연구논문

층화추출과 계통추출을 이용한 효율적인 보조정보 사용

Efficient Use of Auxiliary Information through the Stratified Sampling and Systematic Sampling Design

김관수* · 박민규**
Kim, Gwansu · Park, Mingue

표본설계 단계에서 이용가능한 보조정보가 있는 경우 효율적인 표본추출방법으로 층화추출법이 흔히 고려된다. 특별히 층화변수로 이용할 수 있는 변수가 많은 경우 전체 층의 숫자가 커지게 되며, 이때 각 층으로부터 한 단위를 추출하는 층 표본크기가 1인 층화추출이 효율적임이 알려져 있다. 그러나 각 층으로부터 하나의 추출단위를 추출하는 층 표본크기가 1인 층화추출의 경우 불편 분산 추정량의 계산이 불가능하다. 불편 분산 추정량의 계산은 층의 수를 줄이고 각 층으로부터 두 개의 표본추출단위를 표집하는 층 표본크기가 2인 층화추출에서 가능하나 중요 층화변수가 누락될 경우 층 표본크기가 1인 층화추출에 비해 그 효율성이 떨어진다. 본 연구에서는 Park & Fuller(2008)에 의해 제시된 층 표본크기가 2인 균형층화추출과 호르비츠-톰슨 추정량의 불편 분산 추정량을 살펴보고, 모의실험을 통하여 여러 가지 층화추출법과 계통추출법을 비교한다. 또한 제시된 표본추출법을 2006년 청년패널 자료에 적용하여 그 효율성을 평가한다.

주제어: 층 표본크기가 2인 균형층화추출법, 층화추출법, 계통추출법, 보조정보.

As an efficient sampling design, stratified random sampling is often used when auxiliary information is available at the designing stage. Although one-per-stratum design is an efficient design that can be used when many auxiliary variables are available, it does not provide any unbiased variance estimator. With a two-per-stratum sample in which two elements are selected from each stratum, it is possible to obtain an unbiased variance estimator. However the loss of efficiency could be significant if any important stratification variable is missed. In this study, we investigated a sampling design that uses the all given auxiliary information and also permits an unbiased variance estimator suggested by Park and Fuller(2008). Through a simulation study, we compared several stratified random sampling and systematic sampling design. We also applied the proposed stratified sampling designs

* 고려대학교 통계학과 박사과정

** 교신저자(corresponding author): 고려대학교 통계학과 조교수 박민규.
E-mail: mpark2@korea.ac.kr

to 2007 youth panel data.

Key words: controlled two-per-stratum design, stratified random sampling design, systematic sampling design, auxiliary information.

Ⅰ. 서론

일반적으로 보조정보(auxiliary information)란 현행조사 이외의 자료로부터 얻는 모집단에 대한 정보로서, 관심변수 y의 모수를 추정하기 위해 y와 밀접한 관계를 갖는 보조변수 x가 가지고 있는 정보를 나타낸다(Sarndal et al. 1992; 김영원 외 2000). 보조정보는 관심변수 y의 모수를 추정하기 위해 사용하고, 또한 표본설계를 위한 표본추출방법에 사용할 수 있다(Sarndal et al. 1992). 관심변수 y의 모수를 추정하기 위해 보조정보가 사용되는 추정방법으로는 비추정법(ratio estimation), 회귀추정법(regression estimation), 차이추정법(difference estimation) 등이 있다. 관심변수 y의 모수를 추정함에 있어 보조정보를 사용할 경우, 관심변수 y와 보조변수 x 사이의 선형관계가 강할수록 제시한 추정량들이 표본가중치만을 이용한 호르비츠-톰슨(Horvitz-Thompson: HT) 추정량보다 효율적임이 알려져 있다(Sarndal et al. 1992).

보조정보를 이용한 표본추출방법으로는 층화추출(stratified sampling), 계통추출(systematic sampling: SYS), 확률비례추출(probability proportional sampling) 등이 있다. 이 중 보조정보를 이용한 효율적인 표본추출방법으로 층화임의추출법이 흔히 사용된다. 이용할 수 있는 층화변수가 많은 경우 가능한 층화추출법으로 층 표본크기가 1인 층화추출(one-per-stratum design: OPS)과 층 표본크기가 2인 층화추출(two-per-stratum design: TPS)등을 고려할 수 있다. OPS는 가능한 모든 보조정보를 이용한다는 관점에서 가장 효율적인 방법이지만, 선형추정량의 불편 분산 추정량을 제공하지 못한다. 이웃한 층과 결합하여 분산을 구하는 경우 분산 추정량은 실제 분산보다 과대 추정함이 알려져 있다(Cochran 1977).

OPS와는 다르게 TPS는 층의 수를 줄이고 각 층으로부터 두 개의 표본단위를 표집하므로 불편 분산 추정량이 존재한다. 하지만 주요 층화변수가 누락될 경우 효율성은 OPS에 비해 떨어진다. Park & Fuller(2008)는 TPS의 표본을 추출하는 과정에 추가적인 제약을

함으로써 TPS보다 효율적인 층 표본크기가 2인 균형표본추출(controlled two-per-stratum design: CTPS)을 제시하였다.

층화추출방법과 더불어 계통추출방법도 보조정보를 이용하는 표본추출방법으로 많이 사용된다. 표본을 추출할 때 첫 번째 원소만 임의적으로 추출하고 나머지 원소는 자동적으로 추출하는 계통추출방법은 관심변수 y와 선형관계가 있는 보조변수 x를 이용하여 모집단을 정렬시키므로 그 효율성을 높일 수 있다. 그러나 OPS와 마찬가지로 계통추출법 하에서 선형 추정량의 불편 분산 추정량이 존재하지 않는다.

본 연구에서는 주어진 보조정보를 모두 효율적으로 이용하며, 또한 선형추정량의 불편 분산 추정량을 제공하는 새로운 층화방법을 기존의 보조정보를 이용하는 표본추출법과 비교하고자 한다. 본 논문의 구성은 다음과 같다. 이어지는 2장에서는 Park & Fuller(2008)에 의해 제시된 CTPS의 결과를 요약한다. 3장에서는 가정한 모형 하에서의 CTPS와 계통추출방법의 효율성을 모의실험을 통하여 비교한다. 4장에서는 2006년 청년패널 스크린 조사자료에 여러 가지 층화추출법과 계통추출법을 적용하고 이를 비교한다.

Ⅱ. 층 표본크기가 2인 균형층화추출

층 표본크기가 1인 OPS의 효율성을 유지하면서도 불편 분산 추정량의 계산이 가능한 층화방법을 고려하기 위하여 층 표본크기가 1인 OPS와 2인 TPS를 임의로 각 층에 부여하는 방법을 고려하였다. 이와 같은 층화추출법을 보다 쉽게 설명하기 위해 4개 층으로 구성된 간단한 모집단을 고려한다. 하나의 보조변수를 이용해서 각 층 안에서 관측치를 오름차순으로 정렬하고, 각 층을 크기가 같은 두개의 그룹으로 나눈다. 그룹 1은 보조정보의 값이 작은 그룹, 그룹 2는 보조정보의 값이 큰 그룹으로 한다.

CTPS는 표본을 추출하기 위하여 4개의 층에 인식번호(identification number) 1, 2, 3 그리고 4를 임의로 부여한다. 인식번호 1이 부여된 층에서는 그룹 1에서 두 개의 추출단위를 추출하고, 인식번호 2가 부여된 층에서는 그룹 2에서 두 개의 추출단위를 추출한다. 인식번호 3과 4가 부여된 층에서는 그룹 1과 그룹 2에서 각각 하나씩의 추출단위를 추출한다. 이렇게 표본을 추출하는 방법을 〈표 1〉에 정리하였다. 4개의 층에 인식번호를 부여하는 방법은 24가지가 있으며, 부여하는 모든 방법의 확률은 1/24로 동일하다.

〈표 1〉 층 표본크기가 2인 균형층화추출의 표본 구조

인식번호 (Identification (ID) number)	그룹(Group)	
	1	2
1	XX	
2		XX
3	X	X
4	X	X

층 h의 각 그룹에는 같은 크기 m_h개의 원소들이 있으며, 따라서 층 h의 크기는 $2m_h$이다. CTPS 하에서 h층 j그룹 k번째 개체의 표본포함확률은

$$\pi_{hjk} = \Pr\{(hjk)\in A\} = \sum_{i=1}^{4}\Pr\{(hjk)\in A | ID_h = i\}\Pr\{ID_h = i\} = m_h^{-1} \tag{1}$$

이다. ID_h는 h층에 할당된 인식번호를 나타내고 A는 표본을 나타낸다. 관심변수 y의 모집단 총합에 대한 불편 추정량을 구하기 위한 가중치는 $\pi_{hjk}^{-1} = m_h$이다. 식 (1)과 동일하게 조건부 확률을 이용하는 방법으로 2차 표본포함확률을 구할 수 있다. 2차 표본포함확률은 각 개체가 속해 있는 층과 그룹에 의해 결정된다. 2차 표본포함확률은 다음과 같다.

$$\pi_{hjk,h'j'k'} = \begin{cases} m_h^{-1}, & \text{if} \quad h=h', \quad j=j', \quad k=k', \\ [2m_h(m_h-1)]^{-1}, & \text{if} \quad h=h', \quad j=j', \quad k\neq k', \\ [2m_h^2]^{-1}, & \text{if} \quad h=h', \quad j\neq j', \\ 5[6m_h m_{h'}]^{-1}, & \text{if} \quad h\neq h', \quad j=j', \\ 7[6m_h m_{h'}]^{-1}, & \text{if} \quad h\neq h', \quad j\neq j'. \end{cases} \tag{2}$$

이에 대응되는 두 개체의 공분산은 다음과 같다.

$$\begin{aligned} \delta_{hjk,h'j'k'} &= \pi_{hjk,h'j'k'} - \pi_{hjk}\pi_{h'j'k'} \\ &= \begin{cases} m_h^{-1}(1-m_h^{-1}), & \text{if} \quad h=h', \quad j=j', \quad k=k', \\ (2-m_h)[2m_h^2(m_h-1)]^{-1}, & \text{if} \quad h=h', \quad j=j', \quad k\neq k', \\ -[2m_h^2]^{-1}, & \text{if} \quad h=h', \quad j\neq j', \\ -[6m_h m_{h'}]^{-1}, & \text{if} \quad h\neq h', \quad j=j', \\ [6m_h m_{h'}]^{-1}, & \text{if} \quad h\neq h', \quad j\neq j'. \end{cases} \end{aligned} \tag{3}$$

모집단 평균 $\overline{y_N}$의 불편 추정량인 HT 추정량은 다음과 같다(Horvitz & Thompson 1952).

$$\overline{y_{HT}} = \frac{1}{N}\sum_{h=1}^{4}\sum_{j}\sum_{k}\frac{y_{hjk}}{\pi_{hjk}} = \frac{1}{N}\sum_{h=1}^{4}\sum_{j}\sum_{k} m_h y_{hjk} \tag{4}$$

여기서 y_{hjk}는 h층 j그룹 k번째 관측치이다.

HT 추정량의 분산 추정량을 유도하기 위하여 관측치를 다음과 같이 표현한다.

$$y_{hjk} = \overline{y_N} + a_{N,h} + b_{N,j} + c_{N,hj} + e_{N,hjk} \tag{5}$$

$a_{N,h} = \overline{y_{N,h..}} - \overline{y_N}$, $b_{N,j} = \overline{y_{N,.j.}} - \overline{y_N}$, $c_{N,hj} = \overline{y_{N,hj.}} - \overline{y_{N,h..}} - \overline{y_{N,.j.}} + \overline{y_N} = \overline{y_{N,hj.}} - a_{N,h} - b_{N,j} - \overline{y_N}$, $e_{N,hjk} = y_{hjk} - \overline{y_{N,hj.}}$ 이다. 정의된 a, b, c 에 의하면, $\sum_{h=1}^{4} m_h a_{N,h} = \sum_{j=1}^{2} b_{N,j} = \sum_{h=1}^{4} m_h c_{N,hj} = \sum_{j=1}^{2} c_{N,hj} = 0$ 이다. 여기서 $\overline{y_{N,h..}} = \sum_{j=1}^{2}\sum_{k=1}^{m_h} y_{hjk}/(2m_h)$, $\overline{y_{N,.j.}} = \sum_{h=1}^{4}\sum_{k=1}^{m_h} y_{hjk}/(\sum_{h=1}^{4} m_h)$, $\overline{y_{N,hj.}} = \sum_{k=1}^{m_h} y_{hjk}/(m_h)$이다. 앞으로 표기를 단순화하기 위해 아래첨자 N을 생략한다.

결과 1. (Park & Fuller 2008) CTPS 하에서 HT 추정량 $\overline{y_{HT}}$는 $\overline{y_N}$의 불편 추정량이며 $\overline{y_{HT}}$의 분산은 다음과 같다.

$$\begin{aligned} V_{CTPS}(\overline{y_{HT}}|F) = {} & N^{-2}\left\{\frac{4}{3}\sum_{h=1}^{4}\sum_{j=1}^{2} m_h^2 c_{hj}(2b_h + c_{hj})\right\} + N^{-2}\left\{\frac{1}{6}\left(\sum_{j=1}^{2} b_j^2\right)\sum_{h=1}^{4}\sum_{h'=1}^{4}(m_h - m_{h'})^2\right\} \\ & + N^{-2}\left\{\sum_{h=1}^{4}\sum_{j=1}^{2} m_h^2\left(\frac{2m_h - 3}{2m_h}\right)S_{hj}^2\right\} \end{aligned} \tag{6}$$

여기서 $S_{hj}^2 = (m_h - 1)^{-1}\sum_{k}(y_{hjk} - \overline{y_{hj.}})^2$, $F = \{y_1, \cdots, y_N\}$는 유한 모집단이다.

▶증명. 부록 A.1.

각 층의 개체수가 같은 경우 즉, $m_h = m$, HT 추정량의 분산(6)은 아래와 같다.

$$V_{CTPS}(\overline{y_{HT}}|F) = \frac{1}{64}\left\{\frac{4}{3}\sum_{h=1}^{4}\sum_{j=1}^{2} c_{hj}^2 + \frac{2m-3}{2m}\sum_{h=1}^{4}\sum_{j=1}^{2} S_{hj}^2\right\} \tag{7}$$

결과 2. (Park & Fuller 2008) 다음의 $\widehat{V}_M$은 $\overline{y_{HT}}$의 분산 $V_{CTPS}(\overline{y_{HT}}|F)$의 추정량으로서 불편 추정량이다.

$$\widehat{V}_M = N^{-2}\Big\{\widetilde{m}_1^2(1-2\widetilde{m}_1^{-1})(\widetilde{y}_{111}-\widetilde{y}_{112})^2+\widetilde{m}_2^2(1-2\widetilde{m}_2^{-1})(\widetilde{y}_{221}-\widetilde{y}_{222})^2 + [\widetilde{m}_3(\widetilde{y}_{311}-\widetilde{y}_{321})-\widetilde{m}_4(\widetilde{y}_{411}-\widetilde{y}_{421})]^2\Big\} \quad (8)$$

여기서 $\widetilde{m}_h$는 인식번호 h를 할당받은 층의 원소 개수이고, $\widetilde{y}_{hij}$는 인식번호 h를 할당받은 층 i번째 그룹 j번째 관측값이다.

▶ **증명**. 부록 A.2.

각 층의 개체수가 같은 경우, $m_h = m$, 불편 분산 추정량 (8)은 다음과 같이 표현된다.

$$\widehat{V}_M = \frac{1}{64}\Big\{(1-2\widetilde{m}^{-1})[(\widetilde{y}_{111}-\widetilde{y}_{112})^2+(\widetilde{y}_{221}-\widetilde{y}_{222})^2]+[(\widetilde{y}_{311}-\widetilde{y}_{321})-(\widetilde{y}_{411}-\widetilde{y}_{421})]^2\Big\} \quad (9)$$

기존의 HT 불편 분산 추정량과 Sen-Yates-Grandy 불편 분산 추정량이 음수의 분산 추정치를 제공할 수 있는 문제점을 갖는 데에 비해, (8)의 추정량은 항상 양의 분산 추정량을 제공한다. Park & Fuller(2008)는 기존의 분산 추정량과 제시된 분산 추정량의 효율성을 분산 추정량의 분산과 신뢰구간의 포함비율(coverage rate)을 통하여 비교하였다.

Ⅲ. 모의실험

표본추출 설계 시에 주로 한 두 개의 표본설계변수를 정하고 이 변수를 근거로 하여 효율적인 표본추출법을 유도하게 된다. 층화변수 또는 보조변수와 표본설계변수 사이의 관계를 나타내는 모형을 고려하고 각 모형에서 CTPS와 SYS를 모의실험을 통해 본 장에서 비교한다. SYS는 보조변수를 이용하여 층 내에서 모집단을 정렬한 후 고려할 수 있는 또 다른 효과적인 표본추출방법이다. 층효과가 동일한 상태에서 주어진 보조정보와 관심변수의 관계를 선형모형들을 통해 정의하고 각 관심변수별 CTPS와 SYS의 효율성을 비교했다.

모의실험을 위하여, 층 크기가 300인 4개의 층으로 구성된 3개의 모집단을 세 가지 모형하에서 생성하였다.

모형 1: $y_{hjk} = \mu + a_h + e_{hjk}, \quad e_{hjk} \sim N(0, 2^2)$.

모형 2: $y_{hjk} = \mu + a_h + x_{hjk}\beta + e_{hjk}, \quad e_{hjk} \sim N(0, 2^2), \quad x_{hjk} \sim N(2, 1), \quad \beta = 1$.

모형 3: $y_{hjk} = \mu + a_h + x_{hjk}\beta + e_{hjk}, \quad e_{hjk} \sim N(0, 2^2), \quad x_{hjk} = \left|\sin\left(\frac{\pi}{150}k\right)\right|, \quad \beta = 1$

여기에서 $\mu=0,\ a_1=-2.7,\ a_2=-2.3,\ a_3=1.0,\ a_4=4.0$ 이고, $k=1,\cdots,150$, 그룹 $j=1,2$이다. 모형 1은 보조정보 x없이 관심변수 y와 층효과만 관련된 모형이고, 모형 2와 3은 관심변수 y와 보조정보 x 그리고 층효과가 관련된 모형이다. 모형 2에서는 관심변수 y와 보조변수 x가 선형관계를 갖고 있으며, 모형 3에서도 선형관계를 나타내지만, 보조변수 x가 일정한 주기를 가지고 있어 관심변수 y도 주기를 갖게 된다.

CTPS는 4개의 층에 임의로 인식번호(<표 2.1>)를 부여하여 부여된 인식번호에 맞게 표본을 추출한다. 이렇게 추출된 표본을 가지고 모집단 평균 $\overline{y_N}$에 대한 추정량으로 (2.4)의 HT 추정량을 고려하고, $\overline{y_{HT}}$의 불편 분산 추정량으로 분산 추정량 (2.8)의 $\widehat{V}_M$을 계산한다. 이러한 과정을 독립적으로 10,000번 반복하여 HT 추정량 $\bar{y}_{HT}$과 분산 추정량 $\hat{V}_M$의 Monte Carlo 평균과 분산을 계산하였다.

SYS는 각 층에서 보조정보를 이용해 오름차순으로 정렬시킨 후 크기가 2인 계통추출 표본을 추출한다. 이렇게 추출된 표본으로써 모집단 평균 $\overline{y_N}$에 대한 추정량으로 HT 추정량을 사용한다. SYS 표본은 불편 분산 추정량이 존재하지 않으므로 SYS 표본이 단순임의추출(simple random sample; SRS)의 표본과 근사하다는 가정 하에 SRS의 분산 추정량을 사용한다(김종호 외 2006). SYS도 CTPS와 동일하게 독립적으로 10,000번 반복실험을 통해 얻은 HT추정량과 SRS분산 추정량에 대한 Monte Carlo 평균과 분산을 계산한다.

모형 1에서는 y_{hjk}를 직접 이용하여 각 층 내에서 오름차순으로 모집단을 정렬했다. 관심변수 y를 이용하여 정렬시켰기 때문에 관심변수 y는 각 층 내에서 완전한 선형성을 이루게 된다. 모형 1로부터 생성된 모집단의 평균은 -0.01이다. 모형 2와 3은 보조정보 x_{hjk}를 이용하여 각 층 내에서 오름차순으로 모집단을 정렬했다. 모형 2로부터 생성된 모집단의 평균은 2.04이고, 모형 3으로부터 생성된 모집단의 평균은 0.63이다.

<표 2>는 각각의 모형 하에서 HT 추정량과 분산 추정량의 Monte-Carlo 성질을 나타낸다. 세 가지 모형에서 CTPS와 SYS의 HT 추정량은 모집단 평균에 대한 불편 추정량임을 확인할 수 있다.

모형 1과 2에서 CTPS는 SYS보다 작은 HT 추정량의 분산을 제공하고, 모형 3에서는 HT 추정량의 분산이 비슷하다. 즉, 보조변수와 관심변수 간의 선형관계가 있는 경우 CTPS와 SYS 모두 효율적인 표본추출법으로 고려되며, 선형성이 강할수록 CTPS의 효율성이 더 높음을 알 수 있다.

모든 모형에서 새로운 분산 추정량 $\widehat{V}_M$이 불편 분산 추정량임을 확인할 수 있고, 모집단이 정렬되어 있기 때문에 SYS에서 구한 SRS 분산 추정량은 실제 분산을 과대추정하고 있음을 볼 수 있다.

〈표 2〉 HT 추정량과 분산 추정량의 Monte - Carlo 성질

모형	모집단 평균	추출방법	추정량	평균	분산
모형 1	-0.01	CTPS	$\overline{y_{HT}}$	-0.010	0.173
			$\widehat{V}_M$	0.175	0.027
		SYS	$\overline{y_{HT}}$	-0.008	0.320
			$\widehat{V_{SRS}}$	0.617	0.027
모형 2	2.04	CTPS	$\overline{y_{HT}}$	2.040	0.546
			$\widehat{V}_M$	0.546	0.225
		SYS	$\overline{y_{HT}}$	2.047	0.594
			$\widehat{V_{SRS}}$	0.693	0.237
모형 3	0.63	CTPS	$\overline{y_{HT}}$	0.632	0.483
			$\widehat{V}_M$	0.480	0.174
		SYS	$\overline{y_{HT}}$	0.619	0.476
			$\widehat{V_{SRS}}$	0.517	0.127

Ⅳ. 사례분석

본 장에서는 2007년 청년패널 조사를 위한 표본 설계에서 사용했던 2006년 청년패널 스크린 조사자료를 모집단으로 활용하여 여러 층화추출법들과 계통추출법을 비교하여 보았다. 2006년 청년패널 스크린 조사 자료는 2007년 청년패널 조사를 위한 표본 설계에서 사용했던 자료로서, 2007년 우리나라 청년층의 모집단을 대표하는 데에 한계를 갖고 있는 2001년 청년패널 표본을 2007년 우리나라의 15~29세 연령층을 대표할 수 있는 새로운 청년패널을 구축하기 위해, 2006년 산업·직업별 고용구조조사(OES 조사) 실시와 병행하여 얻어진 조사자료이다(김영원 외 2007).

본 연구를 위하여서는 그 대상을 취업자 청년으로 제한하였다. 이는 스크린 조사에서 미취업 가구원에 대해서는 성별과 연령만이 파악된 상태이고, 교육수준 등 다른 변수는 전혀 조사에 포함되어 있지 않기 때문이다. 표본설계 시 층화변수로 사용할 최종학력과 추정하고자 하는 관심변수 월평균 임금이 조사된 취업자만을 분석대상으로 삼았다. 또한 취업자 중에서도 월평균 임금이 결측값인 관측치를 제거하여 9,109개의 자료를 만들었고, 또한 월평균 임금이 300만원을 초과하는 관측치를 이상치로 판단하여 월평균 임금이 300만원을 초과하는 185개의 관측치를 제거하여 최종적으로 8,952개의 관측치를 가지고 분석하였다.

전체 모집단을 지역(16개)과 최종학력(3범주)을 이용하여 총 48개의 층으로 층화하였으며, 이 48개 층을 4개씩 묶어 하나의 부모집단을 만들어 총 12개의 부모집단을 구성하였다. 4개의 층으로 구성된 부모집단을 구성하기 위하여 지역과 학력 변수를 이용하여 유사한 4개의 층을 묶어 부모집단을 정의하였다. 각 층 내에서의 관측치의 정렬을 위하여 보조변수인 연령을 이용하여 연령이 작은 그룹 1과 연령이 큰 그룹 2로 그룹화하였다.

각 부모집단으로부터 독립된 표본들을 추출하고 이들의 합집합으로 최종표본이 결정된다. 부모집단 내의 표본추출방법으로는 OPS, TPS, CTPS 그리고 SYS를 고려한다. OPS는 층 내의 그룹을 독립된 층으로 간주하고 각 층으로부터 하나의 추출단위를 추출하여 크기가 8인 표본을 추출한다. TPS는 그룹을 무시하고 4개의 층에서 각각 2개의 추출단위를 추출한다. CTPS는 2장에서 소개한 방법으로 표본을 추출한다. SYS는 그룹 1의 크기 m_h를 추출간격으로 하여 SYS 방법으로 표본을 추출한다.

전체 평균은 12개의 부모집단으로부터 계산된 HT추정량들을 가중 평균으로 구하였다. 가중치는 각각의 부모집단의 상대크기를 사용하였다. 모집단의 월평균 임금의 평균은 134.77(단위: 천원)이며 CTPS 하에서 HT 추정량의 분산은 76.11이다. <표 3>은 세 가지의 층화임의추출 및 SYS 하에서 HT 추정량의 10,000번 반복 결과의 Monte Carlo 평균 및 분산이다. CTPS, TPS, OPS 그리고 SYS의 추출방법 하에서 HT 추정량은 모집단 평균의 불편 추정량임을 알 수 있다. TPS의 Monte Carlo 분산이 CTPS, OPS 그리고, SYS의 Monte Carlo 분산 보다 상당히 크게 나타나며 이는 TPS 설계 시 보조변수인 연령을 이용하지 않았기 때문으로 고려된다. CTPS, OPS 그리고 SYS의 Monte Carlo 분산은 비슷하다. 이는 CTPS, OPS 그리고 SYS의 효율성이 비슷함을 나타낸다. 그러나 불편분산 추정량이 존재하는 CTPS가 다른 표본추출법보다는 선호될 수 있다. CTPS 하에서의 HT추정량의 불편 분산 추정량인 $\widehat{V}_M$의 Monte Carlo 평균은 74.87로 실제 분산의 불편 추정량임을 알 수 있다.

<표 3> 각 추출방법 하에서의 HT 추정량의 Monte-Carlo 성질

추출방법	평균	분산
CTPS	134.98	77.41
TPS	134.93	92.53
OPS	134.73	78.97
SYS	134.82	79.06

Ⅴ. 결론

보조정보를 이용한 표본추출방법으로는 층화추출, 계통표집 그리고, 확률비례표집 등이 있다. 층화변수가 많은 경우, 이 중 보조정보를 효율적으로 이용하는 표본추출방법으로 기존에는 층화추출법의 OPS와 TPS가 고려될 수 있다. 본 연구에서는 OPS와 비슷한 효율성을 가지면서 또한, TPS처럼 불편 분산 추정량을 구할 수 있는 장점을 갖는 CTPS를 소개했다.

관심변수 y와 보조변수 x의 선형관계가 강한 경우, 일반적으로 좋은 표본추출방법이라 알려진 SYS보다 CTPS가 더 효율적인 표본추출방법이 될 수 있음을 모의실험을 통해 확인하였다. 즉, 관심변수 y가 보조변수 x와 강한 선형관계를 나타낼수록 CTPS가 SYS보다 효율성이 높은 것으로 고려된다.

층의 개수가 4보다 큰 일반적인 경우, 4개씩 층을 묶어 부모집단을 구성하고 부모집단에 CTPS를 적용함으로서 균형층화추출표본을 얻을 수 있을 것이다. 층의 개수가 4의 배수가 아닌 경우 상위층을 구성하고 남은 층에는 TPS를 적용하는 방법을 고려할 수 있다.

2006년 청년패널 스크린 조사자료를 모집단으로 고려한 경우, 관심변수인 월평균 임금과 보조변수인 연령과의 선형관계로 인하여 CTPS, OPS 그리고 SYS가 TPS에 비하여 효율성이 높은것을 확인하였다. 또한, CTPS, OPS 그리고 SYS의 효율성은 비슷하지만 CTPS만이 불편 분산 추정량이 존재하므로 CTPS가 OPS나 SYS보다 좋은 표본추출방법이 될 것으로 판단된다.

참고문헌

김영원·류제복·박진우·홍기학. 2000. 《표본조사의 이해와 활용》(5판). 자유아카데미.

김영원·박민규·박상현·남기성·장재호·최형아. 2007. 《2007년 청년패널 조사를 위한 표본설계》. 한국고용정보원.

김종호·정재구·류제복·김용기·김주환·홍기학·이기성·남기성·손창균·김선웅·김은주·정영미. 2006. 《표본조사입문》(2판). 자유아카데미.

Sarndal, Carl-Erik, Bengt Swensson, and Jan Wretman. 1992. *Model Assisted Survey Sampling*. New York: Springer - Verlag.

Cochran, W. G. 1977. *Sampling Technique*. New York: Wiley.

Horvitz, D. G. and Thompson, D. J. 1952. "A Generalization of Sampling without Replacement from a Finite Universe." *Journal of the American Statistical Association* 47: 663–685.

Park, M. and Fuller, W. A. 2008. "A Controlled Two-per-Sstratum Design." unpublished manuscript. Korea University.

[접수 2008/10/7, 수정 2008/11/17, 게재확정 2008/11/27]

부록

A.1. 결과1의 증명

HT 추정량의 정의에 의하여 $\overline{y_{HT}}$는 $\overline{y_N}$의 불편 추정량이다. 모평균을 추정함에 있어 HT추정량의 오차는 다음과 같이 표현된다.

$$\begin{aligned}\overline{y_{HT}}-\overline{y_N} &= N^{-1}\sum_{hjk\in A} m_h y_{hjk}-\overline{y_N}\\ &= N^{-1}\{\widetilde{m_1}(\widetilde{y_{111}}+\widetilde{y_{112}})+\widetilde{m_2}(\widetilde{y_{221}}+\widetilde{y_{222}})\\ &\quad +\widetilde{m_3}(\widetilde{y_{311}}+\widetilde{y_{321}})+\widetilde{m_4}(\widetilde{y_{411}}+\widetilde{y_{421}})\}-\overline{y_N}\\ &= N^{-1}\{\widetilde{m_1}[(\widetilde{b_1}+\widetilde{c_{11}})-(\widetilde{b_2}+\widetilde{c_{12}})]-\widetilde{m_2}[(\widetilde{b_1}+\widetilde{c_{21}})-(\widetilde{b_2}+\widetilde{c_{22}})]\\ &\quad +\widetilde{m_1}(\widetilde{e_{111}}+\widetilde{e_{112}})+\widetilde{m_2}(\widetilde{e_{221}}+\widetilde{e_{222}})+\sum_{h=3}^{4}\sum_{j=1}^{2}\widetilde{m_h}\widetilde{e_{hj1}}\Big\}\end{aligned} \tag{A.1}$$

따라서 HT 추정량의 제곱오차의 조건부 기댓값은

$$\begin{aligned}E\{(\overline{y_{HT}}-\overline{y_N})^2|F,ID\} &= N^{-2}\{\widetilde{m_1}[(\widetilde{b_1}+\widetilde{c_{11}})-(\widetilde{b_2}+\widetilde{c_{12}})]-\widetilde{m_2}[(\widetilde{b_1}+\widetilde{c_{21}})-(\widetilde{b_2}+\widetilde{c_{22}})]\}^2\\ &\quad +N^{-2}\Big\{\sum_{l=1}^{2}2\widetilde{m_l^2}(1-2\widetilde{m_l^{-1}})\widetilde{S_{ll}^2}+\sum_{h=3}^{4}\sum_{j=1}^{2}\widetilde{m_h^2}(1-\widetilde{m_h^{-1}})\widetilde{S_{hj}^2}\Big\}\end{aligned} \tag{A.2}$$

이고, 여기서 $\widetilde{S_{hj}^2}=\frac{1}{\widetilde{m_h}-1}\sum_{k=1}^{\widetilde{m_h}}(\widetilde{y_{hjk}}-\widetilde{\overline{y_{hj.}}})^2$, $\widetilde{\overline{y_{hj.}}}=\frac{1}{\widetilde{m_h}}\sum_{k=1}^{\widetilde{m_h}}\widetilde{y_{hjk}}$ 이다.

조건부 분산은 다음의 두 가지 사실로부터 구할 수 있다. 첫 번째, b와 c는 선택된 층에 의존한다. 즉, 같은 층에서는 원소가 바뀌어도 b와 c는 변하지 않는다. 두 번째, 각 층 내에서 표본추출은 단순임의추출법이 사용되었다.

식 (A.2)의 두 번째 부분의 기댓값은 다음과 같다.

$$\begin{aligned}&E\Big\{\sum_{l=2}^{2}2\widetilde{m_l^2}(1-2\widetilde{m_l^{-1}})\widetilde{S_{ll}^2}+\sum_{h=3}^{4}\sum_{j=1}^{2}\widetilde{m_h^2}(1-\widetilde{m_h^{-1}})\widetilde{S_{hj}^2}\Big\}\\ &\quad = \frac{6}{24}\sum_{h=1}^{4}[2m_h^2(1-2m_h^{-1})(S_{h1}^2+S_{h2}^2)+2m_h^2(1-m_h^{-1})(S_{h1}^2+S_{h2}^2)]\\ &\quad = \sum_{h=1}^{4}\sum_{j=2}^{2}m_h^2\left(\frac{2m_h-3}{2m_h}\right)S_{hj}^2\end{aligned} \tag{A.3}$$

식 (A.2)의 첫 번째 부분의 기댓값은

$$
\begin{aligned}
&E\{\widetilde{m}_1[(\widetilde{b}_1+\widetilde{c}_{11})-(\widetilde{b}_2+\widetilde{c}_{12})]-\widetilde{m}_2[(\widetilde{b}_1+\widetilde{c}_{21})-(\widetilde{b}_2+\widetilde{c}_{22})]\}^2 \\
&\quad = \frac{4}{24}\sum_{h}^{4}\sum_{<h'}^{4}(\omega_h-\omega_{h'})^2 \\
&\quad = \frac{1}{6}\left\{16\sum_{h=1}^{4}m_h^2[(b_1+c_{h1})^2-b_1^2]+4b_1^2\sum_{h}^{4}\sum_{<h'}^{4}(m_h-m_{h'})^2\right\} \qquad \text{(A.4)} \\
&\quad = \frac{4}{3}\sum_{h=1}^{4}\sum_{j=1}^{2}m_h^2[(b_j+c_{hj})^2-b_j^2]+\frac{1}{6}\left(\sum_{j=1}^{2}b_j^2\right)\sum_{h=1}^{4}\sum_{h'=1}^{4}(m_h-m_{h'})^2
\end{aligned}
$$

여기서 $\omega_h = m_h[(b_1+c_{h1})-(b_2+c_{h2})]$이며, (A.4)의 결과는

$$
\omega_h\omega_{h'} = \begin{cases} 4m_h^2(b_1+c_{h1})^2, & \text{if } h=h', \\ 4m_hm_{h'}[b_1^2+(c_{h1}+c_{h'1})b_1+c_{h1}c_{h'1}], & \text{if } h\neq h', \end{cases} \qquad \text{(A.5)}
$$

에 의하여 얻어진다. 식 (A.3)과 (A.4)를 이용하여 $\overline{y_{HT}}$의 분산인 식 (2.6)을 얻을 수 있다.

A.2. 결과2의 증명

인식번호 1과 2를 받은 층에 해당하는 $\widetilde{m}_1^2(\widetilde{y}_{111}-\widetilde{y}_{112})^2+\widetilde{m}_2^2(\widetilde{y}_{221}-\widetilde{y}_{222})^2$의 조건부 기댓값은

$$
\begin{aligned}
&E\{\widetilde{m}_1^2(\widetilde{y}_{111}-\widetilde{y}_{112})^2+\widetilde{m}_2^2(\widetilde{y}_{221}-\widetilde{y}_{222})^2|ID\} \\
&\quad = E\{\widetilde{m}_1^2(\widetilde{e}_{111}-\widetilde{e}_{112})^2+\widetilde{m}_2^2(\widetilde{e}_{221}-\widetilde{e}_{222})^2|ID\} \\
&\quad = \frac{\widetilde{m}_1^2}{\widetilde{m}_1(1-\widetilde{m}_1)}\sum_{k}^{\widetilde{m}_1}\sum_{\substack{k'\\k\neq k'}}^{\widetilde{m}_1}(\widetilde{e}_{11k}-\widetilde{e}_{11k'})^2+\frac{\widetilde{m}_2^2}{\widetilde{m}_2(1-\widetilde{m}_2)}\sum_{k}^{\widetilde{m}_2}\sum_{\substack{k'\\k\neq k'}}^{\widetilde{m}_2}(\widetilde{e}_{22k}-\widetilde{e}_{22k'})^2 \qquad \text{(A.6)} \\
&\quad = 2\widetilde{m}_1^2\widetilde{S}_{11}^2+2\widetilde{m}_2^2\widetilde{S}_{22}^2.
\end{aligned}
$$

식 (A.6)의 무조건부 기댓값은

$$
E\{\widetilde{m}_1^2(\widetilde{y}_{111}-\widetilde{y}_{112})^2+\widetilde{m}_2^2(\widetilde{y}_{221}-\widetilde{y}_{222})^2\} = \frac{1}{2}\sum_{h=1}^{4}\sum_{j=1}^{2}m_h^2S_{hj}^2 \qquad \text{(A.7)}
$$

인식번호 3과 4를 받은 층에 해당하는 $[\widetilde{m}_3(\widetilde{y}_{311}-\widetilde{y}_{321})-\widetilde{m}_4(\widetilde{y}_{411}-\widetilde{y}_{421})]^2$의 조건부 기댓값은

$$\begin{aligned} & E\{[\widetilde{m}_3(\widetilde{y}_{311}-\widetilde{y}_{321})-\widetilde{m}_4(\widetilde{y}_{411}-\widetilde{y}_{421})]^2|ID\} \\ &= E\{[(\widetilde{\omega}_3-\widetilde{\omega}_4)+\widetilde{m}_3(\widetilde{e}_{311}-\widetilde{e}_{321})-\widetilde{m}_4(\widetilde{e}_{411}-\widetilde{e}_{421})]^2|ID\} \qquad \text{(A.8)} \\ &= (\widetilde{\omega}_3-\widetilde{\omega}_4)^2+\sum_{h=3}^{4}\widetilde{m}_h^2(1-\widetilde{m}_h^{-1})(\widetilde{S}_{h1}^2+\widetilde{S}_{h2}^2) \end{aligned}$$

식 (A.8)의 무조건부 기댓값은

$$\begin{aligned} & E\{[\widetilde{m}_3(\widetilde{y}_{311}-\widetilde{y}_{321})-\widetilde{m}_4(\widetilde{y}_{411}-\widetilde{y}_{421})]^2\} \\ &= E[(\widetilde{\omega}_3-\widetilde{\omega}_4)^2]+\frac{1}{2}\sum_{h=1}^{4}\sum_{j=1}^{2}m_h^2(1-m_h^{-1})S_{hj}^2 \qquad \text{(A.9)} \end{aligned}$$

식 (A.7), (A.9) 그리고 (A.4)로부터 $\widehat{V}_M$은 $V_{CTPS}(\overline{y_{HT}})$의 불편 추정량이다.