The Korean Communications
in Statistics Vol. 11 No. 2, 2004
pp. 297-311

Sample Size Determination Using the Stratification Algorithms with the Occurrence of Stratum Jumpers

Taekyong Hong[1], Jihun Ahn[2], and Pyong Namkung[3]

Abstract

In the sample survey for a highly skewed population, stratum jumpers often occur. Stratum jumpers are units having large discrepancies between a stratification variable and a study variable. We propose two models for stratum jumpers : a multiplicative model and a random replacement model. We also consider the modification of the L-H stratification algorithm such that we apply the previous models to L-H algorithm in determination of the sample sizes and the stratum boundaries. We evaluate the performances of the new stratification algorithms using real data. The result shows that L-H algorithm for the random replacement model outperforms other algorithms since the estimator has the least coefficient of variation.

keywords : stratum jumper, Lavallée-Hidiroglou stratification algorithm, power allocation, Neyman allocation

1. 서 론

왜도가 큰 모집단에서 층화추출을 할 때 보통 큰 단위에 대해 모든 단위를 추출하는 하나의 전수 층과, 나머지 단위에 대해 확률적으로 추출하는 여러 개의 표본 층을 갖도록 표본설계를 한다. 이 경우에 단위의 값이 커짐에 따라 추출율은 커지고, 표본가중값은 작아지게 된다. 층화설계시 대개 조사변수와 상관이 높은 보조변수를 이용하는데, 실제 분석에서는 층화변수와 조사변수가 일치하지 않는 경우가 발생한다. 이러한 불일치의 한 예가 stratum jumper이다(Rivest 1999). stratum jumper는 층화변수를 이용하여 표본설계를 하는 당시에는 작은 값을 가지던 단위가 짧은 시간내에 급격히 커져 실제 조사시점에서는 큰 값을 갖게 되는 경우에 발생한다. 또한, stratum jumper는 큰 표본가중값을 갖기 때문에 이 단위가 표본에 포함되는 경우 조사추정량에 큰 영향을 준다(Lee 1995). 대개 추정단계에서 이상값 처리 방법을 이용하여 이 문제를 해결한다.

본 논문에서는 stratum jumper의 문제를 설계단계에서 해결할 수 있는 방법을 제시한다. 우선 stratum jumper를 나타내는 두 모형인 승법모형과 확률복원모형에 대해 알아보고, 이 모형을 왜도가 높은 모집단에 대한 층화알고리즘인 Lavallée-Hidiroglou 알고리즘(1988)에 적용하여 벤처

1) Lecturer, Division of Computer Engineering, Hansung University, Seoul.
E-mail : hongstat@skku.edu
2) Researcher, Samsung Electro-mechanics, Co., Ltd., Suwon Kyonggi-Do.
3) Professor, Dept. of Statistics, Sungkyunkwan University, Seoul.

기업 자산자료를 분석하여 효율성을 비교한다.

2. Stratum Jumper

2.1 stratum jumper의 개념

stratum jumper는 잘못된 층화변수에 의해 발생하는 것으로 층화변수가 표본설계 당시에는 적절했으나 조사시점까지의 짧은 시간 동안에 몇몇 단위의 값이 크게 달라져 조사변수의 값이 층화변수와 상당한 차이가 있는 경우를 의미한다. stratum jumper에는 작은 단위가 커지는 경우와 큰 단위가 작아지는 경우가 있는데, 여기서는 층화시점에서 작았던 단위가 조사시점에서 커지는 경우만을 고려한다. 그 이유는 현실 상황에서 가능성이 높고, 단위의 큰 값이 큰 표본 가중값과 결합하여 조사추정값이 과도하게 커지는 문제가 발생할 수 있기 때문이다. 반면, 단위의 값이 작아지는 경우는 단위가 작은 표본 가중값을 갖기 때문에 상대적으로 큰 문제가 되지 않는다. stratum jumper가 있는 경우 이를 고려하지 않고 추정하게 되면 추정량은 신뢰할 수 없을 정도로 변동이 커진다.

stratum jumper가 있는 경우 일반적으로 추정단계에서 이상값(outlier) 처리방법을 사용하는데, stratum jumper가 발생한 단위의 값을 변화시키는 방법인 Winsorization과 그 단위의 표본가중값을 줄여서 추정하는 방법인 가중값 축소가 많이 사용된다.

여기서는 표본설계 시 stratum jumper를 처리하는 방법을 제시하여 Lavallée-Hidiroglou (L-H) 층화알고리즘에 stratum jumper를 나타내는 모형을 적용한다.

2.2 stratum jumper에 대한 모형

stratum jumper에 대한 모형에서는 대부분의 단위에서 두 변수의 값이 같고 몇 개의 단위에서만 값이 크게 다름을 가정한다. 자료 $\{x_i,\ i=1,\ldots,N\}$는 이미 알려진 층화변수 X, $\{y_i,\ i=1,\ldots,N\}$는 알려져 있지 않은 조사변수 Y의 독립적인 N개의 값이고, N은 모집단 크기이다. X와 Y가 연속확률변수이고, $f(x)$, $g(y)$ $x,\ y\in R$이 각각 X, Y의 밀도함수를 나타낸다고 하자. $-\infty = b_0 < b_1 < b_2 < \ldots < b_L = \infty$ 가 층 경계일 때, 층 h는 X값이 구간 $(b_{h-1}, b_h]$ 에 속하는 단위들로 이루어진다.

2.2.1 승법모형

승법모형에서는 $Y=XZ$ 를 고려한다. 여기에서 Z는 X와 독립적으로 분포하는 확률변수로서 다음의 확률함수를 따른다.

$$\Pr(Z=z)=\begin{cases}1-\varepsilon, & \text{if } z=1\\ \varepsilon, & \text{if } z=\mathrm{M}\end{cases} \tag{2.1}$$

ε은 (0 , 1)의 값이다. $M>1$인 경우, M은 승법팽창인자(multiplicative inflation factor)이다.

2.2.2 확률복원모형

승법모형은 두 개의 모수에 의존하지만, 확률복원모형은 단위의 Y의 값이 모집단에서 확률적으로 선택된 단위의 X값과 같을 확률인 ε에만 의존한다.

$$Y=\begin{cases} X & 1-\varepsilon\text{의 확률로} \\ X_{new} & \varepsilon\text{의 확률로} \end{cases} \tag{2.2}$$

여기에서 X_{new}는 X와 독립적인 분포를 가지는 확률변수이다.

3. 층화알고리즘과 층화알고리즘의 일반화

3.1 Lavallée-Hidiroglou 층화알고리즘

본 장에서는 모집단을 하나의 전수층과 여러 개의 표본층으로 나누는 층화알고리즘을 제시하는데, 알고리즘의 목적은 주어진 정도와 표본층에 대한 배분방법을 만족시키는 총 표본크기를 최소화하는 것이다. 각 표본층에 대한 추출방법으로는 비복원 단순임의추출을 가정하고, 배분방법으로는 Bankier(1988)가 제안한 Y-비례 파워 배분(Y-proportional power allocation)과 네이만 배분(Neyman allocation)을 사용한다.

여기서 사용한 층화추출의 기본적인 기호들을 정의하면 L은 층의 수, W_h는 h층에 대한 가중값, $\overline{Y}_h$는 h층의 모평균, $\overline{y}_h$은 h층의 표본평균, S_{yh}^2는 h층의 모분산, a_h는 총 표본크기에 대한 h층에 포함되는 표본크기의 비율, c는 목표변동계수이다.

모평균 $\overline{Y}$에 대한 조사추정량은 $\overline{y}_{st}=\sum W_h \overline{y}_h$이고, 추정량의 분산식을 총표본크기 n에 대해서 정리하면 다음과 같다.

$$n=NW_L+\frac{\sum_{h=1}^{L-1} W_h{}^2 S_{yh}^2/a_h}{\overline{Y}^2c^2+\sum_{h=1}^{L-1} W_h S_{yh}^2/N} \tag{3.1}$$

파워 배분과 네이만 배분에서의 a_h는 식(3.2)와 식(3.3)과 같다.

$$a_h=(W_h\overline{Y}_h)^p/\sum_{k=1}^{L-1}(W_k\overline{Y}_k)^p \tag{3.2}$$

여기서 p는 $(0,1]$ 의 수,

$$a_h = W_h S_{yh} / \sum_{k=1}^{L-1} W_k S_{yk} \tag{3.3}$$

한편 Lavallée-Hidiroglou 알고리즘에서는 $X = Y$ 임을 가정한다. 그런데 확률변수 X는 밀도함수 $f(x)$를 갖기 때문에 범위가 $b_{h-1} < X < b_h$ 이라는 조건을 부여한 Y의 조건부 적률은 다음의 세 가지 식으로 나타내어질 수 있다.

$$W_h = \int_{b_{h-1}}^{b_h} f(x)\,dx \tag{3.4}$$

$$\Phi_h = \int_{b_{h-1}}^{b_h} x f(x)\,dx \tag{3.5}$$

$$\Psi_h = \int_{b_{h-1}}^{b_h} x^2 f(x)\,dx \tag{3.6}$$

X값에 의해 층화가 이루어지며 h층에 속하는 Y의 평균과 분산 대신 단위가 층 h에 속할 때 Y의 조건 평균 $E(Y|b_h \geq X > b_{h-1})$과 조건 분산 $Var(Y|b_h \geq X > b_{h-1})$을 사용한다. 층화를 위해 식(3.1)를 Y에 대한 조건 평균과 분산의 형태로 다시 쓰면 식(3.7)과 같다.

$$n = NW_L + \frac{\sum_{h=1}^{L-1} W_h^2 Var(Y|b_h \geq X > b_{h-1}) / a_{h,X}}{\overline{Y}^2 c^2 + \sum_{h=1}^{L-1} W_h Var(Y|b_h \geq X > b_{h-1}) / N} \tag{3.7}$$

여기에서 $a_{h,X}$는 이미 알려진 X로 나타내어진 배분방법이다. 파워 배분과 네이만 배분에서의 a_h인 식(3.2)와 식(3.3)을 Y의 조건 평균과 조건 분산으로 나타내면 식(3.8), (3.9)와 같다.

$$a_{h,X} = \frac{\{W_h E(Y|b_h \geq X > b_{h-1})\}^p}{\sum_{k=1}^{L-1} \{W_k E(Y|b_k \geq X > b_{k-1})\}^p} \tag{3.8}$$

$$a_{h,X} = \frac{W_h\, Var(Y|b_h \geq X > b_{h-1})^{1/2}}{\sum_{k=1}^{L-1} W_k\, Var(Y|b_k \geq X > b_{k-1})^{1/2}} \tag{3.9}$$

조건 평균과 조건 분산은 조건 적률인 W_h, Φ_h, Ψ_h 로 나타낼 질 수 있다. 따라서 n 역시 조건적률의 식이 된다. n을 최소로 하는 b_h는 $\partial n / \partial b_h = 0$ 을 만족하는 b_h인데 이 편미분식은 연

쇄법칙(chain rule)과 식(3.4), 식(3.5) 그리고 식(3.6)을 이용하면 $h<L-1$에 대해서는 식(3.10)과 같이, $h=L-1$에 대해서는 식(3.11)과 같이 나타난다.

$$\begin{aligned}\frac{\partial n}{\partial b_h} &= f(b_h)\left\{\left(\frac{\partial n}{\partial W_h}-\frac{\partial n}{\partial W_{h+1}}\right)+\left(\frac{\partial n}{\partial \Phi_h}-\frac{\partial n}{\partial \Phi_{h+1}}\right)b_h+\left(\frac{\partial n}{\partial \Psi_h}-\frac{\partial n}{\partial \Psi_{h+1}}\right)b_h^2\right\}\\ &= f(b_h)(\alpha_h+\beta_h b_h+\gamma_h b_h^2)\end{aligned} \tag{3.10}$$

$$\begin{aligned}\frac{\partial n}{\partial b_{L-1}} &= f(b_{L-1})\left\{-N+\frac{\partial n}{\partial W_{L-1}}+\frac{\partial n}{\partial \Phi_{L-1}}b_{L-1}+\frac{\partial n}{\partial \Psi_{L-1}}b_{L-1}^2\right\}\\ &= f(b_{L-1})(\alpha_{L-1}+\beta_{L-1}b_{L-1}+\gamma_{L-1}b_{L-1}^2)\end{aligned} \tag{3.11}$$

위의 두 식에서 편미분은 b_h에 대한 2차방정식에 비례한다. 그러나 b_h의 계수 또한 b_h의 함수이므로 근을 직접 구할 수 없고, Sethi(1963)의 반복과정을 이용하여 방정식의 근을 구한다.

단계1 : $b_1' < b_2' < \ldots < b_{L-1}'$와 같은 임의의 경계에서 시작한다.

단계2 : 단계1에서 정해진 b_h에 대해 식(3.4), 식(3.5) 그리고 식(3.6)을 이용하여 W_h, Φ_h, Ψ_h와 n에 대한 편미분값을 계산한다.

단계3 : 단계2에서 계산된 값을 식(3.12)에 대입하여 새로운 층 경계를 구한다.

$$b_h'' = \frac{-\beta_h' + (\beta_h'^2 - 4\alpha_h'\gamma_h')^{1/2}}{2\alpha_h'} \tag{3.12}$$

단계4 : 연속적인 두 실행에서 얻어진 결과가 같거나 그 차이가 매우 작아질 때까지 단계2와 단계3을 반복한다.

$$\max|b_h'' - b_h'| < \varepsilon$$

여기서 Sethi의 알고리즘을 실행할 때 W_h, Φ_h, Ψ_h에 대한 편미분 값들은 층 h 내에서 x의 0, 1, 2차 적률에 의존하는데, 이는 X에 대한 밀도함수 $f(x)$를 알아야 계산할 수 있다. 그러나 실제 조사에 있어서 층화변수에 대한 밀도함수를 알 수 있는 경우는 드물다. Cochran(1977)이 제시한 바와 같이, 무한모집단에 대한 조건부적률인 (3.4), (3.5), (3.6)은 유한모집단에서 모집단의 N개의 x값들을 사용하여 각각 (3.13), (3.14), (3.15)로 계산될 수 있다.

$$W_h = N_h / N \tag{3.13}$$

$$\Psi_h = \frac{1}{N} \sum_{i:\, b_{h-1} < x_i \le b_h} x_i^2 \tag{3.14}$$

$$\Phi_h = \frac{1}{N} \sum_{i:\, b_{h-1} < x_i \le b_h} x_i \tag{3.15}$$

Lavallée-Hidiroglou 알고리즘에서는 $X = Y$ 임을 가정하므로, 조건부 적률 W_h, Φ_h, Ψ_h를 사용하여 나타내면 조건부평균은 $E(Y | b_h \ge X > b_{h-1}) = \Phi_h / W_h$, 조건부분산은 $Var(Y | b_h \ge X > b_{h-1}) = \Psi_h / W_h - (\Phi_h / W_h)^2$ 이다.

파워 배분에서의 $a_{h,X}$은 조건 적률을 이용하면 $a_{h,X} = \Phi_h^p / \sum_{k=1}^{L-1} \Phi_k^p$ 이고, 이를 식(3.7)에 대입하면 다음과 같다.

$$n = NW_L + \frac{\sum_{h=1}^{L-1} \Phi_h^p \sum_{h=1}^{L-1} (W_h \Psi_h - \Phi_h^2) / \Phi_h^p}{\overline{X}^2 c^2 + \sum_{h=1}^{L-1} (\Psi_h - \Phi_h^2 / W_h) / N} \tag{3.16}$$

네이만 배분하에서의 $a_{h,X}$은 조건 적률을 이용하여 다음과 같이 쓸 수 있다.

$$a_{h,X} = \frac{(\Psi_h W_h - \Phi_h^2)^{1/2}}{\sum_{k=1}^{L-1} (\Psi_k W_k - \Phi_k^2)^{1/2}}$$

따라서 표본크기 n은 식(3.17)과 같다.

$$n = NW_L + \frac{\left\{ \sum_{h=1}^{L-1} (W_h \Psi_h - \Phi_h^2)^{1/2} \right\}^2}{\overline{X}^2 c^2 + \sum_{h=1}^{L-1} (\Psi_h - \Phi_h^2 / W_h) / N} \tag{3.17}$$

3.2 층화 알고리즘의 일반화

$X = Y$를 가정하고 있는 L-H알고리즘에 stratum jumper를 나타내는 모형인 승법모형과 확률복원모형을 적용하여 층화알고리즘을 일반화한다. 모형의 적용은 2장의 각 모형에서 Y의 조건평균과 조건 분산을 계산하여 이를 Y의 평균과 분산 대신 대입하는 방법으로 이루어진다.

3.2.1 승법모형

2.2.1의 승법모형에 대해 다음과 같이 상수 $C_{\varepsilon,M}$을 정의한다.

$$C_{\varepsilon,M}=1+(M-1)^2\varepsilon(1-\varepsilon)/\{1+(M-1)\varepsilon\}^2$$

파워 배분의 경우 식(3.2)를 조건 적률을 사용하여 나타내면 $a_{h,X}=\phi_h^p/\sum_{k=1}^{L-1}\phi_k^p$ 이다. 이를 이용하여 n을 구하면 식(3.18)과 같다.

$$n=NW_L+\frac{\sum_{h=1}^{L-1}\phi_h^p\sum_{h=1}^{L-1}(C_{\varepsilon,M}W_h\Psi_h-\phi_h^2)/\phi_h^p}{\overline{X}^2c^2+\sum_{h=1}^{L-1}(C_{\varepsilon,M}\Psi_h-\phi_h^2/W_h)/N} \tag{3.18}$$

네이만 배분의 경우 다음과 같은 $a_{h,X}$ 를 얻을 수 있다.

$$a_{h,X}=\frac{(C_{\varepsilon,M}W_h\Psi_h-\phi_h^2)^{1/2}}{\sum_{k=1}^{L-1}(C_{\varepsilon,M}W_k\Psi_k-\phi_k^2)^{1/2}}$$

이를 식(3.7)에 대입하면 n은 다음과 같다.

$$n=NW_L+\frac{\left\{\sum_{h=1}^{L-1}(C_{\varepsilon,M}W_h\Psi_h-\phi_h^2)^{1/2}\right\}^2}{\overline{X}^2c^2+\sum_{h=1}^{L-1}(C_{\varepsilon,M}\Psi_h-\phi_h^2/W_h)/N} \tag{3.19}$$

3.2.2 확률복원모형

확률복원모형하에서 Y의 조건 평균과 분산을 이용하여 t_h, m_h를 다음과 같이 정의한다.

$$t_h=W_h^2Var(Y|b_h\geq X>b_{h-1})=(1-\varepsilon)W_h\Psi_h+W_h^2\varepsilon E(X^2)-\{(1-\varepsilon)\phi_h+\varepsilon W_h\overline{X}\}^2$$

$$m_h=W_hE(Y|b_h\geq X>b_{h-1})=(1-\varepsilon)\phi_h+W_h\varepsilon E(X)$$

m_h, t_h 와 조건 적률을 이용하여 파워 배분에서의 $a_{h,X}$와 표본크기 n을 계산하면 다음과

같다.

$$a_{h,X} = m_h^p / \sum_{k=1}^{L-1} m_k^p$$

$$n = NW_L + \frac{\sum_{h=1}^{L-1} m_h^p \sum_{h=1}^{L-1} (t_h / m_h^p)}{\overline{X}^2 c^2 + \sum_{h=1}^{L-1} (t_h / W_h)/N} \tag{3.20}$$

네이만 배분하에서 $a_{h,X} = t_h^{1/2} / \sum_{k=1}^{L-1} t_k^{1/2}$이고 이를 식(3.7)에 대입하면 n을 구할 수 있다.

$$n = NW_L + \frac{\left(\sum_{h=1}^{L-1} t_h^{1/2}\right)^2}{\overline{X}^2 c^2 + \sum_{h=1}^{L-1} (t_h / W_h)/N} \tag{3.21}$$

4. 사례연구

본 논문에서는 층화알고리즘을 이용하여 코스닥 등록기업 중 벤처종목의 자산평균을 추정하고자 한다. 벤처기업은 일반기업에 비해 시장 상황에 민감하게 반응하기 때문에 기업의 생성과 소멸이 빈번하고 자산, 매출, 부채 등의 재무지표들의 변동이 크게 나타난다. 코스닥등록 벤처종목 중 12월 결산인 345개 기업(2001년 12월 기준 등록기업)의 2001년과 2002년 자산을 분석대상으로 하였다. 2001년의 자산이 층화변수, 2002년의 자산이 조사변수로서 2001년의 자산으로 층화표본설계한 후 2002년의 자산을 표본추출하여 자산평균을 추정한다. 층화변수는 조사변수와 높은 상관관계를 가지고 있어야 효율적인 조사 결과를 얻을 수 있는데, 두 변수의 상관계수는 0.911로 높게 나타난다.

알고리즘의 구현은 SAS IML과 MACRO를 이용한다. 층화알고리즘에서는 초기의 층경계점이 마지막 얻어지는 층경계에 큰 영향을 준다. 여기서는 각 층이 같은 수의 단위를 갖도록 초기의 층경계점을 결정하였다. 그리고 Sethi의 반복과정에서는 연속된 실행에서 표본크기의 차이가 0.1보다 작거나 반복이 40회가 넘으면 실행을 중지하도록 하였다.

4.1 알고리즘의 실행 과정

L-H알고리즘은 층수, 목표변동계수, 파워 배분의 p값을 변수로 갖는다. 이 알고리즘의 실행과정을 알아보기 위해 층수는 5, 목표변동계수는 0.05로 하고, p값이 0.5인 파워 배분을 사용하였다. 알고리즘은 33번 반복을 실행하여 최소표본크기 19를 얻었다. 마지막 실행단계에서의 표본배분의 가중값 a_h는 <표 1>과 같다. 5개의 층 중에서 층5는 전수층이므로 a_h는 4개의 표본층에 대해서만 계산되었다. 또한 Sethi의 반복과정을 위해 표본크기에 대한 편미분값 $\partial n/\partial W_h$, $\partial n/\partial \Phi_h$, $\partial n/\partial \Psi_h$을 계산한 결과는 <표 1>과 같다.

<표 1> 각 층의 편미분 값(L-H 알고리즘)

층	a_h	$\partial n/\partial W_h$	$\partial n/\partial \Phi_h$	$\partial n/\partial \Psi_h$
1	0.2069	111.1970	-0.0155	0.000000582
2	0.2653	280.3710	-0.0195	0.000000353
3	0.3068	600.9280	-0.0221	0.000000209
4	0.2210	757.7190	-0.0146	0.000000074
5		887.3798	-0.0081	0.000000020

이 알고리즘에 의해 얻어진 층화표본설계가 <표 2>에 나타나 있다.

<표 2> L-H알고리즘으로 얻어진 층화설계(파워배분(p=0.5), 목표CV=0.05)

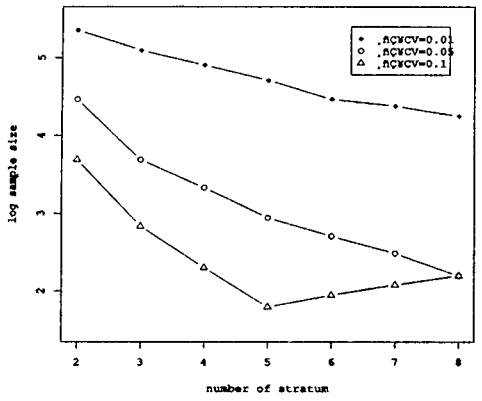
층	층경계	평균	분산	모집단크기	표본크기	추출율	표본가중값
1	19,801	13,259	15,160,883	136	3	0.022	45.33
2	39,109	27,711	26,729,648	107	3	0.028	35.67
3	72,076	52,888	72,488,630	75	4	0.053	18.75
4	161,490	98,006	529,425,480	21	3	0.143	7.00
5	270,350	205,497	1.13×10^9	6	6	1.000	1.00
				345	19		

모닷텔, KH바텍, 한국통신 데이타의 세 기업은 실제로는 층3에 속해야 하지만, 2001년에 자산규모가 작았기 때문에, <표 2>의 층화표본설계에 의해 층1에 속하여 실제의 표본가중값 18.75보다 2.42배 더 큰 45.3의 표본가중값을 갖게 되었다. 이 층화설계로 얻어진 추정량의 변동계수는 0.0924로 목표변동계수 0.05에 비해 상당히 크게 나타나서 추정량이 stratum jumper의 영향을 받고 있음을 알 수 있다.

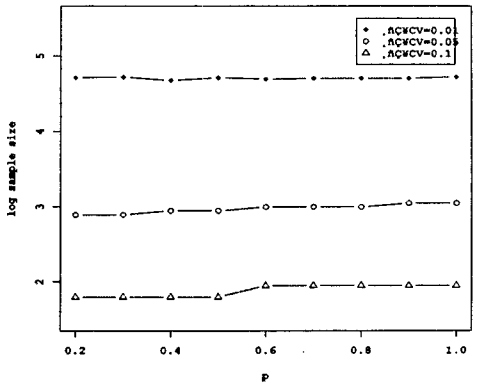
4.2 각 변수 변화에 따른 표본크기의 변화

4.2.1 L-H 알고리즘

Lavallée-Hidiroglou 알고리즘은 층수, 목표변동계수, 파워 배분의 p값을 변수로 갖는다. 층수의 영향이 [그림 1]에, 파워 배분의 p값의 영향이 [그림 2]에 나타나 있다. 각 변수값에 대해 목표변동계수를 0.01, 0.05, 0.1로 변화시켰다. 각 목표변동계수에 따른 표본크기의 차이가 크므로 이를 한 그래프에 나타났을 때 변수값의 변화에 따른 표본크기의 변화 추이가 왜곡될 수 있다. 따라서 각 그래프에서는 표본크기를 로그변환하여 나타내었다.



[그림 1] 층수에 따른 표본크기



[그림 2] p값에 따른 표본크기

[그림 1]을 보면 표본크기는 층수가 증가함에 따라 작아지는 양상을 보인다. p값이 0.5인 파워배분을 사용하고, 층수를 2에서 8까지 증가시키면서 표본크기의 변화를 살펴보았다. 층수가 2에서 8까지 증가함에 따라 표본크기는 목표변동계수가 0.01인 경우 212에서 70으로, 목표변동계수가 0.05일 때 87에서 9로, 목표변동계수가 0.1인 경우 40에서 9로 감소하였다. 목표변동계수가 0.1인 경우에는 층수가 5 이상이 되자 표본크기는 오히려 증가하였다.

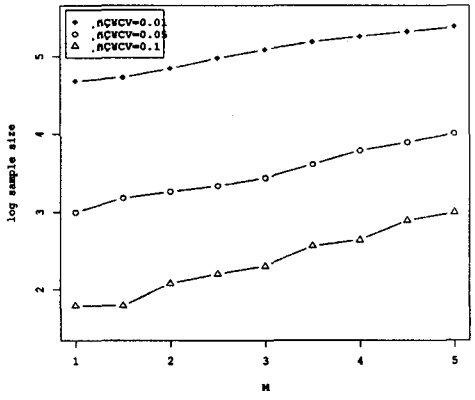
[그림 2]는 p값 변화에 따른 표본크기를 보여준다. 층수를 5로 고정시킨 상태에서 p값을 0.2에서부터 1.0까지 0.1씩 증가시키면서 표본크기를 살펴보았다. 0.01, 0.05, 0.1의 각 목표변동계수 수준에서 표본크기는 p값이 0.1에서 1.0로 증가함에 따라서 크게 변화하지 않았다.

4.2.2 승법모형 적용한 L-H알고리즘

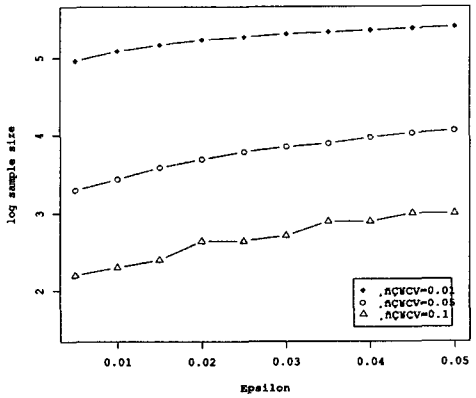
승법모형을 적용한 L-H알고리즘은 M과 ε을 변수로 갖는다.

[그림 3]은 M값의 변화에 따른 표본크기를 나타낸다. 네이만 배분을 사용하고, 층수를 5, ε을 0.01로 고정시킨 상태에서 M을 1부터 5까지 0.5씩 증가시키면서 표본크기의 변화를 살펴보았다. M이 증가함에 따라 표본크기는 증가하였다. 표본크기는 M이 1에서 5로 증가함에 따라 목표변동계수가 0.01일 때는 108에서 218로, 목표변동계수가 0.05일 때에는 20에서 69로, 목표변동계수가 0.1일 때 6에서 25로 증가하였다.

[그림 4]는 ε의 변화에 따른 표본크기의 변화를 나타낸다. 네이만 배분을 사용하고, 층수를 5, M을 3으로 고정시킨 뒤 ε를 0.005에서 0.05까지 0.005씩 변화시키면서 표본크기를 살펴보았다. ε가 증가함에 따라 표본크기는 증가하였다. 표본크기는 ε이 0.005에서 0.05로 증가함에 따라 목표변동계수가 0.01일 때는 143에서 221로, 목표변동계수가 0.05일 때는 27에서 58로, 목표변동계수가 0.1일 때는 9에서 20으로 증가하였다.



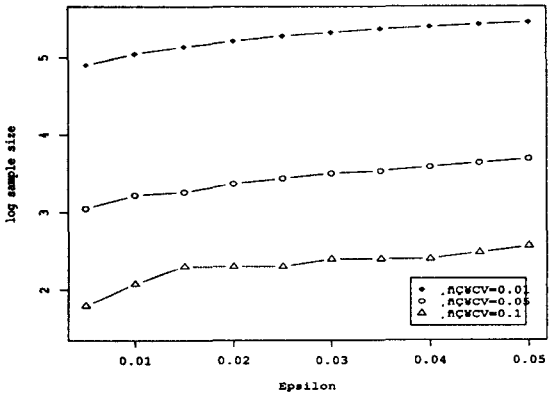
[그림 3] M에 따른 표본크기



[그림 4] ε에 따른 표본크기

4.2.3 확률복원모형 적용한 L-H알고리즘

ε값의 변화에 따라 표본크기가 어떻게 변화하는지 네이만 배분으로 층수를 5로 고정시키고 ε를 0.005에서 0.05까지 0.005씩 증가시키면서 표본크기의 변화를 살펴보았다.



[그림 5] ε에 따른 표본크기

각 ε에 대해 목표변동계수를 0.01, 0.05, 0.1로 변화시켰다. ε이 커짐에 따라 표본크기는 증가하였다. 즉, 표본크기는 ε이 0.01에서 0.05로 증가함에 따라 목표변동계수가 0.01일 때는 135에서 234로, 목표변동계수가 0.05일 때에는 21에서 40으로, 목표변동계수가 0.1일 때에는 6에서 13으로 증가하였다.

4.3 알고리즘의 비교

층화알고리즘에서 stratum jumper의 발생을 고려할 수 있는 세 가지 방법이 있다.

- 방법1 : 작은 목표변동계수를 사용한 L-H 알고리즘 사용
- 방법2 : 승법모형을 적용한 L-H 알고리즘 사용
- 방법3 : 확률복원모형을 적용한 L-H 알고리즘 사용

표본크기가 증가함에 따라 추정량의 변동계수는 작아지므로 단지 추정량의 변동계수만을 가지고 알고리즘의 효율성을 비교하는 것은 적절치 않다. 같은 표본크기에서 가장 작은 변동계수를 갖거나, 같은 변동계수 수준에서 표본크기가 가장 작은 알고리즘이 효율적인 알고리즘이다. 이 절에서는 위의 세 방법을 이용하여 같은 표본크기를 갖도록 한 뒤 자산평균의 변동계수에 대한 비교를 하겠다.

목표변동계수를 0.045로 한 L-H알고리즘, 목표변동계수를 0.07로 한 승법모형을 적용한 L-H 알고리즘, 목표변동계수를 0.07로 한 확률복원모형을 적용한 L-H알고리즘의 실행 결과를 비교하도록 하겠다. 승법모형의 경우 M=3, ε=0.022, 확률복원모형의 경우 ε=0.057로 하였다. 층수를 5로 하였고, 네이만 배분을 사용하였다.

세 방법을 통해 표본크기는 11에서 25로 증가하였다. 일단 각 알고리즘의 가장 큰 차이는 전수층의 크기이다. 낮은 목표변동계수를 사용한 L-H알고리즘의 경우 전수층의 크기가 13이었는데 반해, 모형을 적용한 L-H알고리즘의 경우 전수층의 크기가 둘 다 6이었다. 이는 모형을 적용한 층화알고리즘이 표본설계에서 전수층의 상대적인 중요성을 낮춤을 의미한다.

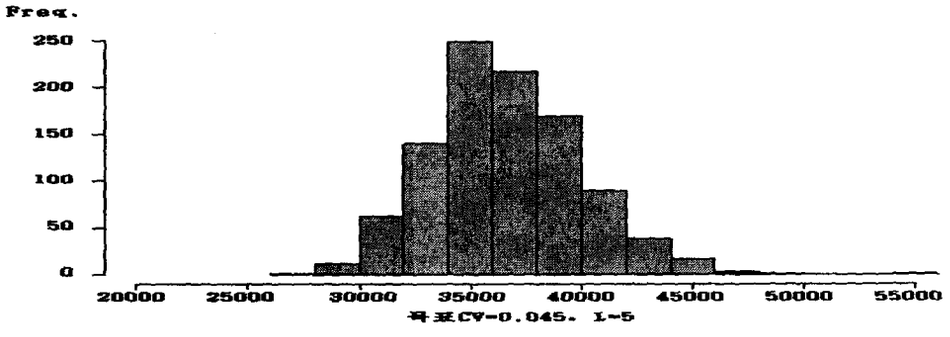
다음으로 최소 표본추출율을 비교해보면 확률복원모형을 사용한 경우 최소 표본추출율이 0.047로, 낮은 목표변동계수를 사용한 L-H알고리즘의 0.021과 승법모형을 적용한 L-H알고리즘의 0.029보다 훨씬 크게 나타났다.

<표 3> stratum jumper를 고려하지 않는 알고리즘과 고려한 알고리즘의 비교

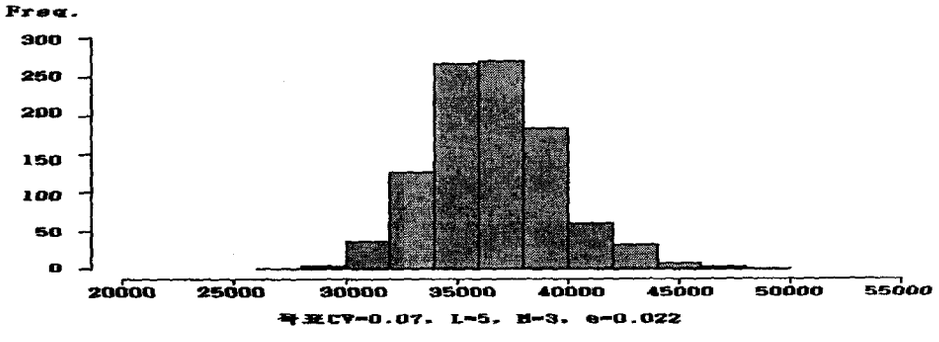
	알고리즘	표본크기	목표CV	CV	최소 표본추출율	최대 표본가중값	전수층 크기
stratum jumper 고려안함	L-H알고리즘	11	0.070	0.1275	0.014	70.00	2
stratum jumper 고려	L-H알고리즘	25	0.045	0.0916	0.021	48.00	13
	승법모형 적용한 L-H알고리즘	25	0.070	0.0782	0.029	34.75	6
	확률복원모형 적용한 L-H알고리즘	25	0.070	0.0696	0.047	21.40	6

이것은 확률복원모형을 적용한 L-H알고리즘을 실행하였을 때 최대 표본가중값이 다른 두 모형보다 상당히 작음을 의미한다. 따라서 확률복원모형을 적용한 L-H알고리즘을 사용하면 층화시점과 조사시점 사이에 단위의 층이 바뀌어도 표본가중값의 변화가 크지 않으므로 stratum jumper의 조사 추정값에 대한 영향은 상당히 줄어들 수 있다. 각각의 표본설계를 이용하여 추정된 2002년 자산평균의 분포는 [그림 6], [그림 7], [그림 8]과 같다. 자산평균에 대한 변동계수는 각각 0.0916, 0.0782, 0.0696으로 stratum jumper의 발생을 고려하지 않았던 경우의 0.1275보다는 목표변동계수

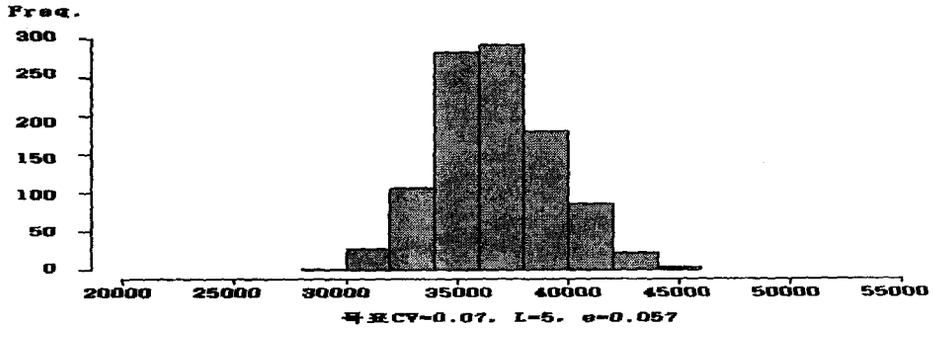
에 근접하게 나타났고, 확률복원모형을 적용한 L-H의 알고리즘의 경우 목표변동계수인 0.07보다 작은 값으로 나와 추정량의 효율이 가장 높다고 할 수 있다.



[그림 6] 2002년 자산평균의 분포(L-H알고리즘 ; 네이만배분)



[그림 7] 2002년 자산평균의 분포(승법모형 적용한 L-H알고리즘)



[그림 8] 2002년 자산평균의 분포(확률복원모형 적용한 L-H알고리즘)

5. 결 론

본 논문에서는 stratum jumper의 문제를 설계단계에서 해결할 수 있는 방법을 제시하였다.

Lavallée-Hidiroglou 알고리즘은 왜도가 있는 모집단에 대하여 주어진 정도조건과 표본배분방법 하에서 총 표본크기를 최소화시키는 층 경계와 표본크기를 얻기 위해 사용되는데, 이 알고리즘은 층화변수와 조사변수가 같음을 가정한다. 이 L-H알고리즘에 stratum jumper를 나타내는 두 모형인 승법모형과 확률복원모형을 적용하였다.

각 조건을 변화시키면서 알고리즘을 실행했을 때 stratum jumper를 설명하기 위해 표본크기가 적정한 비율로 증가했고, 표본크기가 증가함에 따라 조사추정량의 변동계수는 목표변동계수에 가까워졌다. 표본크기에 대한 각 알고리즘의 변수의 영향을 살펴보면 L-H알고리즘 실행에서 층수와 목표변동계수가 증가함에 따라 표본크기는 감소하였고, 파워 배분의 p값은 표본크기에 큰 변화를 주지 않았다. 승법모형을 적용한 L-H알고리즘 실행에서는 층화시점에 비해 조사단위가 몇 배 증가하였는지를 나타내는 변수인 M과 그러한 단위가 발생할 확률인 ε값이 증가함에 따라 표본크기는 증가하였고, 확률복원모형을 적용한 L-H 알고리즘 실행에서는 층화변수와 조사변수의 값이 다를 확률인 ε값이 커짐에 따라 표본크기는 증가하였다. 목표변동계수가 다른 변수에 비해 표본크기에 큰 영향을 주었다.

다음으로 세 가지 알고리즘을 비교하였다. stratum jumper에 대한 해결 방법은 낮은 목표변동계수를 사용한 L-H알고리즘, 승법모형을 적용한 L-H알고리즘, 확률복원모형을 적용한 L-H알고리즘이 있는데, 이를 통해 각 층화설계에서 같은 표본크기를 갖도록 한 뒤 비교를 하였다. 모형을 적용한 두 방법은 L-H알고리즘을 사용한 경우보다 전수층의 표본크기가 더 작아서 전수층의 상대적인 중요성을 낮추는 것으로 보인다. 그리고 확률복원모형을 적용한 L-H알고리즘을 사용한 경우 나머지 두 방법보다 최소 표본추출율이 훨씬 크게 나타났는데, 이는 최대 표본가중값이 작아 stratum jumper가 발생하더라도 조사추정량에 상대적으로 작은 영향을 미친다는 것을 의미한다. 세 방법 모두 stratum jumper를 고려하지 않은 경우보다 낮은 추정량의 변동계수를 얻을 수 있었는데, 확률복원모형을 적용한 L-H 알고리즘을 사용했을 때 목표변동계수보다 더 작은 추정량의 변동계수를 얻을 수 있었고, 분산이 작아서 가장 효율적이라고 판단하였다.

stratum jumper의 문제는 많은 표본조사에서 실제로 빈번하게 발생하는 문제이며 다루기가 쉽지 않다. 여기서 제시한 배분방법과 층화알고리즘의 비교가 모든 조사에 일관되게 적용되지는 않는다. 실제 조사에 있어서는 조사모집단의 성격과 조사목적, 조사여건을 고려하여 배분방법과 층화알고리즘을 사용하면 표본설계단계에서 효과적으로 stratum jumper의 문제를 해결할 수 있다.

참고문헌

[1] Bankier, M. D. (1988). Power Allocations : Determining Sample Sizes for Sub-National Areas. the American Statistician. 42, 174-177.

[2] Cochran, W. G. (1977). *Sampling Techniques*. Third Edition. New York; John Wiley & Sons, Inc.

[3] Hidiroglou, M (1994). Sampling and Estimation for Establishment Surveys: Stumbling Blocks and Progress. *Proceedings of the Section on Survey Research Methods*, American Statistical Association. 153-162.

[4] Hidiroglou, M.A., and Srinath, K.P.(1993). Problems Associated with Designing Subannual Business Surveys. *Journal of Business and Economics Statistics.* 11, 397-405.

[5] Lavallée, P., and Hidiroglou, M. (1988). On the Stratification of Skewed Populations. *Survey Methodology, 14*, 33-43

[6] Lee, H. (1995). Outliers in Business Surveys. *in Business Survey Methods* edited by B. G. Cox, D. A. Binder, B. N. Chinnappa, A. Christianson, M. Colledge, and P. S. Kott, 503-526

[7] Louis-Paul Rivest (1999). Stratum Jumpers; Can We Avoid Them? *Proceedings of the Section on Survey Research Methods,* American Statistical Assocition. 64-72.

[8] Louis-Paul Rivest (2002). A Generalization of Lavallée & Hidiroglou Algorithm for Stratification in Business Surveys. *Survey Methodology,* Vol.28, No2, 191-198.

[9] Louis-Paul Rivest, http://www.mat.ulaval.ca/pages/pr

[10] Särndal, C. -E., Swensson, B. & Wretman, J. H. (1992), *Model Assisted Survey Sampling*, New York, Springer-Verlag.

[11] Sethi, V. K.(1963). A note on the optimum stratification of populations for estimating the population means, Australian Journal of Statistics. 5, 20-33.

[12] Slanta, J., and Krenzke, T. (1994). Applying the Lavallée & Hidiroglou Methods to Obtain Stratification Boundaries for the Census Bureau's Annual Capital Expenditure Survey. *Proceedings of the Section on Survey Research Methods,* American Statistical Association. 693-698.

[2004년 2월 접수, 2004년 5월 채택]