응용통계연구 제 **20**권 3호, *2007*년, *pp. 605–618*

적응집락추출에서 표본크기 결정과 추정량의 효율 비교

남궁 평[1] 원혜경[2] 최재혁[3]

요 약

모집단 단위들이 희박하게 존재하고 접근하기 어려운 경우에 적용하는 적응추출설계에서의 추출과정은 관심변수의 관측값에 의존한다. 동일한 표본크기에서 적응집락추출의 추정량은 단순임의추출의 추정량에 비해 효율이 더 좋다. 적응추출에서 Rao-blackwell의 정리를 적용하여 Murthy의 추정량의 형태로 수정한 한센-휴비ㅡ (HH) 추정량과 호르비ㅡ-톰슨 (HT) 추정량은 기존의 추정량에 비해 작은 분산을 가진다. 본 연구는 초기표본을 바꿔가면서 기대표본크기와 적응추출의 표본크기 하의 단순임의추출의 추정량과 적응추출의 추정량의 효율을 비교하였다.

주요용어: 적응표본설계, 한센-휴비ㅡ (HH) 추정량, 호르비ㅡ-톰슨 (HT) 추정량.

1. 서론

최근 모집단의 밀도, 희귀종의 수 등과 같은 자연 모집단을 추정하기 위한 많은 표본조사 방법들이 연구되고 있다. 자연 모집단에서 전통적인 추출방법을 적용한다면 많은 문제점이 발생할 수 있다. 예를 들어 동물들이 넓은 지역에 흩어져 있는 경우 평균 모집단 밀도는 낮지만 밀도가 높은 특정 집락이 존재하기 때문에 동물들이 전혀 없는 집락이 넓은 지역에 걸쳐 존재하게 된다. 이 경우 전통적 표본추출 방법을 적용하면 많은 지역에서 동물수가 관측되지 않고 관측되는 지역은 매우 좁은 지역으로 존재할 것이다. 따라서 추정량의 분산이 증가하여 정도가 감소하게 된다 (Smith 등, 1995).

이러한 추정의 비효율성을 개선하기 위해 Thompson (1990, 1995, 1996)은 적응표본추출설계 방법을 제시했다. 희귀한 종에 대한 조사에서 사용되는 이 방법은 임의의 장소에서 관측값이 존재하는 경우 인접 장소에 적어도 한 개 이상의 관측값이 존재한다고 가정하고 모집단의 밀도를 추정한다. 즉, 최초의 추출단위에서 관측값이 존재하지 않으면 새로운 추출단위를 추출하고 관측값이 존재한다면 인접단위를 추출단위로 간주하고 관측값을 측정하는 방법이다. 그러나 표본추출의 각 단계가 이전 단계의 관측값에 의존하기 때문에 추출된 단위의 총 크기가 임의추출이라는 가정과 표본추출 시행 횟수에 대한 제약 조건이 필요

1) (110-745) 서울시 종로구 명륜동 3가 53, 성균관대학교 통계학과, 교수
E-mail: namkung@skku.edu
2) (420-844) 경기도 원미구 중동 679-2, KB국민은행, 계장
E-mail: zooty2183@hanmail.net
3) (110-745) 서울시 종로구 명륜동 3가 53, 성균관대학교 통계학과, 박사과정
E-mail: leonash@skku.edu

하다. 추정에서는 전통적인 표본추출에서 사용되는 추정량을 이용하면 편의가 발생하므로 군집형태를 이루고 있는 모집단의 특성을 파악하기 위한 한센-휴비츠 (HH) 추정량과 호르비츠-톰슨 (HT) 추정량을 이용한다.

본 연구에서는 희귀한 모집단에서의 표본설계기법인 적응설계기법과 전통적인 설계기법인 단순임의추출과의 효율을 비교하기 위해 각각의 추정방법을 살펴보고 표본 크기에 따른 추정량의 효율성을 알아보고자 한다.

2. 적응표본설계

적응표본추출은 단위들의 초기집합을 선택하여 선택된 단위가 어떤 기준을 만족할 때마다 이 단위와 인접한 단위들을 표본에 추가하는 표본설계이므로 모집단에서 관찰된 유형의 변화에 영향을 크게 받는다. 추정량들은 네트워크 표본추출 같은 불균등확률 표본설계에서 이용했던 HH 추정량과 HT 추정량과 관련이 있지만 추정량들을 계산하는데 있어 필요한 포함확률을 표본의 모든 단위에서 결정할 수 없으므로 수정된 추정량들을 이용한다.

적응표본추출 설계는 다음과 같은 이웃에 대한 몇 가지의 가정을 필요로 한다 (Thompson, 1990). 단위가 주어졌을 때 동서남북으로 인접한 4개의 표본단위를 이웃이라 하면;

1. 모집단에서 모든 단위 i에 대해 인접 단위들의 집합을 이웃 A_i로 정의한다.
2. 각 단위들의 이웃은 지리적으로 가장 가까운 인접 단위의 집합으로 구성된다.
3. 다른 표본추출 상황에서는 단위들 사이의 사회적 또는 제도적 관계에 의해 정의된다.
4. 이웃 관계는 대칭적이다. 즉, 단위 j가 단위 i의 이웃이라면 단위 i도 단위 j의 이웃이다.

여기서 말하는 이웃은 프레임 안에 있는 각 표본단위와 다른 표본단위가 관련되어 있는 규칙으로 특성화할 수 있다. 이웃 단위들이 표본에 추가될 조건은 관심변수의 범위에서 구간 또는 집합 C에 의해 주어진다 ($C = \{y|y \geq c\}$). 선택된 단위가 이런 가정을 만족할 때 선택된 단위에 모든 이웃 단위들은 표본으로 추가되고 관측된다. 즉, i번째 단위의 초기선택에 의한 설계 하에서 관측되는 모든 단위들의 집합을 고려했을 때 몇 개의 이웃들의 합으로 구성될 수 있는 집합들을 집락이라고 한다. 이 경우에 집락 내의 단위들로 이루어진 부분집합을 네트워크라고 하며 표본에 포함될 조건을 만족하지는 않는 이웃 단위를 테두리 단위 (edge unit)라 한다.

이와 같이 네트워크에 속하는 어떤 임의의 단위를 선택하면 네트워크에 속하는 모든 단위와 이들에 관련된 모든 테두리 단위를 포함시킬 수 있다. 그러나 어떤 임의의 테두리 단위를 선택하는 경우 관측값이 존재하더라도 조건을 만족하지 않는다면 그 단위들을 포함시킬 수 없다. 결국 y값들이 주어졌을 때 조건을 만족하지 않는 임의의 단위를 고려하여 모집단을 네트워크로 분할한다.

3. 수정된 한센-휴비츠 (HH) 추정량과 호르비츠-톰슨 (HT) 추정량

적응집락표본설계에서 선택확률이 표본에 있는 모든 단위들에 대해서 알려져 있지 않기 때문에 다음 식 (3.1)를 이용하여 수정된 HH 추정량의 불편추정량을 계산할 수 있다 (Thompson, 1990).

$$t_{HH^*} = \frac{1}{n_1}\sum_{i=1}^{n_1}\frac{1}{m_i}\sum_{j\in A_i} y_i = \frac{1}{n_1}\sum_{i=1}^{n_1}\bar{y}_i. \tag{3.1}$$

여기서 t_{HH^*}는 모평균에 대한 HH 추정량이고 m_i는 네트워크 안에 있는 단위의 수, $\bar{y}_i$는 초기표본의 i번째 단위를 포함하는 네트워크 안에 있는 관측값들의 평균이다. 따라서 t_{HH^*}의 분산추정량은 다음과 같다.

$$\begin{aligned} \text{비복원} &: \widehat{var}[t_{HH*}] = \frac{N-n}{Nn}\sum_{i=1}^{n}\frac{(\bar{y}_i - t_{HH^*})^2}{n-1}, \\ \text{비복원} &: \widehat{var}[t_{HH^*}] = \frac{1}{n}\sum_{i=1}^{n}\frac{(\bar{y}_i - t_{HH*})^2}{n-1}. \end{aligned} \tag{3.2}$$

적응집락표본설계에서는 표본에 포함된 모든 단위들의 포함확률을 알지 못하기 때문에 불편추정량은 Thompson (1990)이 제안한 수정된 HT 추정량에 의해 계산한다. 만약 초기표본이 단순임의추출이라고 하면 포함확률은

$$\text{비복원} : \pi_i = 1 - \frac{\binom{N-m_i}{n_1}}{\binom{N}{n_1}}, \quad \text{복원} : \pi_i = 1 - \left(1 - \frac{m_i}{N}\right)^{n_1} \tag{3.3}$$

이고, 여기서 m_i는 단위 i에 포함된 네트워크 단위의 수로 어떤 단위가 조건을 만족하지 않으면 $m_i = 1$이다. 이를 이용한 수정된 모평균에 대한 HT 추정량은 다음과 같이 표현 할 수 있다.

$$t_{HT^*} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{y_i I_i}{\pi_i}\right). \tag{3.4}$$

여기서 지시변수 I_i는 표본에 있는 i번째 단위가 조건을 만족하지 않으면 0이라 하고 초기 표본에서 제외하고 그렇지 않은 경우는 $I_i = 1$이다. t_{HT^*}의 분산을 구하기 위해서 개별적인 단위보다는 분할된 모집단의 네트워크를 이용하는 방법을 사용하는 것이 가장 편리할 것이다. 추정량의 분산추정량을 구하기 위해 필요한 초기표본에서 네트워크 i와 j의 최소한 단위를 포함하는 확률 π_{ij}를 구해보면 $\pi_{ii} = \pi_i$이므로 추정량의 분산추정량은 다음

과 같다.

$$\begin{aligned} \text{비복원} : \pi_{ij} &= 1 - \frac{\binom{N-m_i}{n_1} + \binom{N-m_j}{n_1} - \binom{N-m_i-m_j}{n_1}}{\binom{N}{n_1}}, \\ \text{복원} : \pi_{ij} &= 1 - \left\{ \left[1 - \frac{m_i}{N}\right]^{n_1} + \left[1 - \frac{m_j}{N}\right]^{n_1} - \left[1 - \frac{(m_i+m_j)}{N}\right]^{n_1} \right\}, \end{aligned} \tag{3.5}$$

$$\widehat{var}\left[t_{HT^*}\right] = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j \pi_{ij}} \right) y_i y_j. \tag{3.6}$$

4. 테두리 단위를 고려한 추정량

Dryver와 Thompson (2005, 2006)은 기존의 HT, HH 추정량에 Rao-Blackwell 이론 (Blackwell, 1947)을 적용하여 Murthy의 추정량 (Murthy, 1957) 형태로 수정한 추정량을 제안했다. 제안된 추정량들은 초기 표본의 테두리 단위들을 고려한 추정량으로 계산이 매우 쉽다.

최종표본 s는 핵심 부분을 나타내는 s_c와 이를 제외한 나머지 부분을 나타내는 $s_{\bar{c}}$로 나눌 수 있다. 핵심 부분인 s_c는 임의의 조건 $y_i \geq c$를 만족하는 표본에 있는 단위들의 모든 집합을 의미한다. 이를 제외한 나머지 부분인 $s_{\bar{c}}$는 $y_i < c$를 만족하는 부분으로 구성되어 있다. 단위 i에 대해 f_i는 초기표본에 의해 교차된 단위의 네트워크 수라 한다. 즉, f_i는 초기표본에서 단위 i를 포함하는 네트워크 내의 단위들의 수이다. 표본통계량 $d^+(d^+ = \{(i, y_i, f_i) : i \in s_c, (j, y_j) : j \in s_{\bar{c}}\})$에서 교차 빈도인 f_i는 $i \in s_c$만을 포함한다. D^+는 d^+의 값을 나타내는 확률변수이고 d^+에 대한 표본공간으로 정의한다. 초기 표본은 최종 표본과 통계값 d^+의 모든 값에 의해서 결정되므로 $g(s_0{}')$를 초기 표본의 형태로 나타나는 함수라 하고 $s_0{}'$와 d^+의 값에 대하여 다음과 같이 정의된다.

$$I(s_0{}', d^+) = \begin{cases} 1, & \text{if } g(s_0{}') = d^+, \\ 0, & \text{기타.} \end{cases} \tag{4.1}$$

$L(d^+)$를 d^+와 관련된 초기 표본의 수라 하고 $P(d^+)$는$D^+ = d^+$인 확률이라고 하고 S는 모든 가능한 초기 표본을 포함하는 표본공간이라 하면 Murthy의 추정량 형태로 수정된 추정량의 분산을 다음과 같이 구할 수 있다 (Dryver와 Thompson, 2006).

$$var\left[t_{+^*}\right] = E\left[t_* \mid D^+ = d^+\right]. \tag{4.2}$$

4.1. Murthy의 추정량 형태로 수정된 한센-휴비츠 (HH) 추정량

임의의 $i \in s$에 대하여 지시변수 e_i는 다음과 같이 정의한다.

$$e_i = \begin{cases} 1, & y_i < c\text{와 i는 } j \in s_c\text{의 이웃인 경우,} \\ 0, & \text{기타.} \end{cases} \tag{4.3}$$

만약 i가 테두리 단위이고 초기표본에서 선택된 테두리 단위가 네트워크를 이룬다면 $e_i = 1$이다. 따라서 표본에 있는 테두리 단위들의 수 (e_s)와 초기 표본 s_0에서 추출된 단위들의 수 (e_{s_0})는 다음과 같다.

$$e_s = \sum_{i=1}^{\nu} e_i = \sum_{i \in s} e_i, \qquad e_{s_0} = \sum_{i=1}^{n} e_i = \sum_{i \in s_0} e_i. \tag{4.4}$$

여기서 ν는 최종 표본크기이다. 따라서 표본 안에 있는 i번째 단위에 대해 새로운 관심변수 $w_i{}'$는 다음과 같이 정의 할 수 있다.

$$w_i{}' = w_i(1 - e_i) + \bar{y}_e e_i. \tag{4.5}$$

여기서 $\bar{y}_e = \sum_{i=1}^{\nu} e_i y_i / e_s$ 는 최종표본에 있는 테두리 단위인 y값의 평균이다. 변수 $w_i{}'$는 표본의 테두리 단위가 아닐 경우 원래의 $w_i = (1/m_i)\sum_{j \in A_i} y_i$를 나타낸다. 즉, 표본 테두리 단위 $w_i{}'$는 표본 테두리 단위의 평균을 이용하여 계산한다. 따라서 Murthy의 추정량 형태로 수정된 추정량, 추정량의 분산, 비복원에서의 추정량의 분산의 불편추정량은 다음과 같다 (Dryver와 Thompson, 2006).

$$t_{HH+^*} = \frac{1}{n}\sum_{i=1}^{n} w_i{}', \tag{4.6}$$

$$\begin{aligned} var\,[t_{HH+^*}] = {} & \frac{N-n}{Nn(N-1)} \sum_{i=1}^{N} (w_i - \mu)^2 \\ & - \frac{1}{n^2} \sum_{d^+ \in D^+} \frac{P(d^+)}{L(d^+)} \sum_{s_0{}' \in S} I(s_0{}', d^+) \left(\sum_{i \in s_0{}', e_i = 1} y_i - e_{s_0} \bar{y}_e \right)^2, \end{aligned} \tag{4.7}$$

$$\widehat{var}\,[t_{HH+^*}] = \frac{N-n}{Nn(n-1)} \sum_{i=1}^{n} (w_i - \hat{\mu})^2 - \frac{1}{Ln^2} \sum_{s_0{}' \in S} I(s_0{}', d^+) \left(\sum_{i \in s_0{}', e_i = 1} y_i - e_{s_0{}'} \bar{y}_e \right)^2. \tag{4.8}$$

그러나 위의 불편추정량보다 더 효율적인 추정량은 다음과 같다.

$$\begin{aligned} \widehat{var}\,[t_{HH+^*}] &= E\,[\widehat{var}(t_{HH+^*}) \mid d^+] \\ &= \frac{1}{L} \sum_{s_0 \in S} I(s_0{}', d^+) \frac{N-n}{Nn(n-1)} \sum_{i=1}^{n} (w_i - \hat{\mu})^2 \\ &\quad - \frac{1}{Ln^2} \sum_{s_0{}' \in S} I(s_0{}', d^+) \left(\sum_{i \in s_0{}', e_i = 1} y_i - e_{s_0} \bar{y}_e \right)^2. \end{aligned} \tag{4.9}$$

4.2. Murthy의 추정량 형태로 수정된 호르비츠-톰슨 (HT) 추정량

표본에 있는 k번째 네트워크에서의 $i = 1, \ldots, K$에 대한 지시변수 $e_k{}'$ 는 개별단위가 아닌 네트워크에 의해 표시될 수 있다.

$$e_k{}' = \begin{cases} 1, & y_k{}^* < c\text{와 k는 } k' \in s_c\text{의 이웃인 경우,} \\ 0, & \text{기타.} \end{cases} \tag{4.10}$$

여기서 k번째 네트워크에서 y값의 합은 $y_k^* = \sum_{j \in A_i} y_i$이다. s_c는 하나 이상의 모든 네트워크에서 $e_k{}' = 0$이고 s안에 포함되어 있지 않다.

만약 $e_k{}' = 0$이면 $y_k{}'$는 네트워크 안에 있는 y값들의 총합이고 $e_k{}' = 1$이면 테두리 단위 $y_k{}'$는 표본 안에 있는 모든 테두리 단위의 평균이 된다.

$$y_k{}' = \begin{cases} y_k{}^*, & \text{if } e_k{}' = 0, \\ \bar{y}_e, & \text{if } e_k{}' = 1. \end{cases} \tag{4.11}$$

따라서 Murthy의 추정량 형태로 수정된 추정량, 추정량의 분산, 비복원에서의 추정량의 분산의 불편추정량은 다음과 같다 (Dryver와 Thompson, 2006).

$$t_{HT^*} = \frac{1}{N} \sum_{k=1}^{K} \frac{y_k z_k}{\alpha_k}, \tag{4.12}$$

$$\begin{aligned} var\left[t_{HT+^*}\right] = & \frac{1}{N^2} \sum_{k=1}^{K} \sum_{h=1}^{K} \frac{y_k y_h (\alpha_{kh} - \alpha_k \alpha_k)}{\alpha_k \alpha_k} \\ & - \frac{1}{n^2} \sum_{d^+ \in D^+} \frac{P(d^+)}{L(d^+)} \sum_{s_0 \in S} I(s_0, d^+) \left(\sum_{i \in s_0, e_i = 1} y_i - e_{s_0} \bar{y}_e \right)^2 . \end{aligned} \tag{4.13}$$

여기서 네트워크 k의 포함확률은 $\alpha_k = 1 - \binom{N-m_k}{n} / \binom{N}{n}$이고 네트워크 k와h 의 결합확률은 $\alpha_{kh} = 1 - \{\binom{N-m_k}{n} + \binom{N-m_h}{n} - \binom{N-m_k-m_h}{n}\} / \binom{N}{n}$이다. 또한 지시변수 z_k는 초기표본의 어떤 값이 k번째 네트워크와 교차하면 1, 아니면 0의 값을 갖는다. 따라서 분산의 불편추정량은 다음과 같다.

$$\begin{aligned} \widetilde{var}\left[t_{HT+^*}\right] = & \frac{1}{N^2} \sum_{k=1}^{K} \sum_{h=1}^{K} \frac{y_k y_h z_k z_h (\alpha_{kh} - \alpha_k \alpha_h)}{\alpha_k \alpha_h \alpha_{kh}} \\ & - \frac{1}{Ln^2} \sum_{s_0 \in S} I(s_0, d^+) \left(\sum_{i \in s_0, e_i = 1} y_i - e_{s_0} \bar{y}_e \right)^2 . \end{aligned} \tag{4.14}$$

그러나 불편추정량보다 효율적인 분산추정량은 다음과 같다.

$$\begin{aligned}\widehat{var}\,[t_{HT+^*}] &= E\,[\widehat{var}(t_{HT+^*}) \mid d^+] \\ &= \frac{1}{L}\sum_{s_0\in S} I(s_0{}', d^+)\frac{1}{N^2}\frac{y_k y_h z_k z_h(\alpha_{kh}-\alpha_k\alpha_h)}{\alpha_k\alpha_h\alpha_{kh}} \\ &\quad -\frac{1}{Ln^2}\sum_{s_0{}'\in S} I(s_0{}', d^+)\left(\sum_{i\in s_0{}', e_i=1} y_i - e_{s_0{}'}\bar{y}_e\right)^2 .\end{aligned} \tag{4.15}$$

5. 추정량의 효율성 비교

5.1. 위험 모집단에서의 HIV/AIDS

모집단이 주어진 상황에서 초기 표본크기를 증가시키며 적응집락추출과 단순임의추출의 효율성을 비교해 보았다. 모집단은 2004년 미국 50개 주의 HIV/AID (human immunodeficiency virus/Acquired Immune Deficiency Syndrome)에 대해 양성반응을 나타낸 사람들이고 총합을 추정하고자 한다. 이 자료는 WHO에 있는 보건자료에서 표본을 추출하였다. 50개의 주에 대해 감염비율이 희박한 지역과 감염비율이 높은 지역이 편중되게 분포되어 있다. 감염확률은 사회적 연관성이 높고 주위 감염확률에도 영향을 받기 때문에 모집단을 집락으로 설정하였다. 그래프를 통해 모집단을 살펴보면 표본 추출 값이 상대적으로 높게 나타난 지역은 그 주변 지역에서도 높은 값이 나타난다는 것을 볼 수 있다.

기존 연구에서 적응집락추출을 하기 위해 모집단 ($N = 20\times 15 = 300$)은 정방형 구역으로 분할 (인구 100,000명 기준)되어 있다 (Thompson, 2006). 본 연구에서는 비복원 단순임의추출에 의해 선택된 n_1단위의 초기표본을 선택하고 적응집락추출을 시행하여 t_{HH^*},t_{HH+^*} ,t_{HT^*} ,t_{HT+^*} 의 분산을 계산한다. 적응추출의 조건은 $C = \{y \mid y \geq 1\}$을 기본으로 하고 초기표본선택은 $n_1 = 10$부터 $n_1 = 200$으로 시행한다. 여기서 반복은 10,000번 시행한다.

5.2. 기대표본에 따른 추정량 비교

적응집락표본추출이 전통적인 표본설계보다 높은 효율을 갖는 추정량인지는 모집단의 특성을 포함하는 요인의 수, 조사방법에 포함된 표본크기, 비용 등을 고려해서 판단해야 하므로 일반적으로 모든 가능한 모집단의 구조에 대해 어떤 설계가 더 좋다고 단정 지을 수는 없다. 효율을 비교하고자 할 때 초기표본 크기는 총 표본크기보다 적기 때문에 단순히 효율을 비교하는 것은 의미가 없으므로 기대표본에 의한 단순임의추출과 적응집락추출의 최종 표본크기를 적용한 단순임의추출을 통해 효율을 비교하고자 한다.

5.2.1. 포함 확률을 고려한 표본 크기

n을 초기 표본크기가 n_1인 적응집락표본추출에서의 최종 표본크기라 하면 기대표본크기는 $E(n)$이다. 경상비용을 c_0, 단위당 초기추출 비용을 c_1, 적응집락추출로 추가된 단위의 단위당 비용을 c_2라 하면 총 기대비용은 $C = c_0 + c_1 n_1 + c_2[E(n) - n_1]$이다. 이 식에서 비용을 고정하여 기대 최종 표본크기를 계산할 수 있다.

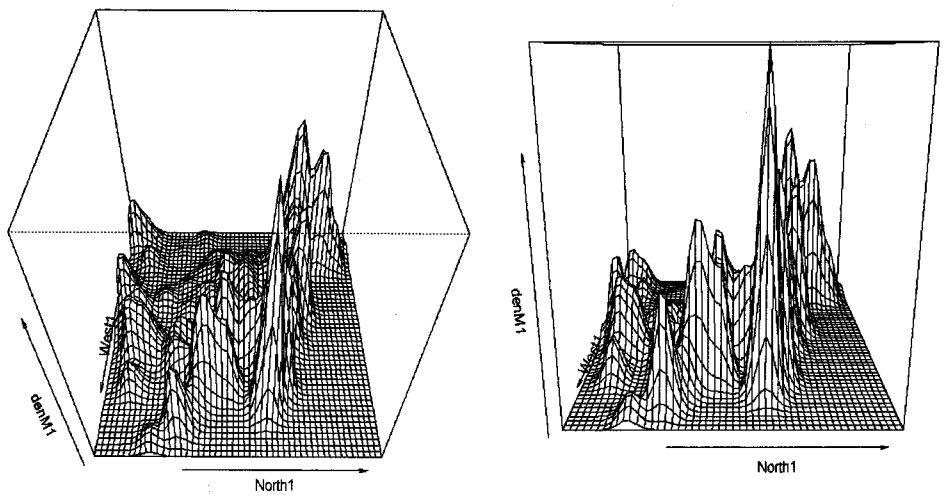

그림 5.1: 밀도에 따른 모집단

비용이 고정된 기대 최종 표본크기를 통한 단순임의추출 (기대표본)과 적응집락추출의 효율을 비교해 보았다. 여기서 $\widehat{var}(\bar{y}; srse)$는 기대표본에 의한 단순임의추출의 분산추정량이다.

수정된 HH 추정량의 경우는 기대표본에 의한 단순임의추출보다 효율이 좋지 않지만 수정된 HT 추정량의 경우는 기대표본에 의한 단순임의추출보다 효율이 더 좋았다. 초기표본이 커질수록 효율이 일관되게 좋아지는 것은 아닌데 그것은 초기표본이 크다고 해서 반드시 최종표본크기가 큰 것은 아니기 때문이다.

표 5.1: 기대표본에 의한 단순임의추출의 분산 추정량과의 효율 비교

n_1	$\widehat{var}(\bar{y}; srse)$	$\widehat{var}(t_{HT^*})$	$eff(t_{HT^*})$	$\widehat{var}(t_{HH^*})$	$eff(t_{HH^*})$
10	1.4796	0.6547	2.2600	3.6868	0.4013
20	0.3773	0.5182	0.7281	1.4451	0.2611
30	0.1808	0.2177	0.8305	0.7394	0.2445
40	0.1059	0.0828	1.2790	0.4641	0.2282
50	0.0672	0.0358	1.8771	0.3108	0.2162
60	0.0464	0.0193	2.4041	0.2236	0.2075
70	0.0349	0.0119	2.9328	0.1697	0.2057
100	0.0170	0.0041	4.1463	0.0867	0.1961
200	0.0045	0.0020	2.2500	0.0152	0.2961

5.2.2. 적응추출을 고려한 표본크기

기대표본을 적용한 단순임의추출과 비교했을 때 수정된 HH 추정량은 상대적으로 낮은 효율을 갖는다. 이번에는 기대표본을 이용하는 경우와 다르게 추출을 하는데 있어 같은 크기의 표본크기를 가지고 추정량을 비교한다. 이 방법은 적응추출을 통해 얻어진 표본크기는 유동적이기 때문에 최종 표본크기와 같은 크기의 단순임의추출의 표본으로 추정량을 얻어 비교하는 방법이다. 기대표본을 적용하는 경우 추출확률에 의해 추정량의 값이 정해진다는 장점이 있고 적응추출을 고려하는 경우는 같은 크기의 표본에 대한 효율을 비교 할 수 있는 장점이 있다. 여기서 $\widehat{var}(\bar{y}; srse)$는 적응표본설계 표본크기에 의한 추정량이다.

적응추출 표본크기를 적용한 단순임의추출의 분산추정량에 비해 수정된 HT 추정량은 상대적으로 매우 높은 효율을 보여주고 수정된 HH 추정량 역시 높은 효율을 갖는 것을 알 수 있다. 그것은 같은 크기의 표본이라도 단순임의추출의 경우는 관찰값이 없는 집락에서 추출이 동등한 확률로 이루어지기 때문에 효율이 떨어지는 것은 당연하다. 따라서 이러한 모집단의 경우는 적응집락추출이 전통적 설계보다 효율이 좋음을 알 수 있다. 또한 이 모의실험 역기 수정된 HH 추정량은 수정된 HT 추정량에 비해 효율이 좋지 않은 것으로 나타났다. 이 결과는 기대표본을 적용한 효율 비교와 같은 결과이다.

표 5.2: 적응집락추출 표본크기에 의한 단순임의추출의 분산 추정량과의 효율 비교

n_1	$\widehat{var}(\bar{y}; srsa)$	$\widehat{var}(t_{HT^*})$	$eff(t_{HT^*})$	$\widehat{var}(t_{HH^*})$	$eff(t_{HH^*})$
10	2.2882	0.6369	3.59	3.6749	0.62
20	1.5288	0.5171	2.96	1.4123	1.08
30	1.4758	0.2172	6.79	0.7536	1.96
40	1.3231	0.0825	16.03	0.4582	2.89
50	1.2724	0.0362	35.15	0.3055	4.17
60	0.2337	0.0191	64.57	0.2214	5.57
70	1.2114	0.0119	101.27	0.1677	7.22
100	1.1904	0.0041	288.88	0.0839	14.17
200	1.1299	0.0020	544.31	0.0153	73.78

5.3. 효율적 적응집락 추출을 통한 모의실험

적응집락추출에서는 조건에 따라서 선택되는 표본들이 달라지므로 특정 조건의 변화에 따라 추정량의 분산 또한 영향을 받는다. 앞 절에서의 모의실험은 모두 $C = \{y \mid y \geq 1\}$ 조건 하에서 적응집락추출을 실시하였다. 표본의 인접단위들이 표본으로 추출되기 위한 조건의 변화에 따라 분산추정량을 살펴보기 위하여 초기 조건의 값에 변화에 따른 결과를 살펴본다. 여기서 $r = 10,000$번의 반복시행 한다.

표본추출 조건의 변화에 따른 추정량을 살펴보면 $C = \{y \mid y \geq 1\}$인 경우에 비해 나머지 조건에 의한 분산추정량은 작게 나타난다. 그러나 초기 표본크기가 클수록 초기 조건에 따른 효율이 더 좋음을 알 수 있다.

단순임의추출과 HT 추정량, HH 추정량과의 분산추정량을 비교하면 조건의 변화 여부에 관계없이 HT 추정량과 HH 추정량이 효율적이다. 특히, 초기 조건이 $c = 20$으로 주어진 경우 분산추정량이 다른 조건에 비해 상당히 작아진다. 초기 표본크기가 같다는 가정 하에 임의의 조건을 변화시켜가며 효율을 높일 수 있는 방향을 정하는 것이 적응집락추출의 효율을 높이는 좋은 방법이 될 수 있다.

표 5.3: $C = \{y \mid y \geq 5\}$ 인 경우의 분산추정량 비교

	$C = \{y \mid y \geq 1\}$		$C = \{y \mid y \geq 5\}$			
n_1	$\widehat{var}(t_{HT^*})$	$\widehat{var}(t_{HH^*})$	$\widehat{var}(t_{HT^*})$	$eff(t_{HT^*})$	$\widehat{var}(t_{HH^*})$	$eff(t_{HH^*})$
10	0.6369	3.6749	0.6017	1.0585	3.5308	1.0408
20	0.5171	1.4123	0.5051	1.0238	1.4235	0.9921
30	0.2172	0.7536	0.2130	1.0197	0.7292	1.0335
40	0.0825	0.4582	0.0805	1.0248	0.4608	0.9944
50	0.0362	0.3055	0.0344	1.0523	0.3074	0.9938
60	0.0191	0.2214	0.0183	1.0437	0.2269	0.9758
70	0.0119	0.1677	0.0115	1.0348	0.1697	0.9882
100	0.0041	0.0839	0.0038	1.0789	0.0842	0.9964
200	0.0020	0.0153	0.0020	1.0000	0.0152	1.0066

표 5.4: $C = \{y \mid y \geq 10\}$ 인 경우의 분산추정량 비교

	$C = \{y \mid y \geq 1\}$		$C = \{y \mid y \geq 10\}$			
n_1	$\widehat{var}(t_{HT^*})$	$\widehat{var}(t_{HH^*})$	$\widehat{var}(t_{HT^*})$	$eff(t_{HT^*})$	$\widehat{var}(t_{HH^*})$	$eff(t_{HH^*})$
10	0.6369	3.6749	0.4934	1.2908	3.0974	1.1864
20	0.5171	1.4123	0.4366	1.1844	1.2632	1.1180
30	0.2172	0.7536	0.1905	1.1402	0.6759	1.1150
40	0.0825	0.4582	0.0748	1.1029	0.4187	1.0943
50	0.0362	0.3055	0.0305	1.1869	0.2892	1.0564
60	0.0191	0.2214	0.0152	1.2566	0.2132	1.0385
70	0.0119	0.1677	0.0094	1.2660	0.1565	1.0716
100	0.0041	0.0839	0.0030	1.3667	0.0796	1.0540
200	0.0020	0.0153	0.0017	1.1765	0.0141	1.0851

표 5.5: $C = \{y \mid y \geq 20\}$ 인 경우의 분산추정량 비교

	$C = \{y \mid y \geq 1\}$		$C = \{y \mid y \geq 20\}$			
n_1	$\widehat{var}(t_{HT^*})$	$\widehat{var}(t_{HH^*})$	$\widehat{var}(t_{HT^*})$	$eff(t_{HT^*})$	$\widehat{var}(t_{HH^*})$	$eff(t_{HH^*})$
10	0.6369	3.6749	0.3375	1.8871	2.0783	1.7682
20	0.5171	1.4123	0.3288	1.5727	0.9866	1.4315
30	0.2172	0.7536	0.1476	1.4715	0.5280	1.4273
40	0.0825	0.4582	0.0553	1.4919	0.3475	1.3186
50	0.0362	0.3055	0.0191	1.8953	0.2423	1.2608
60	0.0191	0.2214	0.0073	2.6164	0.1776	1.2466
70	0.0119	0.1677	0.0036	3.3056	0.1372	1.2223
100	0.0041	0.0839	0.0013	3.1538	0.0678	1.2375
200	0.0020	0.0153	0.0013	1.5385	0.0124	1.2339

5.4. Murthy의 추정량 형태로 수정된 추정량과의 효율성 비교

Rao-Blackwell 정리를 적용하여 Murthy의 추정량 형태로 수정된 HT 추정량과 HH 추정량은 수정된 HT 추정량과 HH 추정량보다 더 작은 분산값을 가진다는 것을 모의실험을 통해 알 수 있다.

HT 추정량에서 $n = 10$, $n = 20$인 경우에서는 추정량의 값이 조금 차이가 나지만 초기 표본크기를 증가함에 따라 추정량의 값이 비슷해져 간다. HH 추정량도 동일한 결과를 나타낸다.

Rao-Blackwell 정리를 이용하여 분산추정량을 계산한 결과 추정량이 약간 작아진다. 수

표 5.6: HT 분산추정량과 HH 분산추정량

n_1	$\widehat{var}(t_{HT^*})$	$\widehat{var}(t_{HT^{+}*})$	$\widehat{var}(t_{HH^*})$	$\widehat{var}(t_{HH^{+}*})$
10	0.4471	0.4457	1.4853	1.4752
20	0.1336	0.1326	1.4684	0.4655
30	0.1635	0.1612	0.6840	0.5605
40	0.0661	0.0552	0.4088	0.3099
50	0.0342	0.0298	0.2459	0.2385
60	0.0483	0.0191	0.1656	0.1029
70	0.0234	0.0193	0.1292	0.0759
100	0.0081	0.0025	0.1310	0.0309
200	0.0046	0.0016	0.0185	0.0076

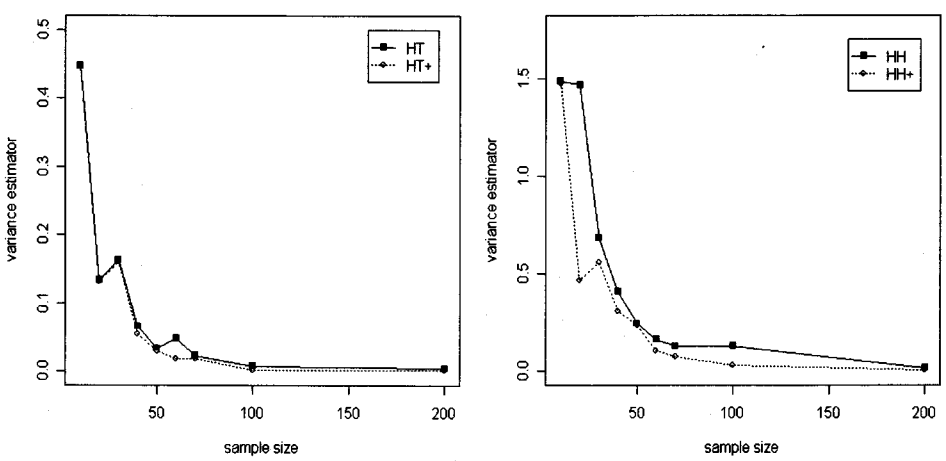

그림 5.2: HT 분산추정량과 HH 분산추정량의 효율 비교

정된 HT 추정량과 HH 추정량보다 더 낮은 분산추정량을 가지는 이유는 초기 표본 선택에서의 테두리 단위들을 고려하기 때문이다. 즉, 테두리 단위와 관련된 표본크기를 동시에 고려하면 효율을 높일 수 있다.

6. 결 론

동일한 표본크기에서 단순임의추출보다 적응표본설계기법이 더 높은 효율을 갖는 것을 살펴보았다. 그러나 일반적으로 모든 가능한 모집단의 구조에 대해 어떠한 방법이 더 좋다고 단정할 수 없다. 그 이유는 모집단의 특성을 포함하는 단위들의 수, 최종 표본크기, 비용에 따라 방법들의 효율이 달라지기 때문이다. 적응표본설계는 내부 네트워크 분산이 모집단의 분산에 근접하고 최종 표본의 일부분이 초기 표본의 일부분에 가까울 때 효율적이다. 모집단 분산과 관련 있는 내부 네트워크 분산의 중요성은 효율을 결정하는데 있어서 중요한 역할을 한다. 적응집락표본추출과 단순임의추출의 상대효율은 모집단에 대한 여러 가지의 특성, 설계와 비용에 영향을 받는다. 단순임의추출에 대한 적응집락추출이 더 효율적인 경우는 다음과 같다. 첫째, 네트워크 내 분산이 모집단 전체 분산과 비슷한 값을 갖는 경우 둘째, 희귀한 모집단인 경우 셋째, 기대 최종 표본 크기가 초기 표본 크기보다 크지 않은 경우 넷째, 연구지역에서 임의로 선택된 단위를 조사하는 것보다 집락 혹은 네트워크에서 단위를 조사하는 경우이다 (Thomson, 1990).

분산추정량을 비교해 보면 HT 추정량이 HH 추정량보다 더 높은 효율을 가지고 Murthy의 추정량 형태로 수정된 추정량을 적용하면 두 추정량 모두 초기표본이 클수록 새로운 추정량의 분산추정량이 더 작게 나타난다 ($\widehat{var}(t_{HT+^*}) \leq \widehat{var}(t_{HT^*}), \widehat{var}(t_{HH+^*}) \leq \widehat{var}(t_{HH^*})$).

따라서 초기 테두리 단위들 고려하는 것이 효율이 더 높으므로 초기 표본은 기대표본크기를 이용하고 초기 테두리 단위를 고려하여 표본추출을 하면 가장 효율적인 조사가 될 것이다. 또한 표본추출 조건의 선택에 따라 효율성이 영향을 받기 때문에 효율적인 추정량을 얻기 위해서는 초기 표본과 설계조건의 적합한 선택이 매우 중요하다.

최근 환경오염 문제와 의료 · 보건 분야의 질병에 대한 관심이 증대되면서 적응표본설계기법은 효율적인 설계기법 중에 하나일 것이다. 표본을 선택함에 있어서 최초 추출된 표본을 제외하고 나머지 추출확률을 가지고 다음 단계의 표본추출을 시행하는 설계기법을 적용한다면 더욱 효율적인 설계기반모형이 될 것이다.

참고문헌

Blackwell, D. (1947). Conditional expectation and unbiased sequential estimation, *Annals of Mathematical Statistics,* **18**, 105–110.

Dryver, A. L. and Thompson, S. K. (2005). Improved unbiased estimators in adaptive cluster sampling, *Journal of the Royal Statistical Society*, Ser. B, **67**, 157–166.

Dryver, A. L. and Thompson, S. K. (2006). Adaptive cluster sampling without replacement of clusters, *Statistical Methodology,* **3**, 35–43.

Murthy, M. N. (1957). Ordered and unordered estimators in sampling without replacement, *Sankhyã*, **18**, 379–390.

Smith, D. R., Conroy, M. J and Brakhage, D. H. (1995). Efficiency of adaptive cluster sampling for estimating density of wintering waterfowl, *Biometrics,* **51**, 777–788.

Thompson, S. K. (1990). Adaptive cluster sampling, *Journal of the American Statistical Association,* **85**, 1050–1059.

Thompson, S. K. (1995). Adaptive sampling, *The Survey Statistician,* **32**, 13–15.

Thompson, S. K. (1996). Adaptive cluster sampling based on order statistics, *Environmetrics,* **7**,123–133.

[2007년 3월 접수, 2007년 6월 채택]

Determination of Sample Size and Comparison of Efficiency in Adaptive Cluster Sampling

Pyong Namkung[1] Hyekyoung Won[2] Jaehyuk Choi[3]

ABSTRACT

Adaptive sampling design is the selection procedure which depends on observed values of the variable of interest. It is the method which could be applied to the rare and unapproachable population. Adaptive cluster sampling strategies are more efficient than simple random sampling on equivalent sample size. Adaptive sampling with new estimators through the Rao-blackwell method have lower variance than Horvitz-Thompson (HT) and Hansen-Hurwitz (HH). Also, to determine suitable sample size, it was used expected sample and the method finding appropriate sample size by changing initial sample size were studied.

Keywords: Adaptive sampling design, Horvitz-Thompson (HT) estimator, Hansen-Hurwitz (HH) estimator, initial sample size.

1) Professor, Department of Statistics, Sungkyunkwan University, Seoul 110-745, Korea
E-mail: namkung@skku.edu
2) Chief, Kookmin Bank, Gyeonggido 420-844, Korea
E-mail: zooty2183@hanmail.net
3) Ph.D. Candidate, Department of Statistics, Sungkyunkwan University, Seoul 110-745, Korea
E-mail: leonash@skku.edu