층화이단 pps 집락 추출시 층별 최적 표본추출율

신민웅[1)]

1. 서 론

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

우리는 주어진 비용아래서 모총계 Y의 추정량의 분산을 최소로하는 층화 이단계 표본추출을 할 때에, 층별 최적 표본추출율을 구하는 문제를 생각한다.

2장에서는 주어진 비용아래서 층별 최적 표본추출율을 구하는 과정을 설명한다. 3장에서는 표본크기가 미리 주어진 경우에 층별 최적 표본추출율을 계산하여 구한다.

2. pps 추출시에 최적의 표본추출율(sampling fraction)과 최적선택확률

우리는 주어진 비용아래서 모총계 Y의 불편 추정량(unbiased estimate) $\hat{Y}_{ST}$의 분산 $V(\hat{Y}_{ST})$을 최소로 하는 n_h와 층별 최적 선택확률 f_{0h}를 구한다.

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 모총계이다. 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_i에 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_{hi}는 h층의 i번째 단위가 추출될 확률로

---

1) 한국 외국어대학교 자연과학대학 통계학과,교수,(449-791)경기도 용인시 모현면 왕산리 산 89 E-mail : mwshin@stat.hufs.ac.kr

$\sum_i z_{hi} = 1$ 이다.

이 논문에서는 일차단위(집락)를 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

$$\begin{aligned} m_0 &= (f_{0h} M_{hi})/(n_h z_{hi}) = (f_{0h} M_{hi})/\pi_{hi} \\ &= (f_{0h} M_{hi})/(n_h \frac{M_{hi}}{M_{h0}}) = f_{0h} M_{h0}/n_h \end{aligned} \tag{2.1}$$

이라고 가정한다. 여기서, 비용함수는

$$C = \sum_h c_{uh} n_h + \sum_h (c_{2h} \sum_i^{n_h} m_{hi}) \tag{2.2}$$

이다. 비용함수에 포함되는 항들은

c_{uh} = h층의 일차단위 당 고정비용

c_{2h} = h층의 부차단위 당 비용

이다.

그런데, (2.1)에서 $n_h m_0 = f_{0h} M_{h0}$이므로

$$E(C) = \sum_{h=1}^{L} c_{uh} n_h + \sum_h (c_{2h} f_{0h} M_{h0}) \tag{2.3}$$

이다.

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

$$\begin{aligned} V(\hat{Y}_{ST}) &= \sum_h^L V(\hat{Y}_h) = \sum_h^L \frac{1}{n_h} \sum_i^{N_h} \left[z_{hi} (\frac{Y_{hi}}{z_{hi}} - Y_h)^2 + \frac{M_{hi}(M_{hi} - m_0)}{z_{hi} m_0} S_{2hi}^2 \right] \\ &= \sum_h^L \frac{1}{n_h} \sum_i^{N_h} [\frac{1}{z_{hi}} (Y_{hi} - z_{hi} Y_h)^2 + \frac{M_{hi}(M_{hi} - m_0)}{z_{hi} m_0} S_{2hi}^2] \end{aligned} \tag{2.4}$$

이다. $d_{hij} = y_{hij} - z_{hi}(\sum_i y_{hij})$로 놓으면 $(Y_{hi} - z_{hi} Y_h) = M_{hi} \overline{D}_{hi}$ 이다. 따라서, $\pi_{hi} = n_h z_{hi}$와 $M_{hi}/n_h z_{hi} m_0 = 1/f_{0h}$ 에서

$$V(\hat{Y}_{ST}) = \sum_h^L \sum_i^{N_h} \left[\frac{M_{hi}^2}{\pi_{hi}} (\overline{D}_{hi}^2 - \frac{S_{2hi}^2}{M_{hi}}) + \frac{M_{hi}}{f_{0h}} S_{2hi}^2 \right] \tag{2.5}$$

이다. 여기서

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

이다. 조건

$$z_{hi} n_h \left(\frac{m_0}{M_{hi}} \right) = f_{0h} \tag{2.6}$$

는 h층의 i번째 집락내의 이차단위가 추출되는 확률이다.
특수한 경우로는 만약 주어진 비용 아래서, f_{0h}가 미리 선택된다면 n_h는 식 (2.3)에서 구할수 있다. 그리고, m_0는 식 (2.1)에서 구할수 있다.
일반적인 경우로 우리는 고정된 평균 비용 (2.3)과

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

인 조건에서, V를 최소화 하는 n_h, f_{0h}를 정하고자 한다. 그러면, Lagrangian 승수법에 의하여, λ와 μ_h를 Lagrangian 승수로 잡고

$$V+\lambda\Big[\sum_{h}^{L} c_{uh}n_h+\sum_{h=1}^{L} c_{2h}f_0M_{h0}-E(C)\Big]+\sum_{h=1}^{L}\mu_h\,(n_h-\sum_{i=1}^{N_h}\pi_{hi}) \tag{2.7}$$

를 최소로 한다.
특히, pps 표본추출의 경우를 생각하여

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

이라 놓자. 여기서 $M_{ho}=\sum_{i=1}^{N_h}M_{hi}$이다. 식 (2.7)를 n_h와 π_{hi}에 관하여 미분하면

$$n_h : \lambda c_{uh}+\mu_h=0. \text{ 즉 } \mu_h=-\lambda c_{uh}$$

$$\pi_{hi} : -\frac{M_{hi}^{\,2}}{\pi_{hi}^{\,2}}\Big(\overline{D}^{2}_{hi}-\frac{S_{2hi}^{\,2}}{M_{hi}}\Big)+\lambda c_{uk}=0 \tag{2.8}$$

그리고 표본추출율 f_{0h}에 관하여 (2.7)을 미분하면

$$\sum_{i}\frac{-M_{hi}}{f_{0h}^{\,2}}\,S_{2hi}^{\,2}+\lambda c_{2h}M_{h0}=0 \tag{2.9}$$

이다. 그러면 (2.9)에서 최적선택확률은

$$f_{0h}^{\,2}=\frac{\sum_{i}M_{hi}S_{2hi}^{\,2}}{\lambda c_{2h}M_{h0}} \tag{2.10}$$

이다. 그러면, (2.8)에서 λ는

$$\lambda=\frac{\sum_{h}^{L}\sum_{i=1}^{N_h}c_{uh}}{\sum_{h}^{L}\sum_{i}^{N_h}\Big[\frac{M_{hi}^{2}}{\pi_{hi}^{\,2}}\Big(\overline{D}_{hi}^{\,2}-\frac{S_{2hi}^{2}}{M_{hi}}\Big)\Big]} \tag{2.11}$$

이다.
그러면 식 (2.10)에 $\pi_{hi}=n_hz_{hi}$를 대입하여 식 (2.3)과 연립하여 풀어서 n_h값을 구할

수 있다. 그리고 식 (2.10)에 앞에서 구한 n_h값을 대입하여 f_{0h}값을 구할 수 있다.

3. 표본크기가 미리 주어진 경우에 최적 선택확률

층화 이단 집락추출시에 전체 표본의 크기 $\sum n_{u}m_{hi}$가 비용에 관계없이 미리 정해져서 층별로 집락의 수, n_h와 이차단위의 수 m_{hi}만 결정하면 되는 경우도 많이 나타난다. 이 때에 최적선택확률을 f_{0h}로 미리 정한다. 그리고 pps로 집락을 표본추출하고, 추출된 각 집락에서 같은 크기의 ssu를 추출한다면, 자체-가중 표본추출이 된다.

3.1 f_{0h}와 m_0가 정해졌을 때

표본크기 $\sum_h n_h m_0$가 주어졌을 때에 f_{0h}와 m_0를 미리 정하여 $m_0 = f_{0h}M_{h0} / n_h$에서 n_h를 구할 수 있다.여기서 m_0는 표본 설계자가 미리 정하는 값이고, f_{0h}는 $\sum_h m_h m_0$, $M_{hi}/n_{hz_{hi}}m_0 = 1/f_{0h}$와 Cochran (11.74)로 부터 정할수 있다. 이 때에

$$\hat{Y}_{ST} = \sum_h \hat{Y}_h$$

$$= \sum_h \left(\frac{1}{n_h}\sum_i^{n_h}\frac{M_{hi}y_{hi}}{m_0 z_{hi}}\right)$$

(2.4)와 $M_{hi} / n_h z_{hi} m_0 = 1/f_{0h}$ 에서

$$V(\hat{Y}_{ST}) = \sum_h^L \sum_i^{N_h}\left[\frac{M_{hi}^2}{\pi_{hi}}\left(\overline{D}_{hi}^{\,2} - \frac{S_{2hi}^2}{M_{hi}}\right) + \frac{M_{hi}}{f_{0h}}S_{2hi}^2\right] \qquad (3.1)$$

4. 결 론

표본조사에서 비용이 미리 정해졌을 때에, 모총계 추정량의 분산을 최소로 하는 층화 이단계 pps 표본 추출을 하는 표본설계에서 층별 최적의 표본추출율과 최적의 선택확률을 정하는 문제를 생각하였다. 또한 전체 표본의 크기가 사전에 정해졌을 때에 층화 이단계 pps 표본추출을 할 경우 층 별 집락의 수를 정하였다.

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.