응용통계연구 제 **14**권 *1*호, *2001*년, *pp. 13–24*

이중추출에서 모평균 추정

김규성[1] 김진석[2] 이선순[3]

요 약

이중추출에서 모평균 추정방법을 고찰하였다. 전통적으로 널리 쓰이는 비추정량과 회귀추정량 그리고 비례배분 및 Rao 배분을 한 후의 층화평균에 대하여 주어진 기대 비용에서 최적의 표본수, 최소분산 및 분산추정량을 살펴보았다. 또한 비추정 및 층화의 효과를 모두 내포하는 결합비 추정량을 제안하고 주어진 기대 비용에서 최적의 표본수 및 최소분산을 유도하였고 분산추정량을 구하였다. 그리고 제한된 모의실험을 통하여 비추정량, 층화평균 및 결합비 추정량의 효율을 비교하였다. 모의실험 결과 비추정량과 층화평균은 경우에 따라 효율이 다르게 나타난 반면, 결합비 추정량은 대체로 두 방법보다 효율이 우수하게 나타나 결합비 추정량이 이중추출에 유용하게 쓰일 수 있음을 보였다.

주요용어: 결합비 추정량, 비추정량, 이중추출, 층화평균.

1. 서론

통계조사에서 관심변수의 모평균을 추정하고자 할 때 관심변수와 연관이 깊은 보조변수가 있는 경우, 보조변수를 이용하여 추정의 효율을 높이는 방법이 여러 방면으로 개발되어 왔다. 보조변수를 이용하는 방법들은 크게 두 가지로 분류할 수 있는데, 하나는 표본추출의 효율을 높이는 데 이용하는 것으로 층화추출법과 확률비례추출법이 대표적이며, 다른 하나는 추정량의 효율을 높이는 데 이용하는 것으로 비추정법이나 회귀추정법이 대표적이다. 그런데 통상적인 의미에서의 이 방법들은 모두 모집단 개개 단위 전체에서 보조변수값을 취할 수 있음을 전제로 하고 있거나 혹은 부분모집단의 특성값을 알고 있어야 함이 전제되어 있다. 따라서 위의 방법들은 보조변수 값들이 사전에 수집되어 있어야 효과적으로 이용될 수 있다.

이러한 보조변수의 구체적인 자료를 사전에 가지고 있지 않을 때, 유용하게 이용할 수 있는 방법이 이중추출법(two-phase sampling 혹은 double sampling)이다. 이 방법은 Neyman (1938)에 의해 처음 제안된 이래 지금까지 표본조사에서 널리 쓰이고 있다. 이중추출법 에서는 관심변수를 조사하기 전에 보조변수를 조사하는 단계가 선행되는데, 우선 다소 큰 1차 표본을 추출하여 보조변수를 조사하고, 1차 표본 중에서 일부를 2차 표본으로 선정을 하여 관심변수의 값을 얻는 것이다. 주어진 총비용에서 관심변수를 조사하는데 쓰일 비용의 일부가 보조변수 획득에 쓰이기 때문에 관심변수를 일부 조사하지 못하는 데서 발생하는 손

1) (130-743) 서울시 동대문구 전농동 90, 서울시립대학교 컴퓨터·통계학과, 부교수
E-mail: kskim@uoscc.uos.ac.kr
2) (151-742) 서울시 관악구 신림동 산 56-1, 서울대학교 통계학과, 박사과정
3) (151-742) 서울시 관악구 신림동 산 56-1, 서울대학교 통계학과, 박사과정

실을 추가로 조사된 보조변수가 보충해줄 수 있을 때 이중추출법은 의미를 지닌다. 따라서 이중추출법에서는 이 방법을 이용할지 여부에 대한 문제와, 또한 1차 표본수와 2차 표본수를 결정하는 문제가 중요한 관건이 된다. 통상적으로 1차 표본에서 보조변수 획득에 소비되는 비용이 적고, 2차 표본에서 관심변수의 정보 획득에 필요한 비용이 상대적으로 클 때 이중추출법은 효과적이다.

이중추출법에 의하여 수집된 보조변수는 2차 표본을 추출하기 위한 층화변수로 사용되거나, 비추정량 혹은 회귀추정량의 보조변수로 사용되어 추정의 효율을 높이는 것이 통상적으로 널리 알려진 방법들이다. 2절에서는 이제까지 알려진 이중추출법에서의 모평균 추정방법들을 비교·분석한다. 3절에서는 보조변수를 층화와 추정량에 동시에 이용하여 추정의 효율을 높이는 방법을 제안한다. 층화 후 비추정량을 이용하는 방법은 결합비 추정량이나 분리비 추정량이 이미 소개되어 있으나 이중추출법에서는 소개된 바 아직 없다. 본 논문의 3절에서는 이중추출에서 결합비 추정량을 이용하여 모평균을 추정하는 방법을 소개하고, 제안된 추정량의 분산과 최소분산을 유도하여 그 성질을 살펴보며 모의실험을 통하여 효율성을 확인해 본다. 마지막으로 4절에서는 간단한 요약과 토의를 할 것이며 향후 연구과제를 살펴보기로 한다.

2. 이중추출에서 모평균 추정

2.1. 비추정량을 이용한 모평균 추정

이중추출법에서 보조변수를 비추정량 구성에 이용하는 방법을 먼저 살펴본다. 크기 N인 모집단에서 뽑힌 1차 단순확률표본을 s'이라하고 크기를 n'이라 하며, 1차 표본에서 단순확률추출된 2차 표본을 s로 표시하고 크기는 n이라 하자. 1차 표본에서는 보조변수 x를 수집하고 2차 표본에서는 관심변수 y를 조사하여 모평균 추정량으로 비추정량을 고려하자.

$$\bar{y}_R = \left(\frac{\bar{y}}{\bar{x}}\right)\bar{x}' = \hat{R}\bar{x}' \tag{2.1}$$

여기서 $\bar{x}$, $\bar{y}$는 2차 표본 s에서의 표본평균이며, $\bar{x}'$는 1차 표본 s'에서의 표본 평균이다. 비추정량 $\bar{y}_R$은 모평균에 대한 설계기반 일치추정량(design based consistent estimator)이며, 설계기반 근사분산은 다음과 같다. (Cochran, 1977).

$$Var\{\bar{y}_R\} \approx \left(\frac{1}{n'} - \frac{1}{N}\right)S_y^2 + \left(\frac{1}{n} - \frac{1}{n'}\right)S_d^2 \tag{2.2}$$

여기서 $(N-1)S_y^2 = \sum_{k=1}^{N}(y_k - \bar{Y})^2$, $(N-1)S_d^2 = \sum_{k=1}^{N}(y_k - Rx_k)^2$ 그리고 $R = \bar{Y}/\bar{X}$이다.

이중추출에서 비추정량의 유용성은 비용함수를 통하여 알아볼 수 있다. 1차 표본의 단위당 비용을 c_0라 하고 2차 표본의 단위당 비용을 c라 할 때, 선형비용함수 $C^* = c_0 n' + cn$ 을 고려하자. 비추정량의 분산(2.2)를 최소로 하는 최적의 1차 표본수와 2차 표본수는 각각 $n_{opt}^{'} = C^*/(c_0 + \sqrt{c_0 c}S_d/S_m)$과 $n_{opt} = C^*/(c + \sqrt{c_0 c}S_m/S_d)$ 이며, 이 때 비추정량의 최소

분산은

$$V_R = \frac{1}{C^*}[\sqrt{c}S_d + \sqrt{c_0}S_m]^2 - \frac{S_y^2}{N} \tag{2.3}$$

이다. 여기서 $S_m^2 = |S_y^2 - S_d^2|$ 이다.

비용 C^*가 주어지면 표본평균에 이용되는 표본수는 $n = C^*/c$ 이므로 표본평균의 분산 $(1/n - 1/N)S_y^2$ 다음과 같이 표현이 가능하며,

$$V_S = \Big(\frac{c}{C^*} - \frac{1}{N}\Big)S_y^2 \tag{2.4}$$

따라서 비추정량이 표본평균보다 효율적일 조건은

$$\frac{c}{c_0} > \Big(\frac{S_m}{S_y - S_d}\Big)^2 \tag{2.5}$$

임을 알 수 있다. 즉, 2차 표본의 단위당 비용 c가 클 때, 혹은 관심변수와 비추정치의 차이 $d_k = y_k - Rx_k$의 산포 S_d^2가 작을 때 비추정량은 효과가 있다.

식(2.2)에 주어진 비추정량의 근사분산에 대한 일치추정량으로는 통상적으로 다음이 이용된다.

$$v_0(\bar{y}_R) = \Big(\frac{1}{n'} - \frac{1}{N}\Big)s_y^2 + \Big(\frac{1}{n} - \frac{1}{n'}\Big)S_d^2 \tag{2.6}$$

여기서 s_y^2, s_d^2은 2차 표본 s에서 구한 y와 d의 표본분산들이다. 위의 식 (2.6)에 제시된 추정량은 설계일치성을 갖기는 하지만 $s' - s$에 포함된 보조변수 x를 이용하지 않기 때문에 개선의 여지가 있다. Rao와 Sitter(1995)는 이점에 착안하여 개선된 분산추정량을 제안하였으며, 또한 잭나이프 방법을 이용한 잭나이프 분산추정량을 제안하였다.

$$v_J(\bar{y}_R) = \Big(\frac{\bar{x}'}{\bar{x}}\Big)^2\Big(\frac{1}{n} - \frac{1}{N}\Big)s_d^2 + 2\Big(\frac{\bar{x}'}{\bar{x}}\Big)\Big(\frac{1}{n'} - \frac{1}{N}\Big)\hat{R}s_{dx} + \Big(\frac{1}{n'} - \frac{1}{N}\Big)\hat{R}^2 s_x'^2 \tag{2.7}$$

여기서 s_{dx}는 x와 d의 s에서의 표본 공분산이며, $s_x'^2$는 x의 s' 에서의 표본분산이다. 분산추정량 v_0는 2차 표본이 치우쳐서 추출될 경우, 즉 $\bar{x}$와 $\bar{x}'$가 현저하게 다를 경우 과소 혹은 과대 추정하는 경향이 있는데 반해, 잭나이프 분산추정량 v_J는 분산식에 $\bar{x}'/\bar{x}$을 포함하고 있어 분산이 편향되게 추정될 가능성이 작다.

이중추출에서 회귀추정량은 $\bar{y}_g = \bar{y} + b(\bar{x}' - \bar{x})$, b는 2차 표본 s에서의 회귀계수, 이며 근사분산은 (2.2)의 식과 동일하게 주어지나 회귀추정량에서는 $S_d^2 = S_y^2\,(1 - \rho^2)$인점이 다르다. 여기서 ρ는 관심변수와 보조변수의 상관계수이다. 또한 최적의 표본수에서 최소분산을 구하여 표본평균의 분산과 비교해 보면, 동일한 비용에서 회귀추정량이 표본평균보다 효과적일 조건은 $\rho^2 > 4(c/c_0)/(1 + c/c_0)^2$임을 알 수 있다. 이는 관심변수와 보조변수의 상관관계가 크거나 2차 표본의 단위당 비용 c가 클 때 회귀추정량은 효과적임을 나타내준다. 회귀추정량의 통상적인 설계일치 분산추정량은 $v_g = (1/n' - 1/N)b^2 s_x^2 + (1/n - 1/N)s_d^2$ 이며, 이를 개선한 분산추정량과 잭나이프 분산추정량이 Sitter(1997)에 의해 제안되었다. 비추정량과 마찬가지로 Sitter가 제안한 잭나이프 분산추정량은 설계일치 추정량이 되며, 조건부 성질이 통상적인 설계일치 추정량보다 우수하다.

이중추출법은 종종 무응답 처리에 이용되는데, 그 이유는 2차 표본을 응답표본으로 간주하면 그 구조가 동일하기 때문이다. (Särndal과 Swensson, 1987; Rao와 Sitter, 1995). 차이점은 이중추출에서는 2차 표본추출확률구조를 알고 있는 반면, 무응답 문제에서는 그 확률구조를 모른다는 것이다. 무응답 단위에 비대체(ratio imputation)을 하는 경우 대체 후 표본평균은 비추정량과 동일하게 되어 이중추출법에서 얻은 결과들을 그대로 이용할 수 있는 편리함이 있다. 이 때 설계일치분산추정량 v_R는 응답패턴이 관심변수와 무관할 때 이용되면 효과적이나 그렇지 않을 때는 편향이 커서 효과적이지 못하다. 잭나이프 분산추정량 v_J는 어느 경우에도 편향의 정도가 작게 나타나는 경향을 보인다.

2.2. 층화평균을 이용한 모평균 추정

관심변수 조사를 위한 층화에 보조변수를 이용할 수 있다. 1차 표본 s'에서 조사한 보조변수를 이용하여 1차 표본을 L개의 층으로 층화하고 층 가중값 $w_h = n'_h/n'$을 구한다. 그리고 각 층에서 2차표본을 독립적으로 추출하여 관심변수를 조사한 후 층화평균 $\bar{y}_{st}$ 를 만든다.

$$\bar{y}_{st} = \sum_{h=1}^{L} w_h \bar{y}_h \tag{2.8}$$

여기서 $\bar{y}_h = \sum_{k\in s_h} y_{hk}/n_h,\ h = 1, ..., L$,은 2차 표본 $s = \cup_{h=1}^{L} s_h$의 각 층에서의 층내 평균이다. 층 가중값 w_h가 모집단 층 가중값의 비편향 추정량이기 때문에 층화평균 $\bar{y}_{st}$는 모평균의 비편향추정량이 된다. 그런데 위에서 언급한 대로 표본을 추출하면 일차 표본수 n'과 2차 표본수 $n = \sum_{h=1}^{L} n_h$는 사전에 정해줄 수 있는 값이나, 일차 표본의 층내 표본수 n'_h는 사전에 알지 못하는 값임에 유의할 필요가 있다. 따라서 2차 표본수 n을 각 층에 배분하기 위해서는 표본수 배분에 대한 기준이 필요하다. 보통 많이 이용되는 표본배분법으로는 비례배분법과, 즉 $n_h = n \times w_h$, Rao(1973) 배분법, $n_h = \nu_h \times n'_h$, ν_h는 사전에 알려진 상수, 이 있다.

2차 표본을 비례배분 했을 때 층화평균의 분산은

$$Var\{\bar{y}_{st}\} = \Big(\frac{1}{n'} - \frac{1}{N}\Big) S_y^2 + \Big(\frac{1}{n} - \frac{1}{n'}\Big) \sum_{h=1}^{L} W_h S_{yh}^2 \tag{2.9}$$

이다. 여기서 W_h 는 모집단의 층 비율이며, S_{yh}^2는 관심변수의 h층에서의 모분산이다. 층화평균의 효율은 선형비용함수 $C = c_0 n' + \sum_{h=1}^{L} c_h n_h$ 를 가정하고 표본평균과 비교해 볼 수 있는데, 이 경우 1차 층내 표본수 n'_h가 변수이기 때문에 총 비용 C가 고정된 값이 아니다. 따라서 기대 총 비용을 고려해 볼 수 있으며, 비례배분을 하면 기대 총비용은 $C^* = E\{C\} = c_0 n' + cn,\ c = \sum_{h=1}^{L} W_h c_h$ 이다. 고정된 기대 총 비용에서 (2.9)을 최소로 하는 최적의 표본수는 각각 $n'_{opt} = C^*/(c_0 + \sqrt{c_0 c} S_{yW}/S_{yB})$, $n_{opt} = C^*/(c + \sqrt{c_0 c} S_{yB}/S_{yW})$ 이며, 이 때 층화평균의 최소분산은

$$V_{EP} = \frac{1}{C^*}(\sqrt{c_0} S_{yB} + \sqrt{c} S_{yW})^2 - \frac{S_y^2}{N} \tag{2.10}$$

이다. 여기서 $S_{yB}^2 = S_y^2 - S_{yW}^2$, $S_{yW}^2 = \sum_{h=1}^{L} W_h S_{yh}^2$. 또한 최소분산 V_{EP}가 표본평균의 분산 V_S 보다 작아질 조건은

$$\frac{c}{c_0} > \left(\frac{S_{yB}}{S_y - S_{yW}}\right)^2 \tag{2.11}$$

임을 알 수 있다. 따라서 2차 표본의 단위당 비용이 높거나 층내 분산 S_{yW}^2이 작고 층간분산 S_{yB}^2이 클 때 층화 평균은 효과적이다.

Rao 배분했을 때 층화평균의 분산은

$$Var\{\bar{y}_{st}\} = \left(\frac{1}{n'} - \frac{1}{N}\right)S_y^2 + \frac{1}{n'}\sum_{h=1}^{L} W_h S_{yh}^2\left(\frac{1}{\nu_h} - 1\right) \tag{2.12}$$

이다. 선형비용함수의 기대 총비용은 $C^* = E\{C\} = c_0 n' + n' \sum_{h=1}^{L} c_h \nu_h W_h$이고, 고정된 기대 총 비용에서 분산 (2.12)를 최소화 하는 최적의 표본수와 상수 ν_h는 각각 $n'_{opt} = C^*/(c_0 + \sqrt{c_0}\sum_{h=1}^{L} W_h\sqrt{c_h}S_{yh}/S_{yB})$과 $\nu_{h,opt} = (S_{yh}\sqrt{c_0})/(S_{yB}\sqrt{c_h})$이며, 이 때의 최소 분산은

$$V_{ER} = \frac{1}{C^*}(\sqrt{c_0}S_{yB} + \sum_{h=1}^{L} W_h S_{yh}\sqrt{c_h})^2 - \frac{S_y^2}{N} \tag{2.13}$$

이다. 또한 최소분산 V_{ER}이 표본평균의 분산 V_S 보다 작아질 조건은

$$\frac{c}{c_0} > \left(\frac{S_{yB}}{S_y - \sum_{h=1}^{L} W_h S_{yh}\sqrt{c_h}/\sqrt{c}}\right)^2 \tag{2.14}$$

이며, Rao 배분 후 층화 평균의 최소분산은 비례배분 후 최소 분산보다 작거나 같다. 이상에서 알 수 있듯이 이중추출에서 층화평균이 표본평균보다 항상 효과적이라고는 할 수 없으며, 비용과 보조변수의 유용성 여부에 그 효과가 따라 달라진다.

식 (2.12)에 주어진 층화평균의 분산에 대한 비편향 추정량의 근사식은 만일 1차 표본수가 충분히 크면 Cochran(1977)으로부터 다음과 같다.

$$v(\bar{y}_{st}) = \sum_{h=1}^{L} w_h s_h^2\left(\frac{1}{n'\nu_h} - \frac{1}{N}\right) + \frac{g'}{n'}\sum_{h=1}^{L} w_h\left(\bar{y}_h - \bar{y}_{st}\right)^2 \tag{2.15}$$

여기서 $g' = (N - n')/(N - 1)$이다.

3. 결합비추정량을 이용한 모평균 추정

3.1. 이중추출에서 결합비 추정량

이제까지 앞 절에서 이중추출에서 보조변수를 비추정량이나 층화에 이용하는 방법을 살펴보았다. 이 절에서는 두 방법을 결합한 더 효과적인 모평균 추정법에 대하여 알아본다.

일반적으로 비추정량이 표본평균보다 효과적이려면, CV_x , CV_y가 x 와 y의 변동계수라 할 때 $\rho_{xy} > CV_x/(2CV_y)$을 만족해야 한다. 또한 관심변수 y와 보조변수 x의 관계가 선형이고 y의 분산이 x에 비례할 때, 즉 $y = \beta x + \epsilon$, $\epsilon \sim (0, x\sigma^2)$일 때, 비추정량은 최량선형

비편향추정량임이 알려져 있다. 그런데 만일 x와 y의 관계가 $E\{y\} = \beta x^g$, $g < 1$ 이라 하면 비추정량이 표본평균보다 효과적이라 하더라도 비추정량은 최선의 추정량이 아니므로 개선의 여지가 있을 수 있다. 따라서 층화의 효과를 비추정량에 가미한 결합비 추정량이나 혹은 분리비 추정량을 이용하면 추정의 효율을 높일 수 있을 것으로 예상된다. 이 점에 착안하여 비추정량이 표본평균보다는 효과적이지만 개선의 여지가 있을 때 이용할 수 있는 결합비 추정량을 제안한다.

이중추출 후에 결합비 추정량 $\bar{y}_{CR}$을 고려하자.

$$\bar{y}_{CR} = \Big(\frac{\bar{y}_{st}}{\bar{x}_{st}}\Big)\bar{x}' \tag{3.1}$$

1차 표본 s'에서 x 와 y의 표본평균을 각각 $\bar{x}'$, $\bar{y}'$라고 할 때 결합비 추정량은 $\bar{y}_{CR} \approx \bar{y}' + (\bar{y}_{st} - (\bar{y}'/\bar{x}')\bar{x}_{st})$로 전개가 가능하므로 비추정량과 유사하게 기대값 및 분산을 구할 수 있다. 결과적으로 제안된 결합비 추정량은 모평균의 설계일치 추정량이며, 비례배분했을 때 결합비 추정량의 근사분산을 얻을 수 있다.

$$Var\{\bar{y}_{CR}\} \approx \Big(\frac{1}{n'} - \frac{1}{N}\Big)S_y^2 + \Big(\frac{1}{n} - \frac{1}{n'}\Big)\sum_{h=1}^{L} W_h S_{dh}^2 \tag{3.2}$$

또한 선형비용함수 $C = c_0 n' + \sum_{h=1}^{L} c_h n_h$ 를 가정하고 기대 총비용 $C^* = E\{C\} = c_0 n' + cn$을 고정했을 때, 분산 (3.2)를 최소로 하는 최적의 표본수는 각각 $n'_{opt} = C^*/(c_0 + \sqrt{c_0 c}\, S_{dW}/S_{dB})$ 와 $n_{opt} = C^*/(c + \sqrt{c_0 c}\, S_{dB}/S_{dW})$임을 보일 수 있으며, 이 때의 결합비 추정량의 최소분산으로 다음을 얻는다.

$$V_{CP} = \frac{1}{C^*}(\sqrt{c_0}S_{dB} + \sqrt{c}S_{dW})^2 - \frac{S_y^2}{N} \tag{3.3}$$

여기서 $S_{dW}^2 = \sum_{h=1}^{L} W_h S_{dh}^2$, S_{dh}^2는 $d_k = y_k - Rx_k$의 층내 분산이며, $S_{dB}^2 = S_y^2 - S_{dW}^2$ 이다. 그리고 최소분산 V_{CP}가 표본평균의 분산 V_S보다 작아질 조건은

$$\frac{c}{c_0} > \Big(\frac{S_{dB}}{S_y - S_{dW}}\Big)^2 \tag{3.4}$$

임을 알 수 있다.

Rao 배분했을 때 결합비 추정량의 분산은 다음과 같다.

$$Var\{\bar{y}_{CR}\} = \Big(\frac{1}{n'} - \frac{1}{N}\Big)S_y^2 + \frac{1}{n'}\sum_{h=1}^{L} W_h S_{dh}^2\Big(\frac{1}{\nu_h} - 1\Big) \tag{3.5}$$

이다. 기대 총비용, $C^* = E\{C\} = c_0 n' + n' \sum_{h=1}^{L} c_h \nu_h W_h$을 고정했을 때 분산 (3.5)를 최소로 하는 최적의 표본수와 주어진 상수ν_h 는 각각 $n'_{opt} = C^*/(c_0 + \sum_{h=1}^{L} \sqrt{c_0 c_h}\, W_h S_{dh}/S_{dB})$과 $\nu_{h,opt} = S_{dh}\sqrt{c_0}/(S_{dB}\sqrt{c_h})$이고, 이 때의 최소분산으로 다음을 얻는다.

$$V_{CR} = \frac{1}{C^*}(\sqrt{c_0}S_{dB} + \sum_{h=1}^{L} W_h S_{dh}\sqrt{c_h})^2 - \frac{S_y^2}{N} \tag{3.6}$$

또한 최소분산 V_{CR}이 표본평균의 분산 V_S보다 작아질 조건을 구하면

$$\frac{c}{c_0} > \Big(\frac{S_{dB}}{S_y - \sum_{h=1}^{L} W_h S_{dh}\sqrt{c_h}/\sqrt{c}}\Big)^2 \tag{3.7}$$

이다.

식 (3.5)에 주어진 결합비 추정량의 분산의 설계기반 비편향 추정량은 층화평균의 분산 추정량을 이용하여 쉽게 구할 수 있다. 결합비 추정량의 분산과 층화평균의 분산의 관계로부터 다음을 얻고, $Var\{\bar{y}_{CR}\} = Var\{\bar{y}_{st}\} + \sum_{h=1}^{L}(1/\nu_h - 1)W_h(S_{dh}^2 - S_{yh}^2)/n'$, 우변 두 번째 항을 층별로 각각 추정하여 대입하고 정리하면 다음과 같은 설계기반 근사 비편향 추정량을 얻을 수 있다.

$$v(\bar{y}_{CR}) = \sum_{h=1}^{L} w_h s_h^2\Big(\frac{1}{n'} - \frac{1}{N}\Big) + \sum_{h=1}^{L} w_h s_{dh}^2 \frac{1}{n'}\Big(\frac{1}{\nu_h} - 1\Big) + \frac{g'}{n'}\sum_{h=1}^{L} w_h(\bar{y}_h - \bar{y}_{st})^2 \tag{3.8}$$

3.2. 모의실험

이중추출에서 모평균 추정량들의 유용성을 알아보기 위하여 모의실험을 실시하였다. 데이터는 다음의 모형에서 생성하였다.

$$y = \beta x^g + \epsilon,\ \epsilon \sim (0, x\sigma^2) \tag{3.9}$$

보조변수 x는 감마분포에서 생성하고 x의 누승은 $0.5 < g < 1$을 고려하며 오차는 정규분포에서 생성하였다. 1,000개의 데이터(x, y)를 발생하여 x의 크기순으로 정렬한 후 층화를 하였는데, 층이 두 개인 경우는 W_1=W_2=0.5, 층이 세 개인 경우는 W_1=0.3, W_2=0.4, W_3=0.3 그리고 층이 네 개인 경우는 W_1=0.2, W_2=W_3=0.3, W_4=0.2로 층의 비율을 정하였다. 비용함수는 선형비용함수를 이용했으며 총비용은 C^* =300, 일차 추출비용은 $c_0 = 1$, 그리고 이차 층내 추출비용은 동일한 것으로 하여 c= $\sum_{h=1}^{L} W_h c_h = 1, 2, 5, 10$을 고려하였다.

비교대상이 되는 추정량은 표본평균을 비롯하며 층화평균, 비추정량 그리고 결합비 추정량이며, 층화를 했을 때의 표본수 배정은 비례배분과 Rao 배분을 고려하였다. 각 추정량의 효율을 비교하기 위하여 주어진 비용에서 최소분산을 구한 후, 설계효과를 계산하였다. 설계효과는 단순확률표본에서 구해진 표본평균의 분산, Var_{SRS} , 과 복합표본에서 구해진 추정량의 분산, Var_C,을 비교하는 것으로 다음의 식으로 정의된다.

$$\mathrm{DEFF}(\bar{y}_C) = \frac{Var_C\{\bar{y}_C\}}{Var_{SRS}\{\bar{y}_s\}}100 \tag{3.10}$$

따라서 설계효과가 100일 때 추정량 $\bar{y}_c$의 분산은 표본평균과 분산이 동일하며 100보다 작으면 추정량 $\bar{y}_c$가 표본평균 보다 효과적이라고 할 수 있고 100보다 크면 표본평균보다 비효과적이라 할 수 있다. 표 3.1에 5가지 추정량의 설계효과가 주어져 있다. 모의실험은 여러 경우를 실시했으나 $g = 0.7, 0.9$인 경우만 모의실험의 결과로 제시한다.

표 3.1: 이중추출에서 비추정량, 층화평균, 결합비추정량의 설계효과

모형	층수	단위비용(c)	비추정량	층화평균		결합비추정량	
				비례배분	Rao배분	비례배분	Rao배분
$g = 0.7$	2	1	225.818	242.767	220.736	218.821	196.216
		2	129.925	152.683	137.298	123.452	108.469
		5	79.194	103.485	91.787	73.394	62.760
		10	60.348	84.545	74.289	54.969	46.088
	3	1	225.818	234.862	207.528	209.663	185.882
		2	129.925	139.420	120.865	115.598	100.018
		5	79.194	88.324	74.720	66.697	55.821
		10	60.348	69.082	57.449	48.902	39.933
	4	1	225.818	224.925	194.052	203.873	180.906
		2	129.925	129.068	108.474	110.884	95.935
		5	79.194	78.410	63.662	62.826	52.491
		10	60.348	59.614	47.216	45.464	37.002
$g = 0.9$	2	1	161.040	242.833	206.578	159.530	153.379
		2	79.542	154.301	128.923	78.518	74.679
		5	39.302	105.792	86.438	38.590	36.103
		10	25.594	87.051	70.051	25.021	23.093
	3	1	161.040	235.233	192.199	157.246	150.275
		2	79.542	139.857	110.721	76.977	72.639
		5	39.302	88.768	67.479	37.524	34.728
		10	25.594	69.516	51.356	24.168	22.008
	4	1	161.040	223.925	178.414	156.058	148.776
		2	79.542	128.121	97.917	76.180	71.655
		5	39.302	77.548	56.074	36.976	34.065
		10	25.594	58.810	40.851	23.730	21.487

모의실험의 결과는 다음과 같이 요약된다. $g = 0.7$과 $g = 0.9$인 모의실험의 대부분의 경우에서 결합비 추정량의 설계효과가 비추정량이나 층화평균의 설계효과보다 작게 나타나 결합비 추정량이 비추정량이나 층화평균보다 효과적임을 알 수 있다. 이는 결합비 추정량이 비추정량의 효과와 층화의 효과를 동시에 내포하고 있기 때문에 나타나는 현상이다. 비추정량과 층화평균의 비교는 모형에 따라 우열이 바뀌는데, $g = 0.7$ 인 모형에서는 비추정량의 분산이 층화평균의 분산보다 대체로 크며 $g = 0.9$ 인 모형에서는 반대로 비추정량의 분산이 대체로 작게 나타난다.

구체적인 경향을 보면, 첫째, 이중추출이 단순확률추출보다 효과적이려면 2차 단위 비용 c_h가 1차 단위 비용 c_0다 상대적으로 커야 한다. 2차 단위 비용이 1차 단위 비용과 동일한 경우($c = 1$)는 모두 2중 추출의 분산이 더 크게 나타나며 2차 비용이 1차 비용의 2배 인 경우도($c = 2$) 대체로 이중 추출의 분산이 더 크다. 2차 단위의 비용이 클수록 2중 추출의 효율은 증가한다. 둘째, 층의 수가 증가할수록 층화평균 및 결합비 추정량의 효율은 증가하는데, 이는 층화의 효과가 커지기 때문이다. 또한 이미 알려진 대로 동일조건에서는 Rao 배분의 효율이 비례배분의 효율보다 높게 나타났다.

모의실험에 이용된 모형, $E\{y\} = \beta x^g$, $Var\{y\} = x\sigma^2$, 에서 추정량들의 대체적인 비교는 다음과 같이 요약된다. $g = 1$인 경우는 비추정량이 최량선형추정량(Best linear unbiased estimator)이므로 비추정량이 가장 우수하게 나타났으며, 또한 $g > 1$인 경우도 비추정량이 우수하게 나타났다. 반면에 $g < 0.5$인 경우도 대체로 층화평균이 우수하게 나타났고, 그 사이구간인 $0.5 < g < 1$ 에서는 대체로 결합비 추정량의 효율이 대체로 높게 나타났다. 이 결과는 모형에 따라 층화평균, 비추정량 그리고 결합비 추정량의 효율이 바뀔 수 있다는 뜻이며, 적절하게 결합비 추정량을 사용해야 추정의 효율을 높일 수 있음을 말해준다.

4. 토의 및 결론

본 논문에서는 이중추출에서 모평균 추정방법을 고찰하였다. 전통적으로 널리 쓰이는 비추정량과 회귀추정량 그리고 층화평균에 대하여 주어진 기대 비용에서 최적의 표본수, 최소분산 및 분산추정량을 알아보았다. 또한 비추정 및 층화의 효과를 모두 반영하는 결합비 추정량을 제안하고 주어진 기대 비용에서 최적의 표본수, 최소분산 및 분산추정량을 유도하였으며 제한된 모의실험을 통하여 효율을 비교하였다. 모의실험 결과 비추정량과 층화평균은 경우에 따라 효율이 다르게 나타난 반면에 결합비 추정량은 대체로 두 방법보다 효율이 우수하게 나타나 이중추출에서 결합비 추정량이 유용하게 쓰일 수 있음을 보였다.

비추정 및 층화의 효과를 내포하는 또 다른 추정량인 분리비 추정량을 이중추출에 이용할 수도 있을 것이다. 이 경우 분리비 추정량은

$$\bar{y}_{SR} = \sum_{h=1}^{L} \left(\frac{\bar{y}_h}{\bar{x}_h}\right) \bar{x}'_h \tag{4.1}$$

와 같이 주어지며 최적의 표본수 및 최소분산의 유도는 결합비 추정량과 유사하게 얻을 수 있을 것이다. 효율 또한 결합비 추정량과 비슷할 것으로 예상된다.

앞에서 고찰한 결과들은 알려진 모집단에서 얻은 것이므로 이 결과를 실제문제에 적용하기 위해서는 몇 가지 작업이 뒤따라야 한다. 우선 비용함수를 설정하고 단위당 비용c_0, c_h를 정하는 일이 선행되어야 하며 표본수를 정하기 위해서는 모집단 특성값, S_y^2 , S_d^2, S_{yW}^2 등이 추정되어야 한다. 따라서 추정된 특성값을 기반으로 한 모평균 추정량들의 효율은 알려진 모집단에서의 그것과 다르게 나타날 수 있다.

본 논문에서는 표본을 단순확률추출한 경우만을 다루었으나 더 일반적인 표본추출을 고려하여 이중추출의 방법을 일반화시킬 수도 있을 것이다. (Särndal, et. al., 1994, 9장). 또한 표본배분에도 일반적인 배분방법을 생각해 볼 수 있을 것이다. (Amstrong과 Wu, 1992). 그리고 이중추출은 보조정보를 활용하는 표본추출 및 추정방법이므로 구체적인 문제에 따라, 보조변수의 성질에 따라 다양한 기법이 개발될 수 있을 것으로 보인다.

참고문헌

[1] Amstrong, J.B. and Wu, C.F.J. (1992). A sample allocation method for two-phase survey designs. *Survey Methodology*, 18, 253-262.

[2] Cochran, W. G. (1977). *Sampling techniques*. 3rd ed. Wiley.

[3] Neyman, J. (1938). Contribution to the theory of sampling human population. *Journal of the American Statistical Association*, 33, 101-116.

[4] Rao, J.N.K. (1973). On double sampling for stratification and analytical surveys. *Biometrika*, 60, 125-133.

[5] Rao, J.N.K. and Sitter, R.R. (1995). Variance estimation under two-phase sampling with application to imputation for missing data. *Biometrika*, 82, 453-460.

[6] Särndal, C.E. Swensson, B. and Wretman, J. (1994). *Model assisted survey sampling*. Springer-Verlag.

[7] Särndal, C.E. and Swensson, B. (1987). A general view of estimation for two phases of selection with applications to two-phase sampling and nonresponse. *International statistical review*, 55, 279-294.

[8] Sitter, R.R. (1997). Variance estimation for the regression estimator in two-phase sampling. *Journal of the American Statistical Association*, 92, 780-787.

[9] Smith, P.J. (1989). Is two-phase sampling really better for estimating age composition? *Journal of the American Statistical Association*, 84, 916-921.

[10] Treder, R.P. and Sedrensk, J. (1993). Double sampling for stratification. *Survey Methodology*, 19, 95-101.

[2000년 7월 접수, 2000년 10월 채택]

Mean Estimation in Two-phase Sampling

Kyuseong Kim[1] Jinseog Kim[2] Sun Soon Lee[3]

ABSTRACT

In this paper, we investigated mean estimation methods in two-phase sampling. Under the fixed expected cost we reviewed the optimal sample sizes, minimum variances and approximate unbiased variance estimators for usual ratio estimator, stratified sample mean with proportional allocation and Rao's allocation of the second phase sample. Also we proposed combined ratio estimator, which uses both ratio estimation and stratification and derived optimal sample size, minimum variance and unbiased variance estimator. Through a limited simulation study, we compared estimators by design effects and came to know that ratio estimator is more efficient than stratified sample mean in some cases and inefficient in the other cases, but combined ratio estimator is more efficient than others in most cases.

Keywords: Combined ratio estimator; Ratio estimator; Stratified sample mean; Two-phase sampling.

1) Associate Professor, Dept. of Computer Science and Statistics, The Univ. of Seoul. E-mail: kskim@uoscc.uos.ac.kr
2) Graduate Student, Department of Statistics, Seoul National University.
3) Graduate Student, Department of Statistics, Seoul National University.