응용통계연구 제 17권 1호, 2004년, pp. 27-34

변수변환에 의한 수정 HORVITZ-THOMPSON추정량

류제복[1)]

요 약

Horvitz-Thompson(H-T)추정량이 확률비례추정량에 비해 효율이 떨어지는 경우가 있다. 이를 극복하기 위해서 2단계 변수변환을 한다. 1단계로는 Midzuno-Sen추출을 적용하기 위해서 보조변수를 변환하고, 이로부터 얻은 포함확률을 H-T추정량에 사용할 때 분산을 줄이기 위해서 2단계로 연구변수를 변환하였다. 이러한 변환을 통해 얻은 추정량과 기존의 PPS추정량들을 비교하였다.

주요용어: Midzuno-Sen추출법, 포함확률, 변수변환, Horvitz-Thompson추정량

1. 머리말

유한 모집단으로부터 표본을 추출할 때, 모집단을 구성하고 있는 단위들의 크기에 비례해서(probability proportional to size; PPS)추출하는 것이 효율적이다. Hansen과 Hurwitz(1943)는 각 층에서 단지 하나의 1차 추출단위(primary sampling unit: PSU)를 추출할 때 각 단위들이 표본으로 추출될 확률을 달리 사용하였다. Horvitz와 Thompson(1952)은 Hansen과 Hurwitz(1943)의 이론을 일반화 시켰다. 즉, 서로 다른 N개의 단위로 구성된 모집단에서 비복원불균등확률로 n 개의 단위를 추출하여 모집단총계에 대한 비편향추정량과 추정량의 분산에 대한 비편향추정량을 제시하였다. 이때 추출확률은 연구변수(Y)의 크기에 비례해서 추출확률을 결정하는 데 통상적으로 연구변수는 미지의 변수가 되므로 이와 상관이 높은 보조변수(X)를 사용한다.

$$\hat{Y}_{HT} = \sum_{i=1}^{n} \frac{y_i}{\pi_i}. \tag{1.1}$$

이 추정량은 비편향추정량이 되며 추정량의 분산은 다음과 같다.

$$\mathrm{Var}(\hat{Y}_{HT}) = \sum_{i}^{N}\sum_{j>i}^{N}(\pi_i\pi_j - \pi_{ij})\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2. \tag{1.2}$$

여기서 y_i는 단위 i의 관찰값이고 π_i는 단위 i가 표본에 포함될 확률이며, π_{ij}는 단위 i와 j가 동시에 표본에 포함될 확률이다. 분산추정량으로 Yates와 Grundy(1953) 그리고 Sen (1953)은 (1.3)식을 제시하였다.

$$var(\hat{Y}_{HT}) = \sum_{i}^{n}\sum_{j>i}^{n}\frac{(\pi_i\pi_j - \pi_{ij})}{\pi_{ij}}\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2. \tag{1.3}$$

1) (360-764) 충북 청주시 상당구 내덕동 36번지, 청주대학교 생명·유전·통계학부, 교수
E-mail: jbryu@chongju.ac.kr

그러나 H-T추정량은 몇 가지 단점을 가지고 있다. 첫째는 분산 추정량이 어떤 표본에 대해서 음수 값을 가질 수 있고, 둘째는 $n > 2$인 경우 π_i와 π_{ij}의 계산이 번거로우며, 셋째는 모집단의 모든 단위에 대해서 y_i/π_i가 일정하지 않으면 추정량의 분산은 커지게 된다. 따라서 이러한 단점을 보완하고 PPS추출의 효율과 실용성을 높이기 위해서 변수변환을 사용한다. Prasad와 Srivenkataramana(1980) 그리고 Srivenkataramana와 Tracy(1979) 등은 연구변수의 변환을 고려하였고, Mohanty(1978), Rao(1988), Shaoo(1997), Bedi와 Agarwal(1999) 등은 포함확률이 추출확률에 비례되게 해주기 위해서 보조변수를 변환하였다. 그리고 Kumar와 Srivenkataramana(1994)는 M-S추출법을 일반화시키고 분산을 줄이기 위해서 변수변환을 이용하였다.

본 논문에서는 H-T추정량을 실제에 편리하게 적용하고 추정량의 효율을 높이기 위해서 1단계로 보조변수를 변환하여 M-S추출법의 사용을 가능하게 하고, 이로부터 얻은 포함확률을 H-T추정량에 사용할 때 분산을 줄이기 위해서 2단계로 연구변수를 변환하였다. 그리고 이러한 변환을 통해 얻은 추정량과 기존의 PPS추정량들을 비교하였다.

2. Midzuno-Sen 추출을 위한 변환

H-T추정량의 분산추정 값이 음수가 되는 문제와 $n > 2$인 경우 π_i와 π_{ij}의 계산을 용이하게 하기 위해서 Midzuno(1952)와 Sen(1953)은 다음과 같은 표본추출방법을 제안하였다.

1단계 : 첫 번째 조사단위는 PPS로 추출하고,

2단계 : 나머지 $(n-1)$개의 조사단위는 비복원균등확률로 추출한다.

보조변수 X를 사용하여 $p_i = x_i/\sum x_i$라 두면, M-S추출법에 의하여 1차와 2차의 포함확률은 다음과 같이 된다.

$$\begin{aligned}
\pi_i &= p_i + \sum_{j\neq i} p_j \frac{n-1}{N-1} \\
&= \frac{n-1}{N-1} + \frac{N-n}{N-1} p_i \\
&= \alpha_1 + \alpha_2 p_i, \qquad (2.1) \\
\pi_i &= p_i \left(\frac{n-1}{N-1}\right) + p_j \left(\frac{n-1}{N-1}\right) + \sum_{k\neq i,j}^{N} p_k \left(\frac{n-1}{N-1}\right)\left(\frac{n-2}{N-2}\right) \\
&= \frac{\alpha_1}{N-2}[(N-n)(p_i+p_j) + (n-2)]. \qquad (2.2)
\end{aligned}$$

여기서, $\alpha_1 = (n-1)/(N-1)$, $\alpha_2 = (N-n)/(N-1)$이다.

(2.1)식과 (2.2)식으로부터 $\pi_i\pi_j > \pi_{ij}$가 되어 분산추정량 (1.3)식은 항상 양수가 된다. M-S추출법은 추출과정이 간단하고 π_i와 π_{ij}의 계산이 쉬워 실제 적용이 편리하다. M-S추출법을 사용한 H-T추정량이 다른 PPS추출법보다 효율이 좋게 되기 위해서는 M-S추출법에 의한 π_i가 p_i와 비례되어야 한다. 이 조건을 만족시키기 위해서 Mohanty(1978), Rao(1988), Shaoo(1997), Bedi와 Agarwal(1999) 등은 보조변수를 변환하였다.

M-S추출법을 사용할 때는 모든 단위의 추출확률 p_i가 (2.3)식을 만족해야 한다. 그런데 이러한 조건은 M-S추출법을 실제조사에 활용하는데 부담이 된다.

$$p_i > \frac{n-1}{n(N-1)}(=k). \tag{2.3}$$

따라서 본 논문에서는 모집단에 있는 일부 단위의 추출확률 p_i가 (2.3)식의 조건을 만족시키지 못할 때 1단계로 보조변수를 적절히 변환시켜서 수정된 추출확률 p_i'가 (2.3)식의 조건을 만족시키도록 한다.

이를 위해서 보조변수를 다음과 같이 변환하여 새로운 변수 u를 얻는다(Reddy와 Rao (1977) 그리고 Rao(1988)).

$$u_i = x_i + d\bar{X}. \tag{2.4}$$

이때,

$$U = \sum_{i=1}^{N} u_i = \sum_{i=1}^{N}(x_i + d\bar{X}) = X(1+d) \tag{2.5}$$

이고 Y와 X의 상관은 Y와 U의 상관과 같으며 변환된 추출확률 p_i'는 다음과 같이 된다(Y와 X의 상관이 음수일 때는 (2.4)식에서 x_i대신에 $-x_i$를 사용한다).

$$p_i' = \frac{u_i}{U} = \frac{Np_i + d}{N(1+d)}. \tag{2.6}$$

(2.6)식의 p_i'도 (2.3)식의 조건을 만족해야하므로 (2.3)식과 (2.6)식으로부터 d값을 구하면,

$$d > \frac{N(k-p_i)}{1-Nk} \tag{2.7}$$

가 된다. 그런데 초기확률 중에서 가장 작은 값을 $p_{(1)}$이라 할 때 $p_{(1)} < k$이면 d의 값은 양수가 되지만 (2.3)식이 만족되지 않아 M-S추출법을 사용할 수 없게 된다. 따라서 (2.7)식을 만족하는 d값을 취해 (2.4)식과 같이 변수변환을 하면 M-S추출법을 사용할 수 있고, 변수변환을 통해 p_i대신 p_i'를 계산해서 (2.1)식과 (2.2)식에 대입하여 다음과 같은 포함확률을 얻는다.

$$\pi_i' = \frac{1}{(N-1)(1+d)X}\left[(n-1)X + (N-n)x_i + \frac{n(N-1)}{N}dX\right], \tag{2.8}$$

$$\pi_{ij}' = \frac{(n-1)}{(N-1)(N-2)(1+d)X}\left[(n-2)X + (N-n)(x_i+x_j) + \frac{n(N-2)}{N}dX\right]. \tag{2.9}$$

(2.8)식과 (2.9)식으로부터 $\pi_i'\pi_j' > \pi_{ij}'$됨을 알 수 있다. 또한 (1.1)식에 π_i대신 (2.8)식에서 얻어진 π_i'를 사용하여 H-T추정량을 얻는다. 여기서 얻은 추정량은 M-S추출법을 사용함으로써 분산추정량이 음수가 되는 것을 방지할 수 있다.

3. 수정된 Horvitz-Thompson 추정량

보조변수를 변환하여 얻은 추출확률 p_i'로부터 구한 π_i'를 (1.1)식에 대입하여 모집단 총계를 추정한다. 그러나 이때 모집단의 모든 단위에 y_i/π_i'가 일정하지 않으면 추정량의 분산은 커지게 된다. 비록 모든 원소 i에 대해서 $y_i = \beta x_i$의 관계가 있다 해도 (2.1)식의 α_1때문에 y_i/π_i'는 상수가 되지 않는다. 이러한 문제를 해결하기 위해서 연구변수를 다음과 같이 변환한다(Prasad와 Srivenkataramana(1980)). 이때, b는 선택된 상수이다.

$$z_i = y_i + \alpha_1\alpha_2^{-1}b, \quad (i = 1, \cdots, N). \tag{3.1}$$

이로부터 모집단 총계 Y의 비편향추정량과 이의 분산은 각각 다음과 같다.

$$\hat{Y}_{RHT(MS)} = \sum_{i=1}^{n}\left(\frac{z_i}{\pi_i'}\right) - N\alpha_1\alpha_2^{-1}b, \tag{3.2}$$

$$\mathrm{Var}(\hat{Y}_{RHT(MS)}) = \mathrm{Var}(\hat{Y}_{HT(MS)}) + b^2\delta_1 + 2b\delta_2. \tag{3.3}$$

(3.3)식으로부터 $\mathrm{Var}(\hat{Y}_{RHT(MS)})$를 최소로 하는 b는 다음과 같이 된다.

$$b_{opt} = -\delta_2/\delta_1. \tag{3.4}$$

여기서,

$$\delta_1 = \left(\frac{\alpha_1}{\alpha_2}\right)^2 \sum_{i>j=1}^{N}(\pi_i'\pi_j' - \pi_{ij}')(1/\pi_i' - 1/\pi_j')^2,$$

$$\delta_2 = \left(\frac{\alpha_1}{\alpha_2}\right) \sum_{i>j=1}^{N}(\pi_i'\pi_j' - \pi_{ij}')(y_i/\pi_i' - y_j/\pi_j')(1/\pi_i' - 1/\pi_j').$$

그러면 $\mathrm{Var}_{min}(\hat{Y}_{RHT(MS)}) = \mathrm{Var}(\hat{Y}_{HT(MS)}) - \delta_2^2/\delta_1$이 된다. 미지의 연구변수에 대한 변환으로 b_{opt}를 구할 수 없는 경우에는 표본자료를 사용하여 b를 추정한다.

4. 수치비교

M-S추출법을 적용하기 위해서 1단계로 보조변수를 변환하고 이로부터 얻은 포함확률을 H-T추정량에 사용하여 모집단 총계를 추정할 때 분산을 줄이기 위해서 2단계로 연구변수를 변환하였다. 보조변수와 연구변수의 변환을 통해 얻은 추정량과 기존의 PPSWR, H-T 추정량들과의 비교를 위해서 Yate와 Grundy(1953)가 사용한 표 4.1의 모집단 A, B, C를 사용한다.

3개 모집단 A, B, C에 M-S추출법을 사용하기 위해서는 (2.3)식이 만족되어야 한다. 그러나 단위 1의 추출확률은 $p_i > (n-1)/[n(N-1)] = 0.167$을 만족하지 못하므로 M-S추출법을 사용할 수 없다. 따라서 (2.4)식과 같이 보조변수를 변환한다.

표 4.1: 모집단 특성

단위	$x_i(=p_i)$	모집단(y_i)		
		A	B	C
1	0.1	0.5	0.8	0.2
2	0.2	1.2	1.4	0.6
3	0.3	2.1	1.8	0.9
4	0.4	3.2	2.0	0.8

3개 모집단의 연구변수와 보조변수가 선형관계에 있다고 할 때, (2.4)식의 변환으로 보조변수에 대한 연구변수의 절편이 작아지면 효율이 증가하게 된다. 3개 모집단에 대해 효율을 증가시키는 d의 범주는 0.8과 1사이가 된다(Mohanty(1978)).

$d = 0.85$이고 $d = 0.95$인 경우 수정된 추출확률을 p_1'와 p_2'라하고 이들로부터 얻은 포함확률을 사용하여 구한 추정량들과 기존의 PPS추정량들의 분산을 비교하였다.

표 4.2: 추정량들의 분산 비교

추정량	A			B			C		
	p	p_1'	p_2'	p	p_1'	p_2'	p	p_1'	p_2'
$\hat{Y}_{PPSWR}$	0.5	2.8969	2.8969	0.5	0.1788	0.1788	0.125	0.2294	0.2294
$\hat{Y}_{HT}$	0.8229	2.5106	2.6233	0.0570	0.2386	0.2688	0.0539	0.1994	0.2076
$\hat{Y}_{HT(MS)}$	-	3.9518	4.1078	-	0.6717	0.6902	-	0.3007	0.3042
$\hat{Y}_{RHT(MS)}$	-	0.0535	0.0532	-	0.0535	0.0537	-	0.0853	0.0852

표 4.2로부터,

i) $p_i > 0.167$을 만족하지 못하므로 M-S추출법을 사용할 수 없다.

ii) 보조변수를 변환하였을 때는 $\hat{Y}_{PPSWR}$(모집단 B 제외)와 $\hat{Y}_{HT}$의 분산은 증가한다.

iii) 보조변수만을 변환하여 M-S추출법을 적용한 $\hat{Y}_{HT(MS)}$의 분산은 $\hat{Y}_{PPSWR}$나 $\hat{Y}_{HT}$의 분산보다 크지만, 2단계로 연구변수를 변환한 $\hat{Y}_{RHT(MS)}$의 분산은 다른 추정량들의 분산에 비해 훨씬 작다.

iv) d의 값에는 크게 영향을 받지 않는다.

위의 결과로부터 2단계의 변수변환을 통해서 H-T추정량의 사용이 편리하게 되고 동시에 추정량의 효율도 높아짐을 알 수 있다. 물론 b_{opt} 대신 추정값을 사용하면 분산은 크게 된다.

5. 맺음말

비복원불균등확률추출에 의한 비편향추정량으로 H-T추정량이 널리 사용되고 있다. 그러나 이 추정량의 분산이 음수 값을 가질 수 있고, π_i와 π_{ij}의 계산이 번거로운 단점이 있다. 이러한 단점을 보완해주기 위해서 M-S추출법을 사용한다. M-S추출법은 모집단을 구성하고 있는 단위들의 추출확률이 (2.3)식을 만족해야만 사용 가능한데, 이 조건은 실제 조사에서 M-S추출법을 사용하는데 큰 제약이 된다. 한편 M-S추출법을 사용한 H-T추정량의 분산은 y_i/π_i가 일정한 값을 갖지 않으면 커지게 된다. 기존의 연구에서는 (2.3)식의 조건이 만족되지 않는 경우에 대해 보조변수를 변환하거나 H-T추정량의 분산을 줄이기 위해서 연구변수를 변환하였다.

본 연구에서는 (2.3)식의 조건을 만족시키지 못하는 경우에 대해 1단계로 보조변수를 변환하여 M-S추출법의 사용을 가능하게 하고 동시에 H-T추정량의 분산을 줄이기 위해서 2단계로 연구변수의 변환을 고려하였다. 결과적으로 H-T추정량의 단점을 극복하는 M-S추출법의 사용을 확대시키고 아울러 추정량의 효율을 증대시킬 수 있었다.

참고문헌

[1] Bedi, P. K. and Agarwal, S. K. (1999). Modified Midzuno scheme of sampling, *Journal of Statistical Planning and Inference*, Vol. 76, 203-214.

[2] Hansen, M. H. and Hurwitz, W. N. (1943). On the theory of sampling from finite populations, *Annals of Mathematical Statistics*, Vol. 14, 333-362.

[3] Horvitz, D. G. and Thompsom, D. J. (1952). A generalization of sampling without replacement from a finite universe, *Journal of the American Statistical Association*, Vol. 47, 663-685.

[4] Kumar, E. V. and Srivenkataramana, T. (1994). A generalization of Midzuno-Sen sampling scheme for finite populations, *Communications in Statistics - Theory and Methods*, Vol. 23, No. 9, 2541-2559.

[5] Midzuno, H. (1952). On the sampling system with probability proportionate to sum of sizes, *Annals of the Institute of Statistical Mathematics(Tokyo)*, Vol. 13, 99-107.

[6] Mohanty, S. (1978). Use of transformation in PPS sampling, *Journal of the Indian Society of Agricultural Statistics*, Vol. 30, No. 1, 127-132.

[7] Prasad, N. G. N. and Srivenkataramana, T. (1980). A modification to the Horvitz-Thompson estimator under the Midzuno sampling scheme, *Biometrika*, Vol. 67, No. 3, 709-711.

[8] Rao, T. J. (1988). Transformation on the auxiliary variate for Midzuno-Sen sampling

scheme, *Journal of the Indian Society of Agricultural Statistics*, Vol. 40, No. 3, 173-177.

[9] Reddy, V. N. (1974). On a transformed ratio method of estimation, *Sankhya*, Vol. 36, Series C. Pt. 1, 59-70.

[10] Reddy, V. N. and Rao, T. J. (1977). Modified pps method of estimation, *Sankhya*, Vol. 39, Series C. Pt. 3, 185-197.

[11] Sahoo, J. (1997). A note on using location shift in the Midzuno-Sen scheme of sampling with a transformed variable, *Statistics and Probability Letters*, Vol. 33, 285-290.

[12] Sen, A. R. (1953). On the estimate of the variance in sampling with varying probabilities, *Journal of the Indian Society of Agricultural Statistics*, Vol. 5, 119-127.

[13] Srivenkataramana, T. and Tracy, D. S. (1979). Transforming the study variate after pps sampling, *Metron*, Vol. 37, 175-181.

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

[2003년 2월 접수, 2003년 7월 채택]

A Modified Horvitz-Thompson Estimator by Transformation of Variables

Jea-Bok Ryu [1)]

ABSTRACT

The Horvitz-Thompson(H-T) estimator is less efficient than PPS estimators in some cases. We use the two-stage variable transformation in order to remove the drawbacks and increase the efficiency of H-T estimator. We transform the auxiliary variable to use the Midzuno-Sen sampling scheme at the first stage. And the next stage, we also transform the study variable to reduce the variance of H-T estimator using the inclusion probability obtained from the first transformation. We compare the efficiency between a suggested modified H-T estimator and PPS estimators.

Keywords: Midzuno-Sen sampling scheme; Inclusion probability; Transformation of variable; Horvitz-Thompson estimator.

1) Professor, Dept. of Statistics, Chongju University, Chongju, 360-764, Korea.
E-mail : jbryu@chongju.ac.kr