光州保健專門大學 論文集 第20輯(1995)
The Journal of Kwangju Health College. Vol. XX.

표본추출방법간의 효율성 비교

임상병리과
전임강사 송 정

Ⅰ. 서 론

모집단에서 표본을 추출 할 때 집락을 추출 단위로하는 집락추출 방법은 각 집락의 크기가 균등하지 않은 경우는 집락의 크기에 비례하여 추출 확률을 부여하는 것이 합리적이다.

불균등 확률을 사용하여 표본을 뽑는 방법은 지금까지 많은 학자들에 의하여 연구되어져 왔고, 불균등 확률 추출로서 가장 많이 사용되는 방법은 확률비례추출이다. 이것은 각 단위를 그 크기에 비례하는 확률로서 추출하는 방법이다.

본 논문은 확률비례추출에서 비복원추출 방법 중 특수한 추정량을 사용하는 방법들에 대해 알아보고 각 추정량의 분산과 확률비례복원추출에 대한 상대효율을 구하여 각 방법들의 효율성을 비교하였다.

Ⅱ. 기 호

다음은 본 논문에서 사용되어질 기호에 대한 정의이다.

N : 모집단의 크기

n : 표본의 크기

Y_i : i 번째 표본추출 단위의 값

Y : Y_i 의 총합 ($Y = \sum_{i=1}^{N} Y_i$)

y_i : i 번째 표본의 값

p_i : i 번째 표본추출단위가 첫번째 시행에서 표본으로 추출될 확률

π_i : i 번째 표본추출단위가 표본으로 뽑힐 확률

π_{ij} : i 번째 표본추출단위와 j 번째 표본추출단위가 표본으로 함께 뽑힐 결합확률

$\hat{Y}$: 모집단 총합에 대한 추정량

$V(\hat{Y})$: $\hat{Y}$ 에 대한 분산

$v(\hat{Y})$: $V(\hat{Y})$ 에 대한 추정량

Ⅲ. 확률비례추출방법

확률비례추출방법에 의하여 집락을 추출하는 경우 추출된 집락의 복원 여부에 따라 복원추출과 비복원추출로 나뉜다. 복원추출은 추정량과 그것의 분산식이 간단하다는 장점은 있으나 같은 집락이 여러번 추출 될 수 있다는 문제가 있다. 비복원추출은 1949년 Madow에 의해서 처음 연구 되어진 이래로 지금까지 많은 방법들이 제시 되어졌다. 많은 방법들 중 본 논문에서는 Horvitz-Thompson 추정량, Das Raj의 방법, Murthy의 방법에 대해서 비교하기로 한다.

1. Horvitz - Thompson 추정량

Horvitz 와 Thompson[7]은 1952년에 비복원추출방법을 사용한 불균등확률로 표본을 뽑는 방법을 일반화 시켰다. Horvitz-Thompson의 모집단 총합에 대한 추정량은

$$\hat{Y}_{HT} = \sum_{i}^{n} \frac{y_i}{\pi_i} \quad \cdots\cdots\cdots\cdots (1)$$

이다. 여기서 y_i 는 i 단위의 관찰값이다.

모집단에 있는 i 번째 표본추출 단위가 표본으로 추출될 확률 π_i 값과, i , j 번째 표본추출 단위가 표본으로 추출될 확률 π_{ij} 값은 다음과 같다.

$$\pi_i = p_i + \sum_{j>i}^{N} \frac{p_i p_j}{1-p_j} \quad \cdots\cdots\cdots\cdots \quad (2)$$

$$\pi_{ij} = p_i p_j \left(\frac{1}{1-p_i} + \frac{1}{1-p_j}\right) \quad \cdots\cdots\cdots\cdots \quad (3)$$

단, p_i , p_j 는 i , j 번째 표본추출 단위가 첫번째 시행에서 뽑힐 확률이다.

<정리 1> $\pi_i > 0$ 인 경우 ($i = 1, \ldots, N$), $\hat{Y}_{HT}$ 는 Y 의 불편추정량이 되고 그 분산은 다음과 같다.

$$V(\hat{Y}_{HT}) = \sum_{i}^{N} \frac{1-\pi_i}{\pi_i} y_i^2 + 2\sum_{i}^{N}\sum_{j>i}^{N} \frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j} y_i y_j \quad \cdots\cdots \quad (4)$$

(증명) $\hat{Y}_{HT}$ 가 Y의 불편추정량임을 증명하기 위하여 확률변수 t_i ($i=1, \ldots, N$)를 도입한다. 만약 i 단위가 선정되면 $t_i = 1$, 그렇지 않으면 $t_i = 0$을 취하는 변수이다. 그러면 t_i는 이항분포를 가지므로 다음 성질을 갖는다.

$$E(t_i) = \pi_i + 0(1-\pi_i) = \pi_i$$

$$V(t_i) = \pi_i(1-\pi_i)$$

$$Cov(t_i, t_j) = E(t_i t_j) - E(t_i)E(t_j) = \pi_{ij} - \pi_i\pi_j$$

만약 i 단위와 j 단위가 동시에 선출되면 $t_i t_j = 1$, 그렇지않으면 $t_i t_j = 0$ 이므로 $E(t_i t_j) = \pi_{ij}$이다. 따라서 y_i를 계수로 보고, t_i를 확률변수로 정의 하면

$$E(\hat{Y}_{HT}) = E\left(\sum_{i}^{N} \frac{t_i y_i}{\pi_i}\right) = \sum_{i}^{N} \frac{y_i E(t_i)}{\pi_i} = Y$$

$$V(\hat{Y}_{HT}) = V\left(\sum_{i}^{N} \frac{t_i y_i}{\pi_i}\right)$$

$$= \sum_{i}^{N} \left(\frac{y_i}{\pi_i}\right)^2 V(t_i) + 2\sum_{i}^{N}\sum_{j>i}^{N} \frac{y_i}{\pi_i}\frac{y_j}{\pi_j} Cov(t_i t_j)$$

이고, 위에서 얻은 $V(t_i)$와 $Cov(t_i, t_j)$를 대입하여 (4)식을 얻게된다.

$\hat{Y}_{HT}$의 분산의 불편추정량은

$$v_1(\hat{Y}_{HT}) = \sum_{i}^{n} \frac{1-\pi_i}{\pi_i^2} y_i^2 + 2\sum_{i}^{n}\sum_{j>i}^{n} \frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j\pi_{ij}} y_i y_j \quad \cdots\cdots\cdots \quad (5)$$

이다. Sen[13]과 Yates, Grundy[14]는 $V(\hat{Y}_{HT})$의 추정량을 다음과 같이 유도 하였다.

$$v_2(\hat{Y}_{HT}) = \sum_{i}^{n}\sum_{j>i}^{n} \frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2 \quad \cdots\cdots\cdots\cdots (6)$$

이들 분산 추정량의 결점은 어떤 표본에서는 음의 값을 취하게 되어 추정량의 신뢰도를 측정 할 수 없게 된다는 문제점이 있다. 이러한 점을 보완해서 Brewer[5], Durbin[6], Samford[12]는 $v_2(\hat{Y}_{HT})$가 항상 양의 값을 갖게하는 표본추출 방법을 제시하였다.

2. Das Raj 방법

Das Raj[9]는 표본을 뽑는 순서를 고려하여 각각의 단계에 새로운 변량을 결합시킨 다음, 이것의 기대치가 원래의 변량의 모수와 같아지도록 하였다. Das Raj의 추정량은

$$\hat{Y}_D = \frac{1}{n}\sum_{i}^{n} t_i \quad \cdots\cdots\cdots\cdots (7)$$

이다. 여기서 t_i는 i 번째까지 추출된 표본단위에 근거한 Y의 불편추정량으로서 다음과 같이 정의한다.

$$t_i = y_1 + y_2 + \ldots + y_{i-1} + \frac{y_i}{p_i}(1 - p_1 - \ldots - p_{i-1}) \quad \cdots\cdots\cdots (8)$$

t_i의 기대값은

$$E(t_i / y_1, y_2, \ldots, y_{i-1}) = y_1 + \ldots + y_{i-1} + \sum{}' y_i = Y \quad \cdots\cdots\cdots (9)$$

이다. 여기서 $\sum'$은 처음 $i-1$ 회 까지 추출된 단위를 제외한 나머지 모집단에 대한 합이 된다. $\hat{Y}_D$는 n개의 불편추정량 t_i의 평균이므로 Y의 불편추정량이 된다. Das Raj는 $\hat{Y}_D$의 분산과 분산의 불편추정량을 다음과 같이 유도 하였다.

$$V(\hat{Y}_D) = \frac{1}{8}\sum_{i}^{N}\sum_{j>i}^{N} p_i\, p_j (2 - p_i - p_j)\left(\frac{y_i}{p_i} - \frac{y_j}{p_j}\right)^2 \quad \cdots\cdots\cdots (10)$$

$$v(\hat{Y}_D) = \frac{1}{n(n-1)}\sum_{i}^{n}(t_i - \bar{t})^2, \qquad \bar{t} = \hat{Y}_D \quad \cdots\cdots\cdots (11)$$

표본의 크기 $n=2$인 경우, 추정량 $\hat{Y}_D$와 $\hat{Y}_D$의 분산의 불편추정량 $v(\hat{Y}_D)$는

$$\hat{Y}_D = \frac{1}{2}(t_1 + t_2)$$

$$= \frac{1}{2}\left[\frac{y_1}{p_1}(1 + p_1) + \frac{y_2}{p_2}(1 - p_1)\right] \quad \cdots\cdots\cdots\cdots (12)$$

$$v(\hat{Y}_D) = \frac{1}{4}(t_1 - t_2)^2$$

$$= \frac{1}{4}(1-p_1)^2\left(\frac{y_1}{p_1} - \frac{y_2}{p_2}\right)^2 \quad \cdots\cdots\cdots\cdots \quad (13)$$

이다.

3. Murthy 방법

Murthy[8]는 표본추출 단위가 뽑히는 순서에 의존하지 않는 추정량을 제시 하였는데 이러한 무순위 추정량은 순위 추정량에 비해 분산이 작아지는 장점을 지니고 있다. 무순위 추정량은 순위 추정량의 추출확률을 가중값으로 하여 가중평균을 구한다.

표본의 크기 $n=2$일때, Murthy의 방법에 의해서 표본으로 y_1과 y_2가 추출 되었다고 하자. 먼저 y_1이 뽑힌 다음 y_2가 뽑힌 표본을 s_1, y_2가 뽑힌 다음 y_1이 뽑힌 경우를 s_2라 하면 s_1과 s_2가 뽑힐 확률과 순위 추정량은

$$p(s_1) = \frac{p_1 p_2}{1-p_1} \quad \cdots\cdots\cdots\cdots \quad (14)$$

$$\hat{Y}_D(s_1) = \frac{1}{2}\left[\frac{y_1}{p_1}(1+p_1) + \frac{y_2}{p_2}(1-p_1)\right] \quad \cdots\cdots\cdots\cdots \quad (15)$$

$$p(s_2) = \frac{p_1 p_2}{1-p_2} \quad \cdots\cdots\cdots\cdots \quad (16)$$

$$\hat{Y}_D(s_2) = \frac{1}{2}\left[\frac{y_2}{p_2}(1+p_2) + \frac{y_1}{p_1}(1-p_2)\right] \quad \cdots\cdots\cdots\cdots \quad (17)$$

이다. 그러므로 뽑히는 순서를 생각하지 않고 y_1과 y_2가 뽑힐 확률은 다음과 같다.

$$p(s) = p(s_1) + p(s_2)$$

$$= \frac{p_1 p_2(2-p_1-p_2)}{(1-p_1)(1-p_2)} \quad \cdots\cdots\cdots\cdots \quad (18)$$

Murthy에 의한 무순위 추정량 $\hat{Y}_M$ 은 다음과 같이 정의 한다.

$$\hat{Y}_M = \frac{\hat{Y}_D(s_1)p(s_1) + \hat{Y}_D(s_2)p(s_2)}{p(s_1)+p(s_2)}$$

$$= \frac{1}{2-p_1-p_2}\left[(1-p_2)\frac{y_1}{p_1} + (1-p_1)\frac{y_2}{p_2}\right] \quad \cdots\cdots \quad (19)$$

$\hat{Y}_M$ 의 분산과 분산의 불편추정량은

$$V(\hat{Y}_M)=\sum_{i}^{N}\sum_{j>i}^{N}\frac{p_i\,p_j(1-p_i-p_j)}{2-p_i-p_j}\left(\frac{y_i}{p_i}-\frac{y_j}{p_j}\right)^2 \quad \cdots\cdots\cdots \quad (20)$$

$$v(\hat{Y}_M)=\frac{(1-p_i)(1-p_j)(1-p_i-p_j)}{(2-p_i-p_j)^2}\left(\frac{y_i}{p_i}-\frac{y_j}{p_j}\right)^2 \quad \cdots\cdots \quad (21)$$

이다.

Ⅳ. 추정량의 비교

지금까지 여러가지 추출방법과 그에 따른 추정량을 고찰 하였다. 이들 추정량의 효율을 이론적으로 비교하는 것은 곤란하므로 가설적인 세개의 모집단에 대하여 수식적인 비교 분석을 한다. Yates와 Grundy는 표 1 과 같은 세개의 모집단 A, B, C를 가정하였다. 모집단의 크기는 $N=4$이고, 각각 2개 단위를 추출 한다.

< 표 1 > 모집단의 예

	A				B				C			
M_i	.1	.2	.3	.4	.1	.2	.3	.4	.1	.2	.3	.4
y_i	.5	1.2	2.1	3.2	.8	1.4	1.8	2.0	.2	.6	.9	.8

표 2는 각 추정량의 분산과 확률비례복원추출에 대한 상대효율이다. 상대효율은

$$RE(\hat{Y}/\hat{Y}_{pps})=[\,V(\hat{Y}_{pps})\,/\,V(\hat{Y})]\times 100 \quad \cdots\cdots\cdots \quad (22)$$

으로 정의된다.

< 표 2 > 표본추출방식의 상대효율

추출방식	추정량	A		B		C	
		$V(\hat{Y})$	RE	$V(\hat{Y})$	RE	$V(\hat{Y})$	RE
pps wr	$\hat{Y}_{pps}$	.500	100	.500	100	.125	100
pps wor	$\hat{Y}_{HT}$	.823	61	.057	877	.059	212
pps wor	$\hat{Y}_{D}$	.365	137	.367	137	.088	142
pps wor	$\hat{Y}_{M}$	.312	160	.312	160	.070	179

표 2를 보면 Horvitz - Thompson 추정량을 제외한 Das Raj, Murthy 추정량은 세 모집단 전부에 있어서 확률비례복원추정량 보다 효율이 높다는 것을 알 수 있다. 또한 Murthy 추정량이 Das Raj 추정량 보다 세 모집단 모두에서 효율이 높게 나타난 점으로 미루어 우리는 순위 추정량 보다 무순위 추정량이 더 효율이 높다는 것을 알 수 있다.

V. 결 론

일반적으로 표본의 크기가 같은 경우 비복원표본추출이 복원표본추출 보다 효율이 높다고 할 수 있다. 그러나 Horvitz - Thompson의 비복원추출 방법은 복원추출 방법 보다 효율성이 낮은 경우도 있으며 분산의 추정량이 음의 값을 가질 수 있다는 단점이 있다.

Murthy의 방법은 다른 방법보다 가장 효율성이 높고 우수한 방법이지만 표본의 수가 커지는 경우 분산의 추정량을 구하기가 어려운 문제점이 있다.

그러므로 표본추출 방법 중 어떤 방법이 절대적으로 좋은 방법이라고 할 수 는 없고, 모집단의 성격이나 크기, 표본의 수 등을 고려하여 가장 효율성이 높은 방법을 그때 그때 적용하는 것이 바람직 하다.

참고문헌

1. 김종호, 표본조사법, 서울, 자유아카데미, (1991)
2. 남궁평 · 김연형, 표본이론, 서울, 박영사, (1986)
3. 박홍래, 통계조사론, 서울, 영지문화사, (1989)
4. Yamane, Taro : Elementary Sampling Theory, Prentice-Hall, Inc., Englewood Cliffs, N. J. (1967)
5. Brewer, K.R.W. and Hanif, M. : "Sampling with unequal probabilities", *Lecture Notes in Statistics*, **15** (1983)
6. Durbin, J. : "Some results in sampling when the units are selected with unequal probabilities", *Journal of the Royal Statistical Society*, **15**, 262 -269 (1953)
7. Horvitz, D. G. and Thompson, D. J. : "A generalization of sampling without replacement from a finite universe", *Journal of the American*

Statistical Association, **47**, 663~685 (1952)

8. Murthy, M.N. : "Ordered and Unordered Estimators in Sampling without Replacement", *Sankhya*, **18**, 379~390 (1957)
9. Raj, D. : "Some Estimators in Sampling with Varying Probabilities without Replacement", *J. R. S. S.*, **51**, 269~284 (1956)
10. Raj, D. : "Variance Estimation in Randomized Systematic Sampling with Probability Proportionate to Size", *J. R. S. S.*, **60**, 278~284 (1965)
11. Rao, J.N.K. and Vijayan, K. : "On Estimating the Variance in Sampling with Probability Proportional to Aggregate Size", *J. A. S. A.*, **72**, 579~584 (1977)
12. Sampford, M.R. : "On Sampling without Replacement with Unequal Probabilities of Selection", *Biometrika*, **54**, 499~513 (1967)
13. Sen, A.R. : "On the Estimate of the Variance in Sampling with Varying Probabilities", *Journal of Indian Social and Agricultural Statistics*, **5**, 119~127 (1953)
14. Yates, F. and Grundy, P.M. : "Selection without Replacement from within Strata with Probability Proportional to Size", *J. R. S. S.*, **15**, 253~261 (1953)

On the comparison of an efficiency for sampling methods

Song, Jeong
Dept. of Clinical Pathology
Kwangju Health College

> Abstract <

The purpose of this paper is to compare several sampling methods using in sampling with probability proportional size.

A sampling with probability proportional size, is divided into two, with replacement and without replacement.

In the case of sampling without replacement, we compare the sampling techniques using special estimators such as Horvitz-Thompson estimator, Das Raj estimator and Murthy estimator.

Based on the comparisons of the variance of estimators and the relative efficiency for the probability proportional size with replacement , Das Raj and Murthy estimators are more useful than other estimators in sampling with probability proportional size with replacement.

Especially, the method of Murthy has a high efficiency than others.