층화 집락추출시 m_{opt}의 결정

신동윤[1] 신민웅[2] 최기철[3]

요 약

표본조사를 하는 경우에 사전에 전체 표본의 크기를 정하여 놓고, 표본설계를 하는 경우가 많다. 이 때에는 조사 비용은 고려의 대상이 안되고 주어진 전체표본 크기로 각 층별로 표본을 할당하여 분산을 최소로 하는 문제가 된다.

이 논문에서는 pps 집락추출과 각 집락에서 같은 크기의 부표본(subsample)을 추출하여 자체 가중이 되도록 표본설계를 하는 경우에 표본의 크기 m_0가 사전에 주어졌을 때에 모총계의 추정량의 분산을 최소로 하는 최적의 표본추출율을 구하고. 이러한 m_0값들 중에서 최적의 m_{opt}값을 구한다.

주요용어 : m_{opt} , pps, 최적의 선택확률, 표본추출율

1. 서론

표본조사를 할 때에 사전에 전체 표본의 크기가 정하여 지는 경우가 많다. 이 때에는 조사 비용은 고려의 대상이 안되고 주어진 표본 크기로 각 층별로 표본을 할당하여 분산을 최소로 하는 문제가 된다. 대규모의 표본 설계를 할 때 층화 이-단 표본 추출법이 주로 사용되고 있다.

많은 사회조사에서는 행정상으로나 조사의 편의를 위하여 원소들을 동질적으로 층화한 후 인근의 조사 단위들을 같은 집락으로 묶는다. 이단 집락추출법이란 Scheaffer(1990)등에 의하면 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 대한 확률표본을 추출하고, 추출된 집락내에서 조사 단위들을 이차로 추출하는 것을 의미한다.

이단 집락 표본 추출을 하는데 문제는 집락들을 적절하게 묶는 것이다. 이 때 고려하여야 할 점은 집락 내에 있는 조사 단위들의 지리적 인접성과 표본조사를 하기 위한 관리의 편리성 등이다.

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

1) (449-791) 경기도 용인시 모현면 한국외국어대학교 정보통계학과 박사과정

2) (449-791) 경기도 용인시 모현면 한국외국어대학교 정보통계학과 교수

3) (608-738) 부산광역시 남구 우암동 55-1. 부산외국어대학교 통계학과 교수

는 지에 따라 달라진다.

큰 집락들은 이질적인 조사 단위들을 포함하는 경향이 있으므로 모집단 모수에 대한 정확한 추정이 요구될 때는 각 집락으로부터 많은 표본을 추출하는 것이 필요하나, 작은 집락들은 상대적으로 동질적인 조사 단위들을 포함하므로 이 경우에는 각 집락으로부터 표본을 적게 추출하여도 집락 특성에 관한 정확한 정보를 얻을 수 있다.

층화 이단계 표본추출은 모집단을 L개의 층으로 나누고, h층의 N_h개의 일차단위(집락)로부터 n_h개의 일차단위를 추출하고, 추출된 h층의 i번째 집락의 크기가 M_{hi}인 집락에서 m_{hi}개의 이차단위(부차단위)를 srs(simple random sampling)로 추출하는 것을 말한다.

이 논문에서는 집락들이 층화되었을 때에 각 층에서 pps로 집락을 일차 추출단위로 뽑고, 추출된 집락내에서 다시 똑같은 크기의 부차단위들을 추출하는 층화 이단계 표본추출을 생각한다. 모집단은 L개의 층으로 이루어졌고, h층의 N_h개의 일차단위(집락)로부터 n_h개의 일차단위를 pps로 추출한다. 그리고, 추출된 h층의 i번째 집락의 크기가 M_{hi}인 집락에서 m_0개의 이차단위(부차단위)를 srs(simple random sampling)로 추출한다. 즉, 모든 집락에서 똑같은 크기의 이차단위를 추출하여 자체-가중이 되도록 한다.

우리는 총 표본의 크기가 사전에 정해졌을 때에 모총계 Y의 추정량의 분산을 구하고, 이 분산을 최소로 하는 층화 이단계 표본추출시에, 층별 최적의 표본크기를 구하는 문제를 생각한다.

2장에서는 전체 표본크기가 미리 정해졌을 때에 층별 최적 표본크기를 구하는 과정을 설명한다. 3장에서는 층화 이단 집락표본 추출시에 m_0 중에서 최적의 m_{opt}를 구한다.

2. 표본크기가 미리 주어진 경우에 최적 선택확률

표본설계의 목적은 전체 표본의 크기가 사전에 주어졌을 때에 pps 집락 추출을 하고, 추출된 집락내에서 다시 똑같은 크기의 부차단위들을 추출하는 층화 이단계 표본추출을 할 때에 모총계 Y의 불편추정량(unbiased estimate) $\hat{Y}_{ST}$의 분산 $V(\hat{Y}_{ST})$을 최소로 하는 층별 표본의 크기, n_h와 층별 최적 선택확률 f_{0h}를 구하는 것이다.

여기서, $\hat{Y}_{ST} = \sum_h \hat{Y}_h$이다. 그리고, y_h는 h층의 표본총계이고, Y_h는 h층의 층총계이다. Y_{hi}는 h층의 i번째 집락의 총계이고, y_{hi}는 h층의 i번째 집락의 표본 총계이다. 그리고 $\hat{Y}_h$는 h층의 총계의 추정량이다. 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(1977)에 의하면 ppz추정량이라 함은 집락을 z_i(집락의 크기를 추정하여 할당하는 확률)에 확률비례하여 추출했을 때에 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.

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

$$V(\hat{Y}_{ST}) = \sum_{h}^{L} V(\hat{Y}_h) = \sum_{h}^{L} \frac{1}{n_h} \sum_{i}^{N_h} \left[z_{hi} \left(\frac{Y_{hi}}{z_{hi}} - Y_h \right)^2 + \frac{M_{hi}(M_{hi} - m_0)}{z_{hi} m_0} S_{2hi}^2 \right]$$

$$= \sum_{h}^{L} \frac{1}{n_h} \sum_{i}^{N_h} \left[\frac{1}{z_{hi}} (Y_{hi} - z_{hi} Y_h)^2 + \frac{M_{hi}(M_{hi} - m_0)}{z_{hi} m_0} S^2_{2hi} \right] \tag{2.1}$$

이다. $d_{hij} = y_{hij} - z_{hi}\left(\sum_i y_{hij}\right)$로 놓으면 $(Y_{hi} - z_{hi} Y_h) = M_{hi} \bar{D}_{hi}$이다.

따라서, $\pi_{hi} = n_h z_{hi}$ 와 $M_{hi}/n_h z_{hi} m_0 = 1/f_{0h}$에서

$$V(\hat{Y}_{ST}) = \sum_{h}^{L} \sum_{i}^{N_h} \left[\frac{M_{hi}^2}{\pi_{hi}} \left(\bar{D}_{hi}^2 - \frac{S_{2hi}^2}{M_{hi}} \right) + \frac{M_{hi}}{f_{0h}} S_{2hi}^2 \right] \tag{2.2}$$

이다. 여기서

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

이다.

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

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

$$m_0 = (f_{0h} M_{hi}) / (n_h z_{hi}) = (f_{0h} M_{hi}) / \pi_{hi}$$

$$= (f_{0h} M_{hi}) / \left(n_h \frac{M_{hi}}{M_{h0}} \right) = f_{0h} M_{h0} / n_h \tag{2.3}$$

이라고 가정한다. 여기서, $\pi_{hi} = n_h z_{hi}$ 로 h층의 i번째 집락이 표본으로 추출 될 확률이다.

3. m_{opt}의 결정

정리 1.

표본추출을 하는데 층화 이단 집락추출시에 전체 표본의 크기 $\sum n_h m_{hi}$가 비용에 관계없이 미리 정해진 경우를 생각한다. 층화 이단 집락추출하는데, 일차단위(집락)을 복원으로 z_{hi}에 확

률비례하여 추출하고, 특히 $z_{hi} = M_{hi}/M_{h0}$ (pps)인 경우를 생각한다. 여기서, $M_{h0} = \sum_i M_{hi}$ 이다.

미리 m_o가 정해졌다면 분산 V를 최소로 하는 층별 n_h는

$$n_h = \frac{M_{h0} f_{0h}}{m_0} = \frac{M_{h0}(k_h f_{0s})}{m_0} \tag{3.1}$$

이다.

미리 m_o가 정해지지 않았다면 최적의 m_o을 m_{opt}라 하자.
그러면

$$n_h = M_{h0} f_{0h}/m_{opt}$$

이다.

증명. 먼저, 아래의 (3.2)식에서 f_{0h}를 구한다.
즉,

$$f_{0h} \propto \frac{1}{\sqrt{c_{2h}}} \sqrt{\sum_i (M_{hi}/M_{h0}) S_{2hi}^2} \tag{3.2}$$

에서 구한 f_{0h}, $h = 1,2,\cdots,L$들 중에서 제일 작은 값을 f_{0s}라 하자. 그러면, $f_{0h} = k_h f_{0s}$, $h = 1,2,\cdots,L$ 이고, k_h는 기지의 상수가 된다.
m_0가 미리 주어졌다면 $\hat{Y}_{ST}$를 전체적으로 자체-가중(self-weighting)으로 만들기 위하여

$$m_0 = (f_{0h} M_{hi})/\pi_{hi}$$

$$= f_{0h} M_{h0}/n_h$$

이라고 가정한다. 여기서, $\pi_{hi} = n_h z_{hi}$ 으로 h층의 i번째 집락이 표본으로 추출 될 확률이다.

총 표본수가 주어졌으므로 (2.4)에서

$$\sum_h^L \frac{k_h f_{0s} M_{h0}}{m_0} = \text{총표본수} \tag{3.3}$$

이다.
(2.6)에서 k_h가 기지이므로 f_{0s}를 구할 수 있다.
그러면 $f_{0h} = k_h f_{0s}$이므로

$$n_h = \frac{M_{h0} f_{0h}}{m_0} = \frac{M_{h0}(k_h f_{0s})}{m_0} \tag{3.4}$$

이다.

미리 m_o가 정해지지 않았다면 최적의 m_o을 m_{opt}라 하자. 최적의 m_{opt}를 구하기 위하여 초기값으로 m_0 = (과거의 표본 설계시 값)으로 놓는다. 행정상의 편리성으로 정한 조건 $a \leqq m_0 \leqq b$ 에서 최적의 m_{opt}를 구한다.

그러면

$$n_h = M_{h0} f_{0h} / m_{opt} \tag{3.5}$$

이다.

참 고 문 헌

[1] Cochran(1977). Sampling Technique, John Willy & 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] Thompson(1992). Sampling, John Wiley & Sons. Inc.