의 이차단위가 추출될 확률이라 하면 $f_{0hi} = z_{hi} n_h m_{hi}/M_{hi}$ 가 된다. 따라서 $\pi_{hi} = n_h z_{hi}$라 하면

$$m_{hi} = (f_{0hi} M_{hi})/(n_h z_{hi}) = (f_{0hi} M_{hi})/\pi_{hi} \tag{2.1}$$

이 된다.

또한 주어진 비용함수는 다음과 같으며

$$C = \sum_{h=1}^{L} c_{uh} n_h + \sum_{h=1}^{L} c_{2h} \sum_{i=1}^{n_h} m_{hi} + \sum_{h=1}^{L} c_{lh} \sum_{i=1}^{n_h} M_{hi}$$

비용함수에 포함되는 항들은

$c_{uh} = h$ 층의 일차단위 당고정비용

$c_{2h} = h$층의 부차단위 당 비용

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

이다. M_{hi} 가 알려져 있을 때 주어진 h층에서

$$E(\sum_{i=1}^{n_h} m_{hi}) = E(\sum_{i=1}^{n_h} f_{0hi} M_{hi}/\pi_{hi}) = \sum_{i=1}^{N_h} \pi_{hi} (f_{0hi} M_{hi}/\pi_{hi}) = \sum_{i=1}^{N_h} f_{0hi} M_{hi}$$

이므로, n_h 단위들의 표본추출 평균비용은

$$E(C) = \sum_{h=1}^{L} c_{uh} n_h + \sum_{h=1}^{L} c_{2h} \sum_{i=1}^{N_h} f_{0hi} M_{hi} + \sum_{h}^{L} c_{lh} \sum_{i=1}^{N_h} \pi_{hi} M_{hi} \tag{2.2}$$

이다. Cochran(1977)은 층이 하나일 경우 $\hat{Y}_{RS}$의 분산을 구했으며 L개의 층이 있는 경우는 쉽게 다음과 같이 됨을 알 수 있다.

$$V(\hat{Y}_{RS}) = \sum_{h=1}^{L} \frac{1}{n_h} \sum^{N_h} \left[\frac{1}{z_{hi}} (Y_{hi} - R_h X_{hi})^2 + \frac{M_{hi}(M_{hi} - m_{hi})}{z_{hi} m_{hi}} S_{d2hi}^2 \right]$$

여기서 $S_{d2hi}^2 = \frac{1}{M_{hi}-1} \sum_{j=1}^{M_{hi}} [(y_{hij} - R_h x_{hij}) - (\bar{Y}_{hi} - R_h \bar{X}_{hi})]^2$, $R_h = \sum_{i=1}^{N_h} Y_{hi} / \sum_{i=1}^{N_h} X_{hi}$ 이고, Y_{hi}, X_{hi}는 h층 i번째 집락의 총계이다. $Y_{hi} - R_h X_{hi} = D_{hi}$라하고 (2.1)식의 π_{hi}, f_{0hi}를 이용하면 분산 식은 다음과 같이 표현된다.

$$V(\hat{Y}_{RS}) = \sum_{h}^{L} \sum_{i}^{N_h} \left[\frac{{D_{hi}}^2}{\pi_{hi}} + \frac{1}{f_{0hi}} - \frac{1}{\pi_{hi}} M_{hi} S_{d2hi}^2 \right] \tag{2.3}$$

고정된 평균 비용 (2.2)와

$$\frac{n_1}{N_1} = \cdots = \frac{n_L}{N_L}, \ \sum_{i=1}^{N_h} \pi_{hi} = n_h, \ h = 1, 2, \ldots, L \tag{2.4}$$

인 조건에서, $V(\hat{Y}_{RS})$를 최소화하는 최적의 n_h, f_{0hi}, π_{hi}, 그리고 M_{hi}를 정한다. 이는 Lagrangian 승수법에 의하여, λ와 μ_h를 Lagrangian 승수로 잡고

$$V(\hat{Y}_{RS}) \quad +\lambda[\sum_{h=1}^{L} c_{uh}n_h + \sum_{h=1}^{L} c_{2h} \sum_{i=1}^{N_h} f_{0hi}M_{hi} + \sum_{h=1}^{L} c_{lh} \sum_{i=1}^{N_h} \pi_{hi}M_{hi} - E(C)]$$

$$+ \sum_{h=1}^{L} \mu_h(n_h - \sum_{i=1}^{N_h} \pi_{hi}) \tag{2.5}$$

를 최소로 하는 n_h, f_{0hi}, π_{hi} 그리고 M_{hi}를 정하는 것과 같게 되며 결과는 다음과 같다.

$$M_{hi}^{opt} = [D_{hi}^2 K_{hi} - c_{uh}][c_{lh} + K_{hi}S_{d2hi}^2]^{-1} \tag{2.6}$$

이에 관한 유도과정은 부록을 참고하기 바란다. 이제 $\sum_{i=1}^{N_h} M_{hi}^{opt} = M_{h0}$가 만족되어야 하므로 최종적인 최적의 M_{hi}^{opt}를 M_{hi}^{f}라 하면

$$M_{hi}^{f} = \frac{M_{h0}}{\sum_{i=1}^{N_h} M_{hi}^{opt}} M_{hi}^{opt} \tag{2.7}$$

가 된다. 또한 이를 이용하면 다음의 결과를 얻는다.

$$n^f = \frac{N \cdot E(C)}{\sum_{h=1}^{L} c_{uh}N_h + \sum_{h=1}^{L} c_{2h}N_h \sum_{i=1}^{N_h} \sqrt{K_{hi}} \sqrt{\frac{S_{d2hi}^2}{c_{2h}}} \frac{(M_{hi}^f)^2}{M_{h0}} + \sum_{h}^{L} c_{lh}N_h \sum_{i=1}^{N_h} \frac{(M_{hi}^f)^2}{M_{h0}}}$$

$$n_h^f = \frac{N_h \cdot E(C)}{\sum_{h=1}^{L} c_{uh}N_h + \sum_{h=1}^{L} c_{2h}N_h \sum_{i=1}^{N_h} \sqrt{K_{hi}} \sqrt{\frac{S_{d2hi}^2}{c_{2h}}} \frac{(M_{hi}^f)^2}{M_{h0}} + \sum_{h}^{L} c_{lh}N_h \sum_{i=1}^{N_h} \frac{(M_{hi}^f)^2}{M_{h0}}}$$

$$m_{hi}^f = \sqrt{K_{hi} \cdot \frac{S_{d2hi}^2}{c_{2h}}} M_{hi}^f$$

여기서 $K_{hi} = \frac{2c_{lh}+c_2^* - \sqrt{(2c_{lh}+c_2^*)^2 - 4c_{lh}^2}}{2S_{d2hi}^2}$ 이다. 이에 관한 유도과정은 부록에 나와있다. 실제 표본 설계에서는 c_{uh}와 c_{lh}가 결과에 영향을 미치지 않을 정도로 작은 경우가 많이 있으며 이 경우에는 $M_{hi}^{opt} \approx [D_{hi}^2][S_{d2hi}^2]^{-1}$가 되어

$$M_{hi}^{f} \approx \frac{M_{h0}}{\sum_{i=1}^{N_h}[D_{hi}^2 S_{d2hi}^{2\ -1}]} \frac{D_{hi}^2}{S_{d2hi}^2} \tag{2.8}$$

을 얻는다. 따라서 h층 i번째 집락에서 D_{hi}^2가 크면 M_{hi}^f를 크게 잡고 큰 S_{d2hi}^2에 대해서는 작은 M_{hi}^f를 잡으면 $\hat{Y}_{RS}$ 의 분산을 줄일 수 있다.

2.2. 결합 비 추정량을 사용하였을 경우

결합 비 추정치(combined ratio estimate)에 대해서도 $\hat{Y}_{RC}$의 분산 $V(\hat{Y}_{RC})$ 을 최소로 하는 n_h, m_{hi}와 M_{hi}를 구한다. 여기서, $\hat{Y}_{RC} = X(\hat{Y}_{st}/\hat{X}_{st})$이고 $\hat{Y}_{st} = \sum_{h=1}^{L} \hat{Y}_h$이며 $\hat{Y}_h$는 h층의

모총계 Y_h의 추정량이다. 또한 $\hat{X}_{st} = \sum_{h=1}^{L} \hat{X}_h$ 그리고 $\hat{X}_h$는 h층의 모총계 X_h의 추정량이다. 이때 $\hat{Y}_{RC}$의 분산은

$$V(\hat{Y}_{RC}) = \sum_{h}^{L} \frac{1}{n_h} \sum_{h}^{N} [z_{hi} \left(\frac{D_{hi}}{z_{hi}} - D_h \right)^2 + \frac{M_{hi}^2 (1 - f_{2hi}) S_{d2hi}^2}{z_{hi} m_{hi}}. \tag{2.9}$$

이다. 여기서 $S_{d2hi}^2 = \frac{1}{M_{hi}-1} \sum_{j}^{M_{hi}} [(y_{hij} - Rx_{hij}) - (\bar{Y}_{hi} - R\bar{X}_{hi})]^2$, $D_{hi} = Y_{hi} - RX_{hi}$ 그리고 $D_h = Y_h - RX_h$이다. 이제 $\pi_{hi} = n_h z_{hi}$, $M_{hi}/n_h z_{hi} m_{hi} = 1/f_{2hi}$라 하고 pps 추출을 고려하여 $z_{hi} = M_{hi}/M_{h0}$라 놓으면 (2.9)은

$$V(\hat{Y}_{RC}) = \sum_{h}^{L} \sum_{i}^{N_h} \left[\frac{D_{hi}^2}{\pi_{hi}} - 2 \frac{M_{hi}}{M_{h0} \pi_{hi}} D_{hi} D_h + \frac{M_{hi}^2}{M_{h0}^2 \pi_{hi}} D_h^2 - \frac{M_{hi}}{\pi_{hi}} S_{d2hi}^2 + \frac{M_{hi}}{f_{2hi}} S_{d2hi}^2 \right] \tag{2.10}$$

이 된다. 전 절에서와 같이 우리의 목적은 고정된 평균비용 (2.2)와 (2.4)의 조건에서, $V(\hat{Y}_{RC})$를 최소화하는 n_h, m_{hi} 그리고, M_{hi}를 정하는 것이다. 최적의 집락 크기를 M_{hi}^{opt}라 하면 부록에 의하여 다음의 결과를 얻는다.

$$\begin{aligned} &\left(- \frac{2D_{hi} D_h}{M_{h0}} + \frac{2M_{hi}^{opt} D_h^2}{M_{h0}^2} - S_{d2hi}^2 \right)^2 G_{hi}^2 \\ &+ \left[2 \left(- \frac{2D_{hi} D_h}{M_{h0}} + \frac{2M_{hi}^{opt} D_h^2}{M_{h0}^2} - S_{d2hi}^2 \right) c_{lh} - \left(\sqrt{\frac{c_{2h}}{S_{d2hi}^2}} + \frac{\sqrt{S_{d2hi}^2}}{\sqrt{c_{2h}}} \right)^2 \right] G_{hi} + c_{lh}^2 = 0 \end{aligned} \tag{2.11}$$

여기서 $G_{hi} = \left[c_{lh} M_{hi}^{opt} + c_{uh} \right] \left[D_{hi}^2 - \frac{2M_{hi}^{opt}}{M_{h0}} D_{hi} D_h + \frac{M_{hi}^{opt\,2}}{M_{h0}^2} D_h^2 - M_{hi}^{opt} S_{d2hi}^2 \right]^{-1}$ 이다. 이 식은 M_{hi}^{opt}의 4차식이 되어 최적 집락의 크기M_{hi}^{opt} 가 구해진다. 구해진 M_{hi}^{opt}와 (2.7)식의 방법을 이용하면 M_{hi}^{f}를 구할 수 있다. 또한

$$n^f = \frac{N \cdot E(C)}{\sum_{h=1}^{L} c_{uh} N_h + \sum_{h=1}^{L} c_{2h} N_h \sum_{i=1}^{N_h} \sqrt{G_{hi}} \sqrt{\frac{S_{d2hi}^2}{c_{2h}}} \frac{(M_{hi}^f)^2}{M_{h0}} + \sum_{h}^{L} c_{lh} N_h \sum_{i=1}^{N_h} \frac{(M_{hi}^f)^2}{M_{h0}}}$$

$$n_h^f = \frac{N_h \cdot E(C)}{\sum_{h=1}^{L} c_{uh} N_h + \sum_{h=1}^{L} c_{2h} N_h \sum_{i=1}^{N_h} \sqrt{G_{hi}} \sqrt{\frac{S_{d2hi}^2}{c_{2h}}} \frac{(M_{hi}^f)^2}{M_{h0}} + \sum_{h}^{L} c_{lh} N_h \sum_{i=1}^{N_h} \frac{(M_{hi}^f)^2}{M_{h0}}}$$

$$m_{hi}^f = \sqrt{G_{hi} \cdot \frac{S_{d2hi}^2}{c_{2h}}} M_{hi}^f$$

이 된다.

3. 사례분석

2절에서 분리 비 추정량을 사용하였을 경우 집락의 최적 크기 M_{hi}^f는 (2.7)식으로 주어졌고 결합 비 추정량을 사용하였을 경우는 4차 식으로 주어졌다. 이 결과는 M_{hi}가 충분히 큰 경우 D_{hi}^2, S_{d2hi}^2가 일정하다는 가정 하에서 얻어진 결과이다. 사례분석을 위하여 서울의 50개 아파트 단지를 선정하고 이 아파트 단지에서 1996년과 1998년의 아파트 평수, 상한가, 하한가를 구하였다. 표 3.1이 모의 실험에 사용된 50개 아파트 단지 목록이다.

표 3.1: 모의 실험에 사용된 아파트 단지

구	동	아파트 이름	동	아파트 이름
강남구	개포동	시영	압구정동	신현대
강동구	명일동	명일 LG	천호동	유원
강북구	번동	주공1단지	우이동	성원
강서구	방화동	개화	염창동	현대
관악구	신림10동	동마	봉천1동	보라매삼성
광진구	구의동	우성	노유동	한강성원
구로구	신도림동	미성	고척동	서울가든
노원구	공릉동	현대	상계동	금호
중랑구	신내동	보광	면목동	두산
중구	신당동	현대		
종로구	명륜동	명륜아남	창신동	쌍용
은평구	녹번동	대림	불광동	미성
	신사동	뉴신성		
용산구	보광동	신동아	이태원동	청화
영등포구	당산동	삼익	여의도동	공작
양천구	목동	벽산	신월동	길훈
송파구	가락동	프라자	송파동	성원
성북구	길음동	삼부	안암동	대광
성동구	금호동	두산	응봉동	현대
서초구	반포동	미주	방배동	신동아
서대문구	연희동	대림	홍은동	극동
마포구	염리동	진주	성산동	대우시영
동작구	본동	신동아	흑석동	한강현대
동대문구	용두동	신동아	휘경동	서울가든
도봉구	창동	동아	도봉동	럭키
금천구	시흥동	한양	독산동	삼승

본 모의 실험에서는 먼저 위에서 언급한 가정이 타당한지 아니면 어떤 선형관계가 있는지 살펴보고 실제 응용에서 사용될 수 있는 방법에 관하여 살펴보았다. 본 논문에서는 모의 실험을 간단하게 하기 위하여 M_{hi}^f 식이 나와있는 분리 비 추정량만을 고려하였다. 또한 M_{hi}^f의 공식을 살펴보면 각 층의 M_{hi}^f는 다른 층의 M_{hi}^f와 무관하므로 모의 실험은 하나의 층만을 고려하였다. 각 가구의 실제 가격은 나와있지 않기 때문에 상한가와 하한가를 이용하여 세대수에 맞게 모의 자료를 만들었다. 이때 아파트 가격은 균등분포를 따른다고 가정하였다. 먼저 M_{hi}가 충분히 큰 경우 S_{d2hi}^2과 D_{hi}^2가 일정하다는 가정이 타당한지를 알아보자. 이를 위하여 각 아파트 단지를 하나의 집락이라 가정하고 S_{d2hi}^2, D_{hi}^2을 구하였다. 여기서 $S_{d2hi}^2 = \frac{1}{M_{hi}-1}\sum_{j=1}^{M_{hi}}[(y_{hij} - R_h x_{hij}) - (\bar{Y}_{hi} - R_h \bar{X}_{hi})]^2$, $R_h = \sum_{i=1}^{N_h} Y_{hi} / \sum_{i=1}^{N_h} X_{hi}$ 이고 $Y_{hi} - R_h X_{hi} = D_{hi}$ 이다. 또한 Y_{hi}, X_{hi}는 h층 i번째 집락의 총계이다. 그림 3.1과 그림 3.2는

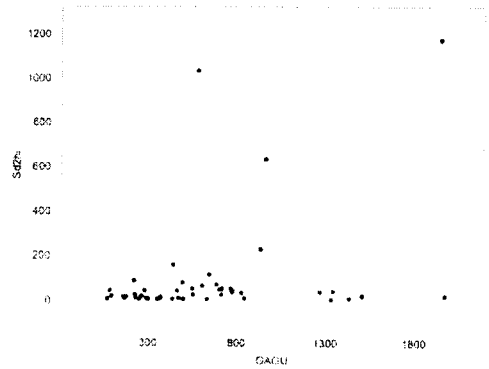

그림 3.1: 가구수에 대한 S^2_{d2hi}

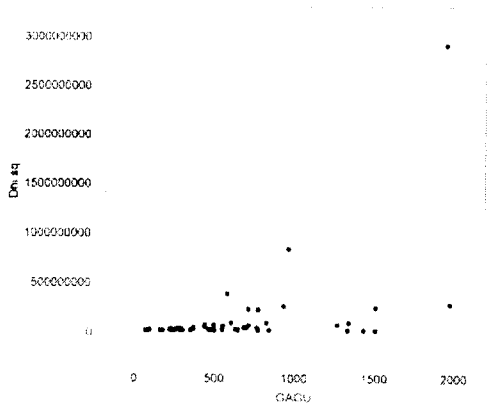

그림 3.2: 가구수와 D^2_{hi}

M_{hi}와 S^2_{d2hi} 그리고 M_{hi}와 D^2_{hi}이 어떤 관계가 있는지를 알아보기 위하여 이들의 산점도를 그린 것이다.

두 그림에서 알 수 있듯이 S^2_{d2hi} 그리고 D^2_{hi}는 가구 수 M_{hi}가 커질 경우에도 일정한 것으로 판단할 수 있다. 이의 관계를 보기 위하여 각각 S^2_{d2hi} 와 D^2_{hi} 를 종속변수로 하고 독립변수는 M_{hi}로 하는 두 개의 회귀식을 추정하였다. 추정 결과 p-값이 각각 0.67 과 0.17로 기울기가 모두 유의하지 않은 것으로 나타났다. 이는 2절에서 M_{hi}가 충분히 큰 경우 S^2_{d2hi} 그리고 D^2_{hi}에 영향을 주지 않는다고 가정하는 것에 큰 무리가 없다는 것을 보여준다. 이제 2절에서 얻어진 결과를 이용하여 최적의 M^f_{hi}를 갖는 방법에 대하여 살펴보자. 실제 자료분석에서 최적의 M^f_{hi}를 구하는 것은 쉽지가 않다. 그러나 c_{uh} 와 c_{lh} 가 결과에 영향을 미치지 않을 정도로 작다고 가정한다면 (2.8)식에서 얻어진 결론을 사용할 수 있다. 즉 D^2_{hi} 이 크면 큰 M^f_{hi}를 선택하고 S^2_{d2hi}이 크면 작은 M^f_{hi}를 선택하는 것이다. 이를 적용하면 최적의 M^f_{hi}는

근사적으로 D_{hi}^2/S_{d2hi}^2에 비례하게 된다. 따라서 주어진 가구 수와 D_{hi}^2/S_{d2hi}^2의 산점도를 그려 이 산점도를 되도록 이면 직선에 가깝게 만드는 것이 좋을 것이다. 문제를 단순화시키기 위하여 주어진 50개 단지에서 10개의 단지를 임의로 추출하였다. 다음의 표 3.2는 선정된 10개 단지이다.

표 3.2: 선택된 10개의 아파트 단지와 단지내의 평수

구	동	아파트 이름	평 수
강남구	개포동	시영	10, 13, 17, 19
강동구	명일동	명일 LG	24, 25, 33, 34, 35
강북구	번동	주공1단지	17, 18, 21, 26, 30, 31
구로구	신도림동	미성	15, 21, 27, 34
중구	신당동	현대	27, 32, 33, 43, 49
종로구	창신동	쌍용	23, 26, 33, 39, 42
은평구	불광동	미성	28, 35, 47
동작구	본동	신동아	19, 25, 35, 40
동작구	흑석동	한강현대	28, 32, 43, 48
금천구	시흥동	한양	16, 25, 35

M_{hi}^f가 D_{hi}^2/S_{d2hi}^2에 비례하도록 만들기 위하여 먼저 각 아파트 단지의 평수를 하나의 집락으로 생각하고 S_{d2hi}^2과 D_{hi}^2를 각각 구한다. 구해진 S_{d2hi}^2과 D_{hi}^2를 서로 비교하여 만약 작은 S_{d2hi}^2값을 가지면 옆의 평수와 합쳐서 하나의 집락을 만들고 또한 D_{hi}^2이 크면 마찬가지로 옆의 평수와 합치는 방법을 취한다.

선택된 10 개의 아파트 단지에서 구한 S_{d2hi}^2과 D_{hi}^2를 표로 만들면 다음과 같다. 표 3.3에서도 알 수 있듯이 D_{hi}^2/S_{d2hi}^2는 큰 차이를 보이고 있으며 주어진 가구 수와 비례하지는 않는다.

표 3.3: 10 개 아파트 단지의 S_{d2hi}^2 과 D_{hi}^2

아파트 이름	S_{d2hi}^2	D_{hi}^2	가구수	D_{hi}^2/S_{d2hi}^2
강남구 시영	15.723	251239929	1970	15979275
강동구 명일 LG	40.930	216082726	772	5279304
강북구 주공 1단지	4.107	3450850	1430	840297
구로구 미성	28.462	71271829	824	2504068
중구 현대	224.466	81042880	932	1113960
종로구 쌍용	15.242	155465	1504	10199
은평구 미성	36.537	81042820	1340	2218111
동작구 신동아	45.334	22457015	765	495366
동작구 한강현대	625.325	801871205	960	1282327
금천구 한양	14.897	242802177	1505	16298248

표 3.4: 10 개 아파트 단지내의 S^2_{d2hi} 과 D^2_{hi}

아파트 단지	평수	S^2_{d2hi}	D^2_{hi}	사구수	D^2_{hi}/S^2_{d2hi}
강남구 시영	10	0.0217	1107	300	50959
	13	0.0213	80203105	1000	3759625720
	17	2.4811	16852795	480	6792591
	19	0.0768	7969307	190	103705202
강동구 명일 LG	24	2.8801	96967	36	33668
	25	0.0874	19926611	274	227886920
	33	0.0864	3267276	44	37824886
	34	0.0927	2475058	58	26695378
	35	0.0918	42819225	360	466325654
강북구 수공 1 단지	17	0.1528	220067	420	1440076
	18	0.0001	5786	90	1439306486
	21	0.3850	587	480	1527
	26	0.0922	479478	100	5201301
	30	0.2059	468	100	2273
	31	5.1936	329636	240	63469
구로구 미성	15	0.3174	500	176	1575
	21	0.3428	4530759	176	13217979
	27	0.4088	5867795	173	14353188
	34	0.0205	14969096	299	7297269716
중구 현대	27	8.4926	3290160	155	387415
	32	2.7543	110677	177	40183
	33	0.0876	1051073	240	12004231
	43	3.4293	77707084	240	22659503
	49	22.217	14637518	120	658843
종로구 쌍용	23	3.2606	422433	257	129557
은평구 미성	28	3.4219	34399631	620	10052825
	33	0.1837	4286003	589	23327492
	39	12.4143	221991	190	17881
	42	0.7413	54886	84	74044
	35	3.9064	27060	540	6927
	47	3.0697	8837304	180	2878839
동작구 신동아	19	5.1733	25264	141	4883
	25	1.0146	3127	132	3082
	35	23.5750	1858525	228	78834
	40	44.0529	9990270	264	226779
동작구 한강현대	28	3.0809	10738844	300	3485578
	32	0.3521	24321428	270	69079755
	43	0.5360	17293227	180	32261971
	48	2.1473	254406549	210	118479929
금천구 한양	16	0.0313	10701304	545	342033342
	25	1.2561	20211625	285	16090289
	35	3.4517	61075660	675	17694167

이제 각 아파트 단지 내의 평수를 하나의 집락으로 생각하자. 표 3.4는 단지내의 평수를 하나의 집락으로 했을 경우의 S^2_{d2hi}과 D^2_{hi} 이다. 이제 표 3.4에서 얻어진 결과를 이용해서 몇 개의 평수를 묶어 하나의 집락으로 만들면 더 좋은 결과 를 얻을 수 있다. 예를 들어 강남구 시영아파트 경우 10 평대의 아파트만이 단지 내에 있다. 이 경우에는 10 평과 13 평을 묶어 하나의 집락으로 하고 17평과 19평을 하나로 묶어 다른 집락을 만든다. 이렇게 함으로써 D^2_{hi}/S^2_{d2hi}을 가구 수와 유사한 비율로 만들 수 있다. 다음 의 표 3.5는 이와 같은 방법을 이용하여 만든 집락과 그에 해당하는 S^2_{d2hi}, D^2_{hi} 그리고 D^2_{hi}/S^2_{d2hi}이다. 표 3.3과 표 3.4에서는 각 가구들 사이의 크기와 D^2_{hi}/S^2_{d2hi}의 크기에는 상당한 차이가 있다. 그러나 표 3.5을 살펴보면 D^2_{hi}/S^2_{d2hi}는 가구 수와는 선형관계가 존재하는 것을 알 수 있다. 표 3.3에서 표 3.5의 가구 수와 D^2_{hi}/S^2_{d2hi} 의 그림이 그림 3.3 에서 그림 3.5에 나와있다. 확연히 나타나지는 않지만 그림 3.5에서만 가구 수와 D^2_{hi}/S^2_{d2hi}는 선형관계가 존재하고 있음을 알 수 있다. 이를 알아보기 위하여 D^2_{hi}/S^2_{d2hi}를 종속변수로 가구 수를 독립변수로 하는 세 개의 회귀식을 추정하였다. 추정 결과 기울기에 관한 p-값이 각각 0.07, 0.08 그리고 0.02로 나와 표 3.5 만이 기울기가 유의한 것으로 나타났다.

표 3.5: 조정된 10 개 아파트 단지내의 평수별 S^2_{d2hi} 과 D^2_{hi}

아파트 이름	평수	S^2_{d2hi}	D^2_{hi}	가구수	D^2_{hi}/S^2_{d2hi}
강남구 시영	10, 13	14.625	79608114	1300	5443369
	17, 19	9.887	48000119	670	4854658
강동구 명일 LG	24, 25, 33	6.417	22803677	310	3553676
	34, 35	49.290	98494368	462	1998262
강북구 주공 1단지	17, 18, 21	0.516	324249	990	628292
	26, 30, 31	8.365	1659504	440	198392
구로구 미성	15, 21, 27, 34	28.462	71271829	824	2504068
중구 현대	27, 32, 33, 43, 49	224.466	250046067	932	1113960
종로구 쌍용	23, 26	5.713	4962539	641	868660
	33, 39, 42	5.955	3361299	863	375341
은평구 미성	28, 35, 47	36.537	81042880	1340	2218111
동작구 신동아	19, 25	3.275	46167	273	14095
	35, 40	43.446	20466723	492	471089
동작구 한강 현대	28, 32	15.248	67382652	570	4419047
	43, 48	697.337	404357273	390	579859
금천구 한양	16, 25, 35	14.897	242802177	1505	16298248

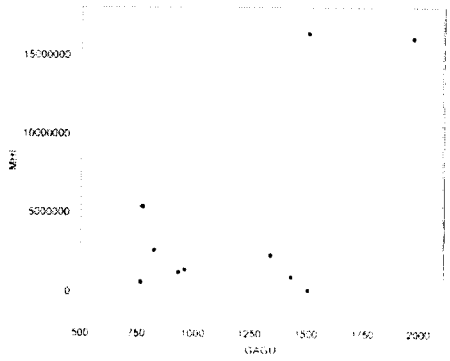

그림 3.3: 10개의 집락을 선택하였을 경우

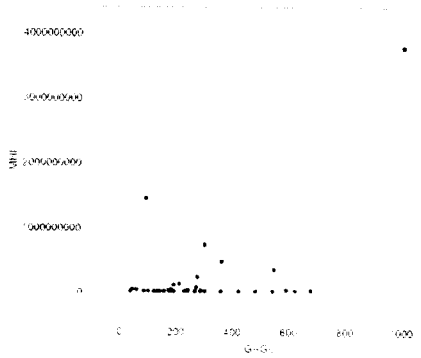

그림 3.4: 단지 내의 평수를 하나의 집락으로 했을 경우

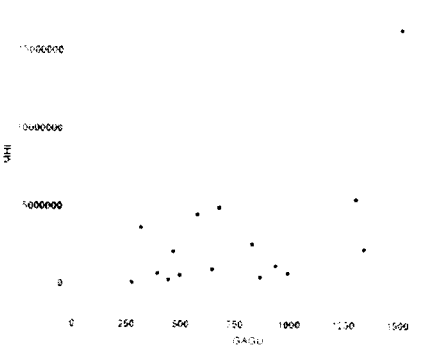

그림 3.5: 조정후의 가구 수와 D_{hi}^2/S_{d2hi}^2

다음으로 표 3.5에 나타난 가구수를 이용하여 m_{hi}^f 구하여 보자. n^f와 n_h^f는 층이 하나이므로 본 모의실험에서는 구할 수가 없다. 그러나 실제 문제에서는 (2.7)식 다음에 나와있는 식을 이용하면 구할 수 있을 것이다.

먼저 $m_{hi}^f = \sqrt{K_{hi} \cdot \frac{S_{d2hi}^2}{c_{2h} M_{hi}^f}}$ 에서 $K_{hi} = \frac{2c_{lh} + c_2^* - \sqrt{(2c_{lh}+c_2^*)^2 - 4c_{lh}^2}}{2S_{d2hi}^2}$ 이므로 $m_{hi}^f = \sqrt{C^* M_{hi}^f}$ 가 된다. 여기서

$$C^* = \frac{2c_{lh} + c_2^* - \sqrt{(2c_{lh} + c_2^*)^2 - 4c_{lh}^2}}{2c_{2h}}$$

이다. 따라서 c_{2h}에 비해 c_{lh}가 상대적으로 작아지면 이에 따라서 C^*의 값도 작아지게 된다. 이제 $c_{2h} = 10,000$를 고정하고 $c_{lh} = 1000, 500, 100$을 대입하자. 그러면 $C^* = 0.1101, 0.00227$, 0.0001이 되고 이에 따른 m_{hi}^f의 값은 각각 $m_{hi}^f = 0.332 M_{hi}^f$, $m_{hi}^f = 0.048 M_{hi}^f$ 그리고 $m_{hi}^f = 0.01 M_{hi}^f$ 이 된다. 따라서 표 3.5에 나타난 가구수를 M_{hi}^f라 하면 이에 따른 m_{hi}^f를 구할 수 있다.

4. 결 론

많은 사회조사에서는 대표본 추출시에 행정상 또는 조사 편의를 위하여 지역별로 또는 동일한 성질을 갖는 단위들을 층화한 후에 동질의 조사단위를 집락화하는 표본 설계를 고려한다. 본 논문에서는 아파트 가격 조사를 위하여 지역별로 그리고 동일한 평형의 아파트를 집락화하여 표본설계를 하는 경우 주어진 비용아래서 모집단을 집락화하여 층화 2-단 표본 추출을 할 때에 집락의 최적 크기를 결정하는 문제를 다루었다. 즉 주어진 비용 하에서 집락의 최적 크기 및 최적의 표본 추출률과 부차 표본 추출률을 구하였다. 모 총계에 대한 추정량으로는 분리 비 추정량과 결합 비 추정량을 사용하였다. 집락의 분산 S_{d2hi}^2과 D_{hi}^2이 집락의 크기 M_{hi}에 영향을 받지 않는 경우 집락의 분산이 클수록 집락의 크기를 작게 잡고 D_{hi}^2 클수록 집락의 크기를 작게 잡는 것이 모총계에 대한 비 추정량의 분산이 작게 되는 것으로 나타났다. 모의 실험 결과 아파트 단지를 집락으로 할 경우 S_{d2hi}^2과 D_{hi}^2이 집락의 크기 M_{hi}에 영향을 받지 않는 것으로 나타났으며 위의 결과를 이용하여 여러 평수를 합침으로써 더 좋은 결과를 얻을 수 있었다.

참고문헌

[1] 신민웅, 최기철, 이주영(1997). 층화 집락 표본추출에서 최적 추출률, 한국통계학회 *proceeding*.

[2] Cochran (1977). Sampling Technique, John Willy Sons.

[3] Durbin, J. (1953). Some results in sampling theory when the units are selected with unequal probabilities, *Jour. Royal Stat. Soc.*, 263-269.

[4] Efron, B. (1986). Double Exponential Families and Their Use in Generalized Linear Regression. em Journal of the American Statistical Association **81**, 709-721.

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

[6] Rao, J.N.K. (1975b). Unbiased variance estimation for multistage designs. Sankhya.

[7] Rao, J.N.K. (1988). Variance estimation in sample survey. In PR, Krishmaish and C.R. Rao(eds), *Handbook of statistics*, **6**(Sampling).

[8] Rao, P.S.R.S. and Rao, J.N.K. (1971). Small sample results for the ratio estimators. *Biometrika*, bf 58, 625-630.

[9] Royall R.M. (1988). The prediction approach to sampling theory. In PR, Krishmaish and C.R. Rao(eds), *Handbook of statistics*, **6** (Sampling)

[10] Scheaffer, R.L., Mendenhall, W. and Ott, L. (1990). Elementary Survey Sampling. Duxbury Press.

[11] Ten Have, T.R. and Chinchilli V.M. (1998). Two-Stage negative Binomial and Overdispersion Poisson Models for Clustered Developmental Toxicity Data With Random Cluster Size. *Journal of Agricultural, Biological, and Environmental Statistics*, **3**, No.1,75-98.

[1999년 10월 접수, 2000년 5월 채택]

부록 A. 분리 비 추정량

먼저 Lagrangian 승수법에 의하여, λ와 μ_h를 Lagrangian 승수로 잡고

$$\begin{aligned} V(\hat{Y}_{RS}) \quad &+ \lambda \left[\sum_{h=1}^{L} c_{uh} n_h + \sum_{h=1}^{L} c_{2h} \sum_{i=1}^{N_h} f_{0hi} M_{hi} \right. \\ &\left. + \sum_{h=1}^{L} c_{lh} \sum_{i=1}^{N_h} \pi_{hi} M_{hi} - E(C) \right] + \sum_{h=1}^{L} \mu_h (n_h - \sum_{i=1}^{N_h} \pi_{hi}) \end{aligned} \tag{4.1}$$

를 최소로 하는 n_h, f_{0hi}, π_{hi} 그리고 M_{hi}를 정하면 주어진 조건하에서 분산을 최소로 만드는 것과 같다. 먼저 n_h, π_{hi} 그리고 f_{0hi}에 관하여 미분하고 "0"으로 놓으면 결과는 다음과 같다.

$$n_h : \lambda c_{uh} + \mu_h = 0 \tag{4.2}$$

$$\pi_{hi} : -\frac{1}{\pi_{hi}^2} D_{hi}^2 + \frac{1}{\pi_{hi}^2} M_{hi} S_{d2hi}^2 + \lambda c_{lh} M_{hi} - \mu_h = 0 \tag{4.3}$$

$$f_{0hi} : -\frac{1}{f_{0hi}^2} M_{hi} S_{d2hi}^2 + \lambda c_{2h} M_{hi} = 0 \tag{4.4}$$

이제 M_{hi}에 대한 미분을 고려하자. 본 논문에서는 집락의 크기가 충분히 커서 D_{hi}^2와 S_{d2hi}^2이 M_{hi}의 크기에 영향을 받지 않고 일정한 경우를 살펴보자. M_{hi}의 미분은 다음과 같다.

$$M_{hi} : \left(\frac{1}{f_{0hi}} - \frac{1}{\pi_{hi}} \right) S_{d2hi}^2 + \lambda c_{lh} \pi_{hi} + \lambda c_{2h} f_{0hi} = 0 \tag{4.5}$$

(4.2), (4.3)에 의해서

$$\lambda = \frac{1}{\pi_{hi}^2} (D_{hi}^2 - M_{hi} S_{d2hi}^2)(c_{lh} M_{hi} + c_{uh})^{-1} = \frac{1}{\pi_{hi}^2 K_{hi}} \tag{4.6}$$

가 된다. 여기서

$$K_{hi} = (c_{lh} M_{hi} + c_{uh})(D_{hi}^2 - M_{hi} S_{d2hi}^2)^{-1} \tag{4.7}$$

이다. 또한 (4.6)에 의해서

$$f_{0hi}^2 = \pi_{hi}^2 K_{hi} \cdot \frac{S_{d2hi}^2}{c_{2h}} \tag{4.8}$$

이므로 (4.5)의 양변에 π_{hi}를 곱하면 $\left(\frac{\pi_{hi}}{f_{0hi}} - 1 \right) S_{d2hi}^2 + \lambda c_{lh} \pi_{hi}^2 + \lambda c_{2h} f_{0h} \pi_{hi} = 0$이 되고 이 식에 (4.6), (4.8)를 대입하면 다음과 같다.

$$\left(\frac{\sqrt{c_{2h}}}{\sqrt{K_{hi} \cdot S_{d2hi}^2}} - 1 \right) S_{d2hi}^2 + \frac{c_{lh}}{K_{hi}} + \frac{1}{\sqrt{K_{hi}}} \sqrt{\frac{S_{d2hi}^2}{c_{2h}}} = 0$$

이 식은 $\frac{c_{1h}}{S^2_{d2hi}} \le K_{hi}$인 K_{hi}의 2차 식이 되므로 쉽게 다음의 결과를 얻는다.

$$K_{hi} = \frac{2c_{1h} + c_2^* - \sqrt{(2c_{1h} + c_2^*)^2 - 4c_{1h}^2}}{2S^2_{d2hi}}$$

여기서 $c_2^* = c_{2h} + 2 + \frac{1}{c_{2h}}$이다. 이제 K_{hi}가 정해 지면 (4.7)에 의해서 최적의 M_{hi}^{opt}가 다음과 같이 정해진다.

$$M_{hi}^{opt} = [D_{hi}^2 K_{hi} - c_{uh}][c_{1h} + K_{hi} S^2_{d2hi}]^{-1} \tag{4.9}$$

$$n^{opt} = \frac{N \cdot E(C)}{\sum_{h=1}^{L} c_{uh} N_h + \sum_{h=1}^{L} c_{2h} N_h \sum_{i=1}^{N_h} \sqrt{K_{hi}} \sqrt{\frac{S^2_{d2hi}}{c_{2h}}} \frac{\left(M_{hi}^{opt}\right)^2}{M_{h0}} + \sum_h^L c_{1h} N_h \sum_{i=1}^{N_h} \frac{\left(M_{hi}^{opt}\right)^2}{M_{h0}}}$$

이제 최적의 n_h, m_{hi}를 구하자. 먼저 (2.2) 식에 (4.8)식의 f_{0hi}, (2.4)식의 $n_h = nN_h/N$ 그리고 $\pi_{hi} = n_h M_{hi}^{opt}/M_{h0}$을 이용하면 최적의 n을 구할 수 있다.

$$n^{opt} = \frac{N \cdot E(C)}{\sum_{h=1}^{L} c_{uh} N_h + \sum_{h=1}^{L} c_{2h} N_h \sum_{i=1}^{N_h} \sqrt{K_{hi}} \sqrt{\frac{S^2_{d2hi}}{c_{2h}}} \frac{\left(M_{hi}^{opt}\right)^2}{M_{h0}} + \sum_h^L c_{1h} N_h \sum_{i=1}^{N_h} \frac{\left(M_{hi}^{opt}\right)^2}{M_{h0}}}$$

구해진 n^{opt}과 (2.1), (2.4) 를 이용하면 최적의 n_h, m_{hi}는 다음과 같다.

$$n_h^{opt} = \frac{N_h \cdot E(C)}{\sum_{h=1}^{L} c_{uh} N_h + \sum_{h=1}^{L} c_{2h} N_h \sum_{i=1}^{N_h} \sqrt{K_{hi}} \sqrt{\frac{S^2_{d2hi}}{c_{2h}}} \frac{\left(M_{hi}^{opt}\right)^2}{M_{h0}} + \sum_h^L c_{1h} N_h \sum_{i=1}^{N_h} \frac{\left(M_{hi}^{opt}\right)^2}{M_{h0}}}$$

$$m_{hi}^{opt} = \sqrt{K_{hi} \cdot \frac{S^2_{d2hi}}{c_{2h}}} M_{hi}^{opt}$$

부록 B. 결합 비 추정량

$$V(\hat{Y}_{RC}) + \lambda[\sum_{h=1}^{L} c_{uh} n_h + \sum_{h=1}^{L} c_{2h} \sum_{i=1}^{N_h} f_{2hi} M_{hi} + \sum_{h=1}^{L} c_{lh} \sum_{i=1}^{N_h} \pi_{hi} M_{hi} - E(C)]$$

$$+ \sum_{h=1}^{L} \mu_h (n_h - \sum_{i=1}^{N_h} \pi_{hi})$$

를 최소로 하는 n_h, m_{hi} 그리고, M_{hi}를 정하면 된다. 먼저 n_h와 π_{hi} 그리고 f_{2hi}에 관하여 미분하면

$$n_h : \lambda c_{uh} + \mu_h = 0,\ \mu_h = -\lambda c_{uh} \tag{4.10}$$

$$\pi_{hi} : -\frac{1}{\pi_{hi}^2}\left(D_{hi}^2 - \frac{2M_{hi}}{M_{h0}} D_{hi} D_h + \frac{M_{hi}^2}{M_{h0}^2} D_h^2 - M_{hi} S_{d2hi}^2\right) + \lambda c_{lh} M_{hi} - \mu_h = 0 \tag{4.11}$$

$$f_{2hi} : -\frac{M_{hi}}{f_{2hi}^2} S_{d2hi}^2 + \lambda c_{2h} M_{hi} = 0 \tag{4.12}$$

이 된다. (4.10)을 (4.11)에 대입하여 풀면 다음의 식을 얻는다.

$$\lambda = \frac{1}{\pi_{hi}^2}\left(D_{hi}^2 - \frac{2M_{hi}}{M_{h0}} D_{hi} D_h + \frac{M_{hi}^2}{M_{h0}^2} D_h^2 - M_{hi} S_{d2hi}^2\right)[c_{lh} M_{hi} + c_{uh}]^{-1} = \frac{1}{\pi_{hi}^2 G_{hi}}$$

여기서 $G_{hi} = [c_{lh} M_{hi} + c_{uh}][D_{hi}^2 - \frac{2M_{hi}}{M_{h0}} D_{hi} D_h + \frac{M_{hi}^2}{M_{h0}^2} D_h^2 - M_{hi} S_{d2hi}^2]^{-1}$ 이다. 그러면, (4.12)에서

$$f_{2hi}^2 = \frac{S_{d2hi}^2}{\lambda c_{2h}} = \pi_{hi}^2 G_{hi} \frac{S_{d2hi}^2}{c_{2h}} \tag{4.13}$$

가 된다. 문제를 간단히 하기 위하여 집락의 크기, M_{hi}가 충분히 커서 D_{hi}, D_h 그리고S_{d2hi}^2가 영향을 받지 않고 일정한 경우에 M_{hi}에 관하여 미분하면 다음의 결과를 얻는다.

$$M_{hi} : -\frac{2D_{hi} D_h}{M_{h0} \pi_{hi}} + \frac{M_{hi} D_h^2}{M_{h0}^2 \pi_{hi}} - \frac{S_{d2hi}^2}{\pi_{hi}} + \frac{S_{d2hi}}{f_{2hi}} + \lambda c_{lh} \pi_{hi} + \lambda c_{2h} f_{0hi} = 0 \tag{4.14}$$

양변에 π_{hi}를 곱하고 G_{hi}에 관하여 풀면 다음의 식이 된다.

$$\left(-\frac{2D_{hi} D_h}{M_{h0}} + \frac{2M_{hi} D_h^2}{M_h 0^2} - S_{d2hi}^2\right)^2 G_{hi}^2$$
$$+ \left[2\left(\frac{-2D_{hi} D_h}{M_{h0}} + \frac{2M_{hi} D_h^2}{M_{h0}^2} - S_{d2hi}^2\right) c_{lh} - \left(\frac{\sqrt{c_{2h}}}{\sqrt{S_{d2hi}^2}} + \frac{\sqrt{S_{d2hi}^2}}{\sqrt{c_{2h}}}\right)^2\right] G_{hi} + c_{lh}^2 = 0 \tag{4.15}$$

이 식은 M_{hi}의 4차식이 되어 M_{hi} 가 구해진다. 구해진 최적의 크기를M_{hi}^{opt}라 하자. 그러면 (2.2) 식에 (4.13)식의 f_{2hi}, (2.4)식의 $n_h = nN_h/N$ 그리고 $\pi_{hi} = n_h M_{hi}^{opt}/M_{h0}$을 이용하면 다음과 같은 최적의 n을 구할 수 있다.

$$n^{opt} = \frac{N \cdot E(C)}{\sum_{h=1}^{L} c_{uh} N_h + \sum_{h=1}^{L} c_{2h} N_h \sum_{i=1}^{N_h} \sqrt{G_{hi}} \sqrt{\frac{S_{d2hi}^2}{c_{2h}}} \frac{\left(M_{hi}^{opt}\right)^2}{M_{h0}} + \sum_{h}^{L} c_{1h} N_h \sum_{i=1}^{N_h} \frac{\left(M_{hi}^{opt}\right)^2}{M_{h0}}}$$

구해진 n^{opt}과 (2.1), (2.4) 를 이용하면 최적의 n_h^{opt} ,m_{hi}^{opt}는 다음과 같다.

$$n_h^{opt} = \frac{N_h \cdot E(C)}{\sum_{h=1}^{L} c_{uh} N_h + \sum_{h=1}^{L} c_{2h} N_h \sum_{i=1}^{N_h} \sqrt{G_{hi}} \frac{\sqrt{S_{d2hi}^2}}{c_{2h}} \frac{\left(M_{hi}^{opt}\right)^2}{M_{h0}} + \sum_{h}^{L} c_{1h} N_h \sum_{i=1}^{N_h} \frac{\left(M_{hi}^{opt}\right)^2}{M_{h0}}}$$

$$m_{hi}^{opt} = \sqrt{G_{hi} \cdot \frac{S_{d2hi}^2}{c_{2h}}} M_{hi}^{opt}$$

A Optimal Cluster Size in Stratified Two-Stage Cluster Sampling

Minwoong Shin [1] Key-Il Shin [2]

ABSTRACT

Generally cluster size is predetermined when we use the stratified two-stage cluster sampling. But in case that the sizes of clusters vary greatly one may want to make the sizes to be about equal. In this paper we study the optimal cluster size in stratified two-stage cluster sampling. Also we find the optimal primary sampling unit sizes and optimal secondary sampling unit sizes under the given cost restriction.

Keywords: Two-Stage cluster sampling; Combined ratio estimate; separate ratio estimate; Optimal cluster size.

1) Professor, Department of Statistics, Hankuk University of Foreign Studies.
E-mail: mwshin@stat.hufs.ac.kr
2) Associate Professor, Department of Statistics, Hankuk University of Foreign Studies.
E-mail: keyshin@stat.hufs.ac.kr