The Korean Journal of Applied Statistics (2013) DOI: http://dx.doi.org/10.5351/KJAS.2013.26.6.1033
26(6), 1033–1041

Efficiency of Variance Estimators for Two-stage PPS Systematic Sampling

Young-won Kim[a,1] · Yeny Kim[a] · Hye-eun Han[a] · Eun-sun Kwak[a]

[a]Department of Statistics, Sookmyung Women's University

(Received November 13, 2013; Revised December 4, 2013; Accepted December 4, 2013)

Abstract

In this paper, we investigate several variance estimators for pps systematic sampling. Unfortunately, there is no unbiased variance estimators for a systematic sample because systematic sampling can be regarded as a random selection of one cluster. This study provides guidance on which variance estimator may be more appropriate than others in several circumstances. We judge the efficiency of variance estimators for systematic sampling based on of their relative biases and relative mean square error. Also, we investigate variance estimation problems for two-stage systematic sampling applied for the Food Raw Material Consumption Survey and the Establishment Labor Force Survey simulation study, in order to consider the popular two-stage pps systematic sample design for establishment and household survey in Korea.

Keywords: Two-stage systematic sampling, variance estimator, pps systematic sampling, relative bias, mean square error.

1. 서론

계통추출(systematic sampling)은 단순확률추출(simple random sampling)보다 추출작업이 쉽고 효과적으로 활용하는 경우 단순확률추출보다 정도(precision)가 높은 추정결과를 얻을 수 있기 때문에 널리 사용되는 표본추출방법이다. 계통추출은 집락추출의 일종으로 전체 추출단위를 k(추출간격)개의 집락으로 나누고 그 중 하나의 집락을 추출하는 것으로 볼 수 있다. 따라서 계통추출은 실행이 간단한 반면, 하나의 계통표본으로는 추정량의 분산에 대한 불편추정량을 구하는 것은 이론적으로 불가능하다는 문제가 있다. 흔히 계통표본을 단순확률표본으로 간주하고 단순확률추출의 분산공식을 사용하여 계통표본의 분산을 추정한다. 하지만 이 경우 분산추정량은 편향을 가질 수 있을 뿐만 아니라 구간추정 등 통계적 추론을 하는데 문제를 발생시킨다.

비복원 불균등확률추출에서 가장 널리 사용되는 방법인 크기비례(probability proportional to size; pps) 계통추출 역시 분산에 대한 불편추정량을 구할 수 없고, 또한 pps 계통표본의 경우 비복원 불균등확률추출에서 널리 사용하는 Yates와 Grundy (1953)의 분산추정량을 사용하는 데 어려움이 있다.

본 논문의 2절에서는 계통추출된 하나의 표본을 가지고 편향이 적은 분산추정량을 구현하기 위한 기존 연구들을 정리하고 각 추정량들의 장·단점에 대해 간단히 논의했다. 그리고 실제 표본조사에서 많이 활

This Research was supported by the Sookmyung Women's University Research Grants 2011.
[1]Corresponding author: Professor, Department of Statistics, Sookmyung Women's University, 52 Hyochangwon-gil, Yongsan-gu, Seoul 140-742, Korea. E-mail: ywkim@sm.ac.kr

용하는 2단계 불균등계통추출에서 분산 추정방법을 정리했다. 마지막으로 3절에서는 실제 사업체 및 가구 대상 조사 자료를 이용한 모의실험을 통해 2단계 pps 계통추출에서 각 분산추정량의 효율성을 비교했다.

2. 계통표본의 분산 추정

계통추출은 집락추출의 일종이라 할 수 있다. 다시 말하면 계통표본은 추출간격에 해당하는 k개의 집락 중에서 하나의 집락을 추출하는 특수한 형태의 집락추출법으로 볼 수 있다. 그러므로 하나의 계통표본을 가지고 추정량의 분산에 대한 불편추정량을 구하는 것은 이론적으로 불가능하다. 본 장에서는 하나의 계통표본을 가지고 모평균이나 총계 추정량의 분산 추정에 대한 기존 연구 결과들을 우선 정리하고자 한다 (Wolter, 2007).

2.1. 균등확률계통표본의 분산 추정

균등확률계통표본은 추출간격이 k일 때, k개의 가능한 표본 중에서 $1/k$의 확률로 하나의 계통표본이 추출되는 것을 말하며, 계통표본의 모평균에 대한 분산을 추정하는 가장 간단한 방법은 단순확률추출의 분산추정량을 이용하는 것이며 표본평균에 대한 분산추정량은 다음과 같다.

$$v_1(\bar{y}) = (1-f)\left(\frac{s^2}{n}\right), \tag{2.1}$$

여기서 $s^2 = \Sigma_{i=1}^{n}(y_i - \bar{y})^2/(n-1)$, $f = n/N = p^{-1}$이다. 모집단 단위들이 랜덤하게 정렬된 경우 계통추출의 분산추정량과 단순확률추출의 분산추정량은 근사적으로 같다.

모집단이 n개의 층으로 구성되어 있고 각 층의 크기는 k(추출간격)라고 가정하면, 계통표본은 각 층에서 표본단위를 하나씩 추출한 층화표본이라고 할 수 있다. 분산을 추정하기 위하여 두 개의 층을 하나로 묶어서 모집단을 총 $n/2$개의 층으로 간주하면, 계통표본을 각 층에서 단위를 2개씩 추출한 것으로 보고 다음과 같이 분산을 추정할 수 있다 (Cochran, 1946; Park, 2000).

$$v_2(\bar{y}) = \frac{N-n}{Nn^2}\sum_{i=1}^{\frac{n}{2}}(y_{2i} - y_{2i-1})^2, \tag{2.2}$$

여기서 v_2는 조사단위의 중복 사용을 허용하지 않은 것이고, 만약 조사단위를 중복해서 사용함으로써 분산추정량의 자유도를 증가시키는 경우 분산추정량은 아래와 같다.

$$v_3(\bar{y}) = (1-f)\left(\frac{1}{n}\right)\sum_{i=1}^{n-1}\frac{(y_{i+1} - y_i)^2}{2(n-1)}. \tag{2.3}$$

하나의 계통표본을 2개 이상의 크기가 같은 부차계통표본으로 나누어 표본평균들의 분산을 구함으로써 계통표본의 분산을 추정할 수도 있다. 크기가 n인 계통표본을 p개의 크기가 같은 부차계통표본으로 나누는 경우 α번째($\alpha = 1, \ldots, p$) 부차표본의 평균과 분산추정량은 다음과 같다. 참고로 본 논문의 각 분산추정량에 대한 사례분석 모의실험에서는 v_4는 이용하지 않았음을 미리 밝혀둔다.

$$v_4(\bar{y}) = \frac{1}{p(p-1)}\sum_{\alpha=1}^{p}(\bar{y}_\alpha - \bar{y})^2, \quad \text{여기서 } \bar{y}_\alpha = \frac{p}{n}\sum_{i=1}^{\frac{n}{p}} y_{p(i-1)+\alpha} \tag{2.4}$$

한편 어떤 경우에서나 최적인 계통추출에 대한 분산추정량은 존재하지 않는다. Wolter (2007) 등이 지적한 것처럼 v_1은 선형추세나 층화효과가 강한 모집단에서는 효율성이 떨어진다는 한계를 갖고 있고,

1차 차분을 근거로 한 v_2와 v_3은 강한 선형추세나 층화효과가 존재하는 모집단을 포함한 대부분의 모집단에서 좋은 분산 추정결과를 얻을 수 있다. 따라서 모집단에 대한 정보가 충분하지 않은 경우에 v_2와 v_3을 사용할 것을 권장한다. 특히 v_2는 표본크기가 작은 경우 유용하게 사용할 수 있다.

한편, 2차 이상의 고차 차분에 근거하여 얻은 분산추정량 역시 선형추세나 층화효과가 있는 모집단에서 우수한 분산 추정을 제공한다. 하지만 1차 차분을 통해 얻은 v_2나 v_3에 비해 분산추정량의 분산이 크며, 1차 차분을 통해 얻은 추정량보다 계산이 복잡하다는 단점이 있다. 계통표본을 여러 개의 부차계통표본으로 나누어 분산을 추정하는 v_4는 편향이 매우 심하며 분산추정량의 분산이 크기 때문에 분산의 최적 추정량이라고 할 수 없다. 그리고 v_4는 부차표본의 개수를 늘리면 분산은 작아지나 반대로 편향은 커지기 때문에 v_4를 사용하기 위해서는 편향과 MSE를 동시에 고려하여 적절한 부차표본의 개수를 결정해야 한다. Wolter (2007)은 모든 계통표본에서 항상 최적인 분산추정량은 존재하지 않지만, 모집단에 대한 정보가 충분하지 않다면 v_2 혹은 v_3의 사용을 권장하고 있다.

2.2. 크기비례계통표본의 분산 추정

크기비례(pps)계통추출은 비복원 불균등확률추출 방법 중 가장 널리 사용하는 표본추출방법 중 하나이다. 비복원 불균등확률추출의 분산 추정은 복원 불균등확률추출에 비해 훨씬 복잡하기 때문에 비복원추출을 실시하고도 복원추출의 분산 공식을 사용하는 경우가 많으며, 균등확률계통추출과 마찬가지로 분산에 대한 불편추정량이 존재하지 않는다. pps 계통표본의 총계추정량에 대한 분산을 추정하는 기존의 연구 결과를 정리하면 다음과 같다.

비복원 불균등확률추출에서 흔히 사용하는 Yates와 Grundy (1953)이 제안한 분산추정량은 다음과 같다.

$$\mathrm{Var}(\hat{Y}) = \sum_{i=1}^{n}\sum_{j>i}^{n}\left(\frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}}\right)\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2. \tag{2.5}$$

pps계통표본의 경우 랜덤출발점이 다른 두 추출단위의 결합포함확률은 0이 되고, 비복원 불균등확률추출의 경우에는 결합포함확률의 계산이 매우 복잡하기 하기 때문에 식 (2.5)를 이용하는데 어려움이 있다. 따라서 식 (2.5)의 추정량에 Hartley와 Rao (1962)의 결합포함확률의 근사식

$$\pi_{ij} = \frac{n-1}{n}\pi_i\pi_j + \frac{n-1}{n^2}\left(\pi_i^2\pi_j + \pi_i\pi_j^2\right) - \frac{n-1}{n^3}\pi_i\pi_j\sum_{i=1}^{N}\pi_j^2$$

을 대입하여 정리하면 다음의 분산추정량을 얻을 수 있다.

$$v_5(\hat{Y}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{i<j}^{n}\left(1 - \pi_i - \pi_j + \sum_{j=1}^{N}\frac{\pi_j^2}{n}\right)\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2. \tag{2.6}$$

이 분산추정량은 앞에서 언급한대로 편향이 존재한다. 하지만 랜덤정렬된 모집단에서 π_{ij}의 근사가 잘 이루어진 경우에는 그 편향이 작아 유용하게 사용할 수 있다.

두 번째 분산추정량은 pps계통표본을 복원추출을 이용하여 얻은 pps 표본으로 간주하고 분산을 추정하는 것으로 다음과 같다.

$$v_6(\hat{Y}) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{y_i}{p_i} - \hat{Y}\right)^2. \tag{2.7}$$

이 분산추정량 역시 편향을 갖지만, 단위들이 랜덤정렬된 모집단의 크기가 충분히 큰 경우 그 편향은 상당히 작아지므로 유용하게 사용할 수 있다.

크기비례계통표본을 크기가 동일한 $n/2$개의 각 층에서 2개의 표본단위씩 계통추출한 층화표본으로 보고 다음과 같이 분산을 추정할 수 있다.

$$v_7(\hat{Y}) = \frac{1}{n}\sum_{i=1}^{\frac{n}{2}}\left(\frac{y_{2i}}{p_{2i}} - \frac{y_{2i-1}}{p_{2i-1}}\right)^2 /n. \tag{2.8}$$

층화표본의 분산공식을 이용하여 계통표본의 분산을 추정한 v_7은 조사단위를 중복해서 이용하지 않고, 두 표본단위의 차이를 이용하여 얻은 것이다. 이와 같은 방법을 사용하되 표본단위를 중복해서 사용하면 (2.9)의 분산추정량을 얻을 수 있다.

$$v_8(\hat{Y}) = \frac{1}{n}\sum_{i=2}^{n}\left(\frac{y_i}{p_i} - \frac{y_{i-1}}{p_{i-1}}\right)^2 /2(n-1). \tag{2.9}$$

표본단위의 차이를 이용하여 총계추정량에 대한 분산을 추정하는 v_7과 v_8은 모집단에 선형관계가 존재하는 경우나 표본의 크기가 작은 경우 유용하게 사용할 수 있다.

pps계통표본의 총계 추정량에 대한 분산을 추정하는 또 다른 방법은 하나의 계통표본을 여러 개의 부차계통표본으로 나누어 총계추정량들의 분산을 이용하여 추정하는 것이다. 하나의 크기비례계통표본을 크기가 m인 p개의 계통표본으로 나누었다고 가정하면, 총계추정량 $\hat{Y}$에 대한 분산추정량은 다음과 같다.

$$v_9(\hat{Y}) = \frac{1}{p(p-1)}\sum_{\alpha=1}^{k}\left(\hat{Y}_\alpha - \hat{Y}\right)^2, \tag{2.10}$$

여기서 $\hat{Y}_\alpha$는 α번째 부차표본 총계에 대한 Horvitz-Thompson추정량이다. 이 추정량은 몇 개의 부차계통표본으로 나누느냐에 영향을 받는다. Wolter (2007)에 의하면 v_9의 분산은 부차표본의 수와 반비례한다. 즉, 부차 표본의 수가 적을수록 MSE는 커지고 부차표본의 수가 적을수록 편향은 작아진다. 따라서 v_9은 MSE와 편향을 동시에 고려하여 분산추정량을 선택할 때 항상 최적의 추정량이 될 수 없다.

마지막으로 크기비례계통표본에서 추정량의 분산을 추정하기 위해 잭나이프(jackknife) 방법을 이용할 수 있다. 계통추출에 의해 n개의 개체로 이루어진 표본이 선택되고 총계 추정량이 $\hat{Y}$일 때 표본을 이루는 개체들 중 j번째 개체를 제외하고 구한 총계추정치를 $\hat{Y}_{(j)}$이라고 하면, 총계에 대한 잭나이프 분산추정량은 아래와 같다.

$$v_{10}(\hat{Y}) = \frac{n-1}{n}\sum_{j=1}^{n}\left(\hat{Y}_{(j)} - \hat{Y}\right)^2. \tag{2.11}$$

한편, 실제 조사에서 표본추출의 효율을 높이기 위해서 집락을 추출한 뒤 표본집락 내의 모든 조사단위를 전부 조사하는 대신 각 표본집락 내에서 조사단위를 추출하는 2단 집락추출법을 많이 사용한다. 2단 집락추출에 있어서 1차추출단위를 비복원 불균등확률추출할 경우 총계의 Horvitz-Thompson 불편추정량은 다음과 같다.

$$\hat{t}_{HT} = \sum_{i\in S}\frac{\hat{t}_i}{\pi_i} = \sum_{i=1}^{N} Z_i\frac{\hat{t}_i}{\pi_i},$$

여기서 Z_i는 i번째 1차추출단위가 표본에 포함되는 경우 1의 값을 갖는 지시변수(indicator variable)이며 $\hat{t}_i$는 i번째 집락(PSU)에서의 총계추정량이다. 모집단 총계에 대한 Horvitz-Thompson 추정량의 분산추정량은 다음과 같다.

$$V\left(\hat{t}_{HT}\right) = \sum_{i=1}^{N} \frac{1-\pi_i}{\pi_i} t_i^2 + \sum_{i=1}^{N}\sum_{k\neq i}^{N} \frac{\pi_{ik}-\pi_i\pi_k}{\pi_i\pi_k} t_i t_k + \sum_{i=1}^{N} \frac{V(\hat{t}_i)}{\pi_i}$$
$$= \frac{1}{2}\sum_{i=1}^{N}\sum_{k=1,k\neq i}^{N} (\pi_i\pi_k - \pi_{ik}) \left(\frac{t_i}{\pi_i} - \frac{t_k}{\pi_k}\right)^2 + \sum_{i=1}^{N} \frac{V(\hat{t}_i)}{\pi_i}. \quad (2.12)$$

위 분산추정량에서 첫 번째 항은 1차추출단위 간의 분산이며 두 번째 항은 각 1차추출단위의 총계 t_i의 참값이 아닌 추정값을 사용함으로써 추가적으로 발생하는 2차추출단위 간의 분산이다. 따라서 식 (2.12)을 아래와 같이 표현할 수도 있다. 본 연구에서는 PSU와 SSU를 각각 계통추출한 것을 전제로 분산을 추정하게 된다.

$$V(\hat{t}_{HT}) = \mathrm{Var}_{psu} + \mathrm{Var}_{ssu}.$$

3. 사례분석 모의실험

모의실험을 통해 우리나라에서 사업체와 가구를 대상으로 한 표본설계에서 흔히 적용되는 2단계 계통집락추출에서의 효과적인 분산 추정 문제를 살펴보기로 하자. 앞서 살펴본 계통표본에서의 분산추정량들을 사업체 및 가구 부문 실제 조사 자료에 적용하여 2단계 pps계통표본에서의 효율적인 분산추정방법에 대해 알아보고자 한다. 이를 위해 2008년 사업체근로실태조사와 2011년 식품원료소비실태조사 자료를 이용하여 가상적인 모집단을 구성했다. 가상 모집단을 대상으로 한 모의실험을 통해 1차 추출단위(사업체, 조사구)를 pps계통추출함으로써 발생하는 분산을 $v_5, v_6, v_7, v_8, v_9, v_{10}$으로 추정하는 경우 각 분산추정량의 편향과 MSE를 비교하였다. 또한 v_1, v_2, v_3를 이용하여 2차 추출단위(근로자, 가구)를 계통추출함으로써 발생하는 분산을 추정하고, 각 추정량들의 편향과 MSE를 살펴보았다.

3.1. 사업체조사 계통표본 사례 모의실험

본 연구에서는 2008년 사업체근로실태조사의 조사 대상 사업체 중 도매 및 소매업에 해당하는 사업체 총 3,161개소에서 근로자 10인 미만을 고용하고 있는 사업체(1,858개소)는 제외하고, 근로자 10인 이상을 고용하고 있는 모든 사업체(1,303개소)와 이들 사업체에 고용된 전체 근로자(45,527명)를 모집단으로 보고 모의실험을 실시하였다. 2008년 사업체근로실태조사자료는 사업체규모와 각 사업체에 종사하는 근로자의 임금, 경력 등이 포함되어 있으며, 이 자료를 이용하여 근로자의 임금에 대한 분산추정량을 계산했다.

2단계 계통추출을 위한 모의실험의 1단계에서는 전체 1,303개 사업체 중에서 pps계통추출을 이용하여 100개의 사업체를 사업체규모(근로자수)에 비례하도록 추출하였다. 2단계에서는 각 사업체에서 근로자를 경력년수에 따라 나열한 후 10명의 근로자를 계통추출하였다. 모집단을 구성하는 1,303개의 사업체 규모를 나타내는 근로자수를 기준으로 추출간격(k)를 구하면 42이고, 따라서 모든 가능한 42개 계통표본에서 얻은 분산추정량의 편향과 MSE를 계산하여 분산 추정량의 효율성을 비교하고자 한다. v_9는 부차표본 개수에 영향을 받는데 본 모의실험에서는 크기가 100인 하나의 표본에서 계통추출을 이용하여 크기가 10인 표본 10개를 추출했다.

사업체(psu)간의 분산 추정값들을 이용하여 상대 편향(relative bias), 상대 MSE(relative MSE)를 구하고, 이를 기준으로 분산추정량을 비교하고자 한다. 분산추정량 v의 상대편향은 다음과 같이 계산한

Table 3.1. Comparison of variance estimation for systematic sampling of PSU(establishment)

	v_5	v_6	v_7	v_8	v_9	v_{10}
$E\{v\}$	16373.70	17718.50	14682.16	13735.43	11118.24	15212.98
Rel Bias$\{v\}$	0.23	0.33	0.11	0.03	−0.16	0.15
Rel MSE$\{v\}$	11.98	15.32	14.65	6.08	12.62	4.42

Table 3.2. Comparison of variance estimation for systematic sampling of SSU(laborer)

	v_1	v_2	v_3
$E\{v\}$	360.80	276.03	277.71
Rel Bias$\{v\}$	0.18	−0.10	−0.10
Rel MSE$\{v\}$	5.45	2.88	3.19

다.

$$\text{Rel Bias}(v) = \frac{E\{v\} - \text{Var}(\overline{y}_{sys})}{\text{Var}(\overline{y}_{sys})}.$$

분산추정량의 기댓값은 $E\{v\} = \Sigma_s v(s)/k$로 계산할 수 있다. 여기서 k는 모든 가능한 표본의 개수를 의미하고, s는 구간 $(0, p]$에서 랜덤하게 선택한 랜덤출발점을 의미하므로 주어진 분산추정량의 기댓값은 모든 가능한 표본의 분산추정량들의 단순평균과 같다. 분산추정량 v의 상대 MSE는 다음과 같이 계산한다.

$$\text{Rel MSE}(v) = \frac{E\left\{(v - \text{Var}(\overline{y}_{sys}))^2\right\}}{\left\{\text{Var}(\overline{y}_{sys})\right\}^2},$$

여기서 $E\{(v - \text{Var}(\overline{y}_{sys})\} = \Sigma_s\{v(s) - \text{Var}(\overline{y}_{sys})\}^2/k$이며 $\text{Var}(\overline{y}_{sys})$는 모집단에서 구한 평균추정량의 분산이다. 주어진 가상 모집단에서 $\overline{y}_{sys}$의 분산 $\text{Var}(\overline{y}_{sys})$은 13,273이다. 각 분산추정량 $v_5, v_6, v_7, v_8, v_9, v_{10}$의 기댓값, 상대편향, 상대 MSE는 Tabel 3.1과 같다.

Table 3.1을 보면 주어진 표본을 10개의 부차계통표본으로 나누어 표본평균의 분산으로 추정한 v_9를 제외한 나머지 추정량들이 실제 값보다 과대추정하고 있음을 확인할 수 있다. 인접한 표본 단위의 차이를 이용한 분산추정량 v_8이 가장 작은 편향을 갖고 있으며, 주어진 표본을 복원 크기비례추출을 통해 뽑은 표본으로 간주하고 분산을 추정한 v_6의 편향이 가장 크다. Hartely와 Rao (1962)의 결합포함확률 근사식을 이용한 분산 공식 v_5 또한 큰 편향을 가짐을 확인할 수 있다.

상대 MSE를 기준으로 분산추정량들을 비교해보면 잭나이프 방법을 이용한 분산추정량 v_{10}의 MSE가 확연하게 작음을 확인할 수 있다. 반대로 주어진 표본을 복원 크기비례추출로 간주하고 분산을 추정한 v_6가 가장 큰 MSE를 가짐을 볼 수 있다.

2단계 계통추출에 대한 모의실험을 위해 1단계 크기비례계통추출에 따른 100개 사업체를 추출한 후에, 2단계에서는 표본으로 추출된 각 사업체(psu)의 근로자(ssu)를 경력년수를 기준으로 나열한 후 10명의 근로자를 계통추출했다. 2절에서 정리한 계통표본의 분산추정량을 이용하여 2차 계통추출에 대한 분산을 추정하면 Table 3.2와 같고 분산의 참값 $\text{Var}(\overline{y}_{sys})$은 306.89이다.

Table 3.2를 보면 단순확률표본으로 간주하고 구한 v_1은 분산을 과대 추정했고, 반면에 추정량 v_2와 v_3은 분산을 과소 추정했다. v_1에 비해 v_2와 v_3의 분산 추정값이 더 편향이 작다는 것을 알 수 있다. MSE를 기준으로 분산추정량들을 비교해 보면 표본단위의 중복을 허용하지 않은 차이를 이용한 분산추정량 v_2의 MSE가 가장 작으며 반대로 v_1의 MSE가 가장 크다. 결과적으로 편향과 MSE를 기준으로 했을 때, 표본단위의 차이를 이용한 분산추정량 v_2이 가장 좋은 결과를 보였다.

Table 3.3. Comparison of variance estimation for systematic sampling of PSU(enumeration district)

	v_5	v_6	v_7	v_8	v_9	v_{10}
$E\{v\}$	195.72	215.41	68.91	59.18	113.23	48.16
Rel Bias$\{v\}$	14.46	16.02	4.44	3.68	7.94	2.80
Rel MSE$\{v\}$	226.16	275.78	26.79	15.10	110.90	242.57

3.2. 가구조사 계통표본 사례 모의실험

가구 모집단에서 2단 pps계통표본의 분산을 어떤 방법으로 추정하는 것이 효율적인지 분석하기 위해 2011년 식품원료소비실태조사 자료 중 가구부문 자료를 기초로 모집단을 설정하고 모의실험을 수행했다. 모의실험을 위해 2011년 식품원료소비실태조사에 응답한 가구 중에서 동일한 조사구 내에서는 20개의 가구를 복원 단순확률추출하고, 해당 조사구가 속한 층에서 나머지 가구를 복원 단순확률추출하여 조사구의 가구수가 모집단의 가구수와 일치하도록 가상적인 모집단을 구성하였다. 새롭게 구성한 가상적인 모집단은 400개 조사구, 38,412가구이며, 이 모집단을 이용해 모의실험을 실시했다.

모의실험을 통해 전체 가구의 연간 쌀 소비량의 평균을 추정하기 위해 2단 계통집락추출을 이용하여 가구를 추출하였다. 조사구당 평균 가구수는 96가구이지만 조사구에 최소 47가구에서 최대 233가구로 가구수의 차이가 크기 때문에 1단계에서는 전체 400개 조사구 중에서 pps계통추출을 이용하여 40개의 조사구를 조사구내 가구수에 비례하도록 계통추출하였다. 2단계에서는 각 조사구에서 가구를 가구원수에 따라 정렬한 후 계통추출을 이용하여 10가구를 추출했다. 2단 크기비례계통추출을 적용해 연간 쌀 소비량의 평균을 추정하고, 사업체 자료 분석과 마찬가지로 조사구(psu) 추출에 따른 분산과 조사구내 가구(ssu) 추출에 따른 분산을 추정하고, 각 추정량들의 편향과 MSE를 비교했다.

먼저 모집단에서 pps계통추출을 사용하여 40개의 조사구를 추출했다. 여기서 규모 변수(M_i)로는 조사구의 가구수를 사용했다. 모집단 내에 400개 조사구의 크기 변수를 고려해 40개 조사구를 계통추출하는 경우 추출간격(k)은 28인 것으로 나타났고, 따라서 이 경우 모든 가능한 계통표본이 28개이다. 2절에서 정리한 분산추정량 중 하나인 v_9을 구하기 위하여 본 모의실험에서는 크기가 40인 하나의 표본에서 계통추출을 이용하여 크기가 8인 표본 5개를 추출하였다.

모집단에서 $\bar{y}_{sys}$의 분산 Var($\bar{y}_{sys}$)은 12.69이고, 각 분산추정량 $v_5, v_6, v_7, v_8, v_9, v_{10}$의 기댓값, 상대편향, 상대 MSE는 Table 3.3과 같다.

Table 3.3을 보면 모든 추정량들이 실제 값보다 분산을 과대 추정하고 있다. 그러나 잭나이프 방법에 의한 분산추정량 v_{10}이 가장 작은 편향을 갖고 있으며, 인접한 표본 단위의 차이를 이용한 분산추정량 v_8이 상대적으로 작은 편향을 가짐을 볼 수 있다. 따라서 편향을 기준으로 보았을 때 가구 모집단 모의실험에서 분산추정량으로는 v_{10}과 v_8이 좋은 추정량이라고 할 수 있다. 상대 MSE를 기준으로 분산추정량들을 비교해보면 v_8의 MSE가 확연하게 작음을 볼 수 있다. 반면에 잭나이프 방법으로 추정한 v_{10}의 MSE가 매우 커짐을 확인할 수 있다. 따라서 종합적으로 보면 사업체 부문과 마찬가지로 v_8이 가장 바람직한 분산 추정량이라고 할 수 있다.

가구부문 2단계 계통추출에 대한 모의실험을 위해 1단계에서 크기비례계통추출로 40개 조사구를 표본으로 추출한 후에, 2단계에서는 표본 조사구(psu)의 가구(ssu)를 가구원수로 나열한 후 10 가구를 계통추출하였다. 2절에서 정리한 계통표본의 분산추정량 v_1, v_2, v_3을 이용하여 2단계 계통추출에 대한 기댓값과 상대편향, 상대 MSE를 구하면 Table 3.4와 같고, 모집단에서 구한 분산의 참값 Var($\bar{y}_{sys}$)은 10.03이다.

Table 3.4를 보면 세 가지 분산추정량 모두 분산을 과소 추정하고 있으며, v_1이 v_2, v_3에 비해 편향이 작

Table 3.4. Comparison of variance estimation for systematic sampling of SSU(household)

	v_1	v_2	v_3
$E\{v\}$	4.18	3.00	2.87
Rel Bias$\{v\}$	−0.58	−0.70	−0.71
Rel MSE$\{v\}$	0.34	0.49	0.51

았고 MSE도 v_1이 v_2, v_3에 비해 작다는 것을 알 수 있다.

4. 결론 및 시사점

본 연구에서는 계통표본추출에서 활용이 가능한 다양한 분산 추정 방법에 대해서 살펴보고, 국내에서 흔히 접하게 되는 사업체조사 및 가구조사에서 pps계통추출을 적용한 사례를 염두에 둔 모의실험을 통해 어떤 분산추정방법이 국내 가구 또는 사업체 조사의 계통추출에서 효과적인지를 살펴보았다.

모의실험에서는 사업체근로실태조사에서 근로자 평균 임금 추정에 대한 분산 추정 문제와 식품원료소비실태조사에서 가구당 연평균 쌀 소비량 추정에 대한 분산 추정 문제를 다루었다. 두 경우 모두 PSU와 SSU를 계통추출한 경우 분산추정방법에 따른 차이를 비교하고 효과적인 분산추정법이 어떤 것인지 살펴보았다.

모의실험 결과 PSU의 pps계통추출에 대한 분산추정에 있어서는 가구부문이나 사업체부문 모두 표본단위 간의 차이를 이용하여 분산을 추정한 방법이 편향이나 MSE 관점에서 효율적이라는 결론을 얻을 수 있었다.

아울러 모의실험 결과를 보면, 2단계 계통추출에서 전체 분산 중 PSU 계통추출에 따른 분산이 차지하는 비중이 전체 분산 중 사업체근로실태조사 모의실험의 경우 98%, 식품원료소비실태조사 모의실험의 경우 97%를 차지하기 때문에 실제 SSU 계통추출에 대한 분산추정은 가급적 단순한 방식으로 추정을 하더라도 큰 문제가 없다는 결론도 함께 얻을 수 있었다. 결과적으로 국내에서 흔히 사업체나 조사구를 PSU로 사용한 2단계 pps계통추출의 경우 PSU 추출에 대한 분산은 v_7이나 v_8과 같은 인접한 단위들간의 차이를 이용한 분산추정법이 효과적인 것으로 나타났고, SSU 추출에 대한 분산은 전체 분산 중 차지하는 비중이 작기 때문에 간단한 형태의 분산추정방법을 사용하더라도 큰 차이가 없다는 결론을 내릴 수 있다.

References

Cochran, W. (1946). Relative accuracy of systematic and stratified random samples for a certain class of populations, *Annals of Mathematical Statistics*, **17**, 164–177.

Hartley, H. O. and Rao, J. N. K. (1962). Sampling with unequal probabilities and without replacement, *Annals of Mathematical Statistics*, **33**, 350–374.

Park, H.-R. (2000). *Theory of Statistical Survey (revised)*, YoungJi-Moonhwa-sa.

Wolter, K. M. (2007). *Introduction to Variance Estimation Second Edition*, Springer-Verlag, New York.

Yates, F. and Grundy, P. M. (1953). Selection without replacement from within state with probability proportional to size, *Journal of the Royal Statistical Society, Series B*, **15**, 253–261.

2단 크기비례 계통추출법의 분산추정량 효율성 비교

김영원[a,1] · 김예니[a] · 한혜은[a] · 곽은선[a]

[a]숙명여자대학교 통계학과

(2013년 11월 13일 접수, 2013년 12월 4일 수정, 2013년 12월 4일 채택)

요 약

본 논문에서는 크기비례 계통추출법에서 적용할 수 있는 다양한 분산추정 방법들을 정리하고 각 분산추정 방법들의 통계적 특성에 대해서 논의하였다. 이론적으로 하나의 계통표본을 가지고 비편향 분산추정량을 구하는 것은 불가능하지만 실제 표본자료 분석에 있어서 어떤 대안이 있을 수 있는지 살펴보고, 다양한 분산추정 방법들의 성질을 상대편향 및 상대평균제곱오차 관점에서 비교해 보았다. 또한 우리나라 가구나 사업체 표본설계에서 흔히 발생하는 2단 크기비례 계통추출 표본에서 적용 가능한 효과적인 분산추정 방법을 알아보기 위해 2008년 사업체근로실태조사 자료의 근로자 평균임금과 2011년 식품원료소비실태조사 자료의 가구당 연평균 쌀 소비량의 분산 추정 문제를 기초로 모의실험을 수행하였다.

주요용어: 2단 계통추출법, 분산추정량, 크기비례 계통추출법, 상대편향, 상대평균제곱오차.

이 논문은 2011년도 숙명여자대학교의 교내연구비에 의하여 수행되었음.

[1]교신저자: (142-742) 서울특별시 용산구 청파로47길 100, 숙명여자대학교 통계학과, 교수.
E-mail: ywkim@sm.ac.kr