응용통계연구 (2011)
24(6), 1115–1128
DOI: http://dx.doi.org/10.5351/KJAS.2011.24.6.1115

복합추정량을 이용한 절사표본 총합 추정에 관한 연구

김지학[1] · 신기일[2]

[1]한국외국어대학교 통계학과, [2]한국외국어대학교 통계학과

(2011년 8월 접수, 2011년 10월 채택)

요 약

절사표본 설계는 관심변수의 분포가 오른쪽으로 치우쳐진 경우에 모집단 일부를 표본조사에서 제외시켜 조사의 효율을 높이는 방법이다. 그러나 전체 모집단의 총합 추정을 위해서는 버려진 절사 층에 관한 추론이 필요하게 된다. 기존에 사용하고 있는 많은 방법들은 절사층에서 표본조사를 하지 않고 알려진 보조정보와 조사자료를 이용하여 절사층 총합을 추정하며 이를 이용하여 전체 모집단의 총합을 추정한다. 본 논문에서는 절사층을 완전히 표본 층에서 제거하지 않고 조사의 편리성 및 효율성을 고려하여 최소의 표본을 추출한 후 모집단 전체 총합을 추정하는 방법을 제안하였다. 또한 모의실험을 통하여 제안된 추정법과 기존 방법의 우수성을 비교하였다.

주요용어: 표본조사, 비표본오차, 절사표본, 비추정.

1. 서론

절사표본(cut-off sampling)이란 모집단 일부를 표본에서 제외시켜 표본설계를 하는 방법으로 관심변수의 분포가 오른쪽으로 치우친 경우가 대부분인 사업체 조사에서 주로 사용되고 있다. 사업체 조사인 경우 대규모 사업체들만을 조사에 포함하고도 모총합 기준으로 전체 사업체들을 상당부문 대변할 수 있는 경우가 많이 있다. 또한 소규모 사업체의 경우 잦은 휴, 폐업으로 표본틀 관리에 어려움이 있고 상대적으로 높은 무응답으로 비표본오차가 크게 발생하고 있다. 따라서 이들 소규모 사업체를 모집단에서 제외함으로써 조사의 효율성을 높일 수도 있다. 그러나 이러한 조사의 효율성과 편리성에도 불구하고, 전체 모집단 총합을 추정하기 위해서는 버려진 절사층에 관한 추정이 필요하게 된다.

절사표본에 관한 연구는 크게 두 가지로 요약될 수 있다. 첫 번째는 절사층과 표본층을 나누는 기준인 절사점의 최적점을 구하는 것이고 두 번째는 전수층과 표본층을 최적으로 나누는 최적점을 구하는 것이다. 흔히 두 가지 상이한 기준점이 혼용되어 절사점이란 이름으로 사용되기도 한다.

Hidiroglou (1986)는 수정절사법(modified cut-off sampling)을 제안하였으며 완전히 조사하지 않는 절사층(take-nothing stratum) 없이 표본층과 전수층 만을 고려한 방법을 제안하였다. 이 방법은 신민웅과 이상은 (2001)에 자세히 설명되어 있으며 표본층과 전수층을 나누는 최적의 절사점을 구하는 방법을 자세히 설명하고 있다. 또한 최적점을 구하는 SAS code도 제공되어 있다. 최근 Lee (2011)은 Hidiroglou (1986)가 제안한 수정절사법의 절사점 추정 방법을 확장하여, 주어진 허용오차를 만족하는 최소의 표본 수가 주어졌을 때 MSE(mean squared error)를 최소로 하는 새로운 절사점 추정 방법을 제안하였다. 또한 Lee와 Shin (2011)은 이에 관한 이론과 그 타당성을 제시하였다.

이 논문은 2011년도 한국외국어대학교 학술연구비 지원에 의해 이루어진 것임.
[2]교신저자: (449-791) 경기도 용인시 모현면 산 89, 한국외국어대학교 통계학과, 교수.
E-mail: keyshin@hufs.ac.kr

다른 한편에서는 조사하지 않는 절사층의 총합 추정을 위한 연구가 활발히 진행되었다. 이 연구에서는 절사층에서 관심변수가 얻어지지 않았지만 행정자료나 다른 조사에서 이미 보조변수는 조사되어 그 결과가 얻어졌다는 것을 가정하고 있다. 먼저 Elisson과 Elvers (2001)는 스웨덴의 기업체 매출액 자료를 이용하여 절사층에서 얻어진 보조변수와 표본층에서 얻어진 관심변수와 보조변수의 비추정값을 이용하여 절사층의 총합을 추정하는 방법을 연구하였다. 여기에 추가하여 전년도 또는 전월에서 구해진 절사층과 표본층의 비추정값으로 앞서 구한 비추정 총합 추정값을 보정하는 방법도 연구하였다. Clark과 Kinyon (2007)은 미국 통계국(US. Census Bureau)의 2005년 소매업조사(Annual Reatil Trade Survey; ARTS)와 서비스업조사(Service Annual Survey; SAS)의 새로운 표본설계를 실시하면서 응답자 부담과 자료수집비용을 절감하는 방안으로 절사법을 연구하였다. 이 논문에서는 총합 추정을 위해 행정자료를 이용한 직접대체, 기존 센서스자료를 보조변수로 하는 회귀추정 또는 시계열 자료 형태인 경우에는 전년대비 비율 조정을 통한 추정 등을 연구하였다. 또한 McDowney (2004)는 모형기반이론(model-based theory) 또는 예측이론을 통한 절사추정법을 이용하여 미국에너지정보국(Energy Information Administration; EIA)의 월별 전기 매출액 및 전기 발전 자료의 총 매출액을 추정하였다.

이와 같이 많은 연구가 이루어졌고 실제 표본조사의 총합추정에 절사층에서 표본을 조사하지 않는 절사법이 사용되고 있다. 그러나 절사층의 총합을 추정하는데 있어 비록 조사 부담성 및 응답 부담성 등을 고려하더라도 완전히 조사를 하지 않은 것에 비해 일부의 표본을 절사층에 포함하여 조사하였을 경우 더 좋은 결과가 얻어질 수 있을 것으로 판단된다. 따라서 최소의 표본을 절사층에서 조사하는 방법을 연구할 필요가 있으며 그 효율성을 살펴볼 필요가 있다.

본 논문의 구성은 다음과 같다. 먼저 2절에서는 Benedetti 등 (2010)의 내용을 이용하여 절사표본조사의 일반적 구조와 기호를 설명하고 기존에 사용되었던 절사층 추정법을 간단히 살펴보았다. 이때 기존의 절사표본 총합 추정법의 내용은 박인호 (2007)를 참조하였다. 또한 일부의 표본을 절사층에서 조사할 경우에 사용할 수 있는 새로운 추정량을 제안하였다. 3절에서는 모의실험을 통하여 기존의 절사층 추정량과 새롭게 제안한 복합추정량을 여러 비교 통계량을 기준으로 살펴보았다. 특히 절사층에 배정된 표본 수에 따른 효율성을 살펴보았다. 4절에 결론이 있다.

2. 절사표본

2.1. 기존의 절사표본

2.1.1. 절사표본 및 기호 절사표본 조사방법을 살펴보기 위해 Benedetti 등 (2010)의 설정과 기호를 사용하였다. 먼저 모집단은 3개의 층으로 구분된다. 크기가 N인 전체 모집단, U는 모든 개체를 조사하는 전수층(take-all or census stratum; U_C), 일부만을 표본조사하는 표본층(take-some stratum; U_S) 그리고 표본조사에서 제외하는 절사층(take-nothing or exclusion stratum; U_E)의 3개로 분할 또는 층화된다. 따라서 조사변수 y의 모총합은 각 층 총합을 더한 다음의 형태로 표기할 수 있다.

$$t_y = t_{yC} + t_{yS} + t_{yE}, \tag{2.1}$$

여기서 t_y는 모총합이고 t_{yC}, t_{yS}, t_{yE}는 전수층, 표본층, 절사층 각각의 층 총합들이다. 또한 전수층과 표본층을 합하여 조사에 포함(inclusion)된 층인 조사포함층(혹은 광의의 표본층)의 총합 t_{yI}는 다음과 같이 정의된다.

$$t_{yI} = t_{yC} + t_{yS}.$$

물론 표본층은 표본설계에 따라 하나 이상의 세부층($U_S = U_{h=1}^{H} U_h$)으로 나뉠 수도 있다. 다음으로 조

사포함층, U_I는 다음과 같다.

$$U_I = U_C \cup U_S.$$

또한 모집단으로부터 추출된 전체 표본 S_U는 S를 표본집합이라 할 때 S는 전수층과 표본층을 합한 다음의 형태이다.

$$S_U = U_C \cup S \cup \emptyset.$$

다음으로 모집단 각 개체에 대하여 보조변수 값을 알고 있다고 가정하면, 모집단 및 분할층의 보조변수 x-총합은 $t_x, t_{xC}, t_{xS}, t_{xE}, t_{xI}$로 나타낼 수 있다. 또한 절사표본추출에서 각 개체들이 표본에 포함될 확률(inclusion probability)은 다음과 같다.

$$\pi_k = \begin{cases} 1, & k \in U_C, \\ > 0, & k \in U_S, \\ 0, & k \in U_E. \end{cases}$$

2.1.2. 모총합 추정전략 연구를 단순화하기 위해 무응답이 없다고 가정하자. 표본층 총합 추정량은 전통적인 설계기반이론(design-based theory)하에서의 Horvitz-Thompson 추정량인 $\hat{t}_{yS} = \sum_{k\in S} y_k/\pi_k$이 사용되며 이는 t_{yS}의 불편추정량(unbiased estimator)으로 알려져 있다. 이제 식 (2.1)의 총합 추정을 위한 경우를 살펴보자.

Case 1: t_y에 대한 t_{yE}의 기여도가 무시할 정도로 작은 경우

이 경우는 $t_y \approx t_{yI}$이므로 $\hat{t}_{yS}$에 t_{yC}를 더하여 t_{yI}의 불편추정량을 구할 수 있다. 즉

$$\hat{t}_{yI} = t_{yC} + \hat{t}_{yS} = \sum_{k\in S_U} \frac{y_k}{\pi_k} \tag{2.2}$$

이 된다. 이제 $A_y = t_y/t_{yI}$라 하면 $t_{yE}/t_y \approx 0$이고 $A_y \approx 1$이 된다. 흔히 이 추정량을 단순절사추정량이라 부른다. 여기서 $A_y = t_y/t_{yI}$는 조사변수 y를 기준으로 한 조사포함층 U_I에 대한 모집단 U의 상대크기를 나타낸다. 절사층을 고려하지 않은 단순절사추정량의 상대편향(relative bias; rbias)은

$$\text{rbias}\left(\hat{t}_{yI}\right) = \frac{E(\hat{t}_{yI}) - t_y}{t_y} = -\frac{t_{yE}}{t_y} = A_y^{-1} - 1$$

으로 음의 값임을 알 수 있으며 결론적으로 이는 과소추정을 의미한다.

Case 2: 절사층에 관한 보조 정보가 있는 경우

조사에 포함되지 않는 절사층에 대해서 사전에 알고 있는 정보, 예를 들면, 표본틀 상의 보조변수값 x_k나 조사 이전에 파악한 x_k와 y_k의 관계 혹은 표본을 통해 추측할 수 있는 보조 정보가 있다면 이 정보를 이용하여 절사층 총합을 추정할 수 있다. 즉 모형을 이용하여 적절한 총합 추정량을 구한 후 이를 모두 합하여 모집단 전체의 모총합 추정량을 얻을 수 있다.

이때 절사층의 총합을 직접적으로 추정하는 방식인

$$\hat{t}_y = t_{yC} + \hat{t}_{yS} + \hat{t}_{yE} \tag{2.3}$$

을 사용하거나 절사층의 상대적 크기를 보정해 주는 방식인

$$\hat{t}_y = \hat{A}_y \hat{t}_{yI} \tag{2.4}$$

의 형태를 고려할 수 있다. 만약, $\hat{A}_y = (t_{yC} + \hat{t}_{yS} + \hat{t}_{yE})/(t_{yC} + \hat{t}_{yS})$이면, 위의 두 추정량은 결국 동일한 추정량이 된다. 여기서 $\hat{A}_y$와 같이 $\hat{t}_{yI}$에 일정한 수를 곱하여 전체를 추정하는 계수를 확대계수라고 부른다.

2.1.3. 모총합 추정량

Benedetti-Bee-Espa(BBE) 방법

Benedetti 등 (2010)은 조사 이전에 알고 있는 보조변수를 이용하여 식 (2.4)의 확대계수 $\hat{A}_y$ 값을 대체하는 방식으로 다음의 추정량을 제시하였다.

$$\hat{t}_y^{BBE} = A_x \hat{t}_{yI}, \tag{2.5}$$

여기서 $A_x = t_x/t_{xI}$이고 t_x와 t_{xI}는 각각 모집단과 조사포함층의 알려진 보조변수 x의 총합이다. 이는 $A_y \approx A_x$일 때 적합한 모형으로 대표적인 예로는 두 변수 y와 x가 비례 관계를 갖는 경우이다. 특히 $A_x \approx 1$인 경우에는, $\hat{t}_y^{BBE} \approx \hat{t}_{yI}$임을 알 수 있다.

Sarndal-Swansson-Wretman(SSW) 방법

모집단과 조사포함층의 두 변수 y와 x사이의 총합비율 혹은 상대적 크기를 모집단과 조사포함층에서 각각 $R_{yx} = t_y/t_x$과 $R_{yxI} = t_{yI}/t_{xI}$으로 정의하자. 이 두 값이 매우 근사한 경우, 즉

$$R_{yxI} \approx R_{yx}$$

인 경우를 가정하여 Sarndal 등 (1992)은 다음의 추정량을 연구하였다.

$$\hat{t}_y^{SSW} = \hat{R}_{yxI} t_x, \tag{2.6}$$

여기서 $\hat{R}_{yxI} = \hat{t}_{yI}/\hat{t}_{xI}$는 표본으로부터 추정한 조사포함층에서의 총합비율을 나타낸다. 이는

$$\hat{t}_y^{SSW} = A_x^{'} \hat{t}_{yI}$$

의 형태로도 표현될 수 있다. 즉 확대계수 $A_x^{'} = t_x/\hat{t}_{xI}$는 식 (2.5)의 확대계수 A_x와는 분모에만 차이가 있을 뿐이다.

BLUP 방법

McDowney (2004)는 조사 변수값 y_k가 확률변수 Y_k의 실현 값이며 모형 $y_k = \beta x_k + \sqrt{x_k}\epsilon_k$, $\epsilon_k \sim (0, \sigma^2)$을 따른다고 가정하자. 그러면 모총합 t_y의 최량선형불편추정량(best linear unbiased predictor; BLUP)은 다음과 같이 표현된다.

$$\hat{T}_y^{BLUP} = \hat{t}_{yI} + R_{YxS_U} \sum_{S_U^c} x_k, \tag{2.7}$$

여기서 $R_{YxS_U} = T_{yS_U}/t_{xS_U}, T_{yS_U} = \sum_{S_U} y_k, t_{xS_U} = \sum_{S_U} x_k$이다. 이는 모총합 추정에서 모형기반 논리에 따른 표본설계의 추출확률을 고려하지 않고 조사변수 값을 단순히 합하여 추정한 것으로 모형이 실제자료에 맞지 않으면 편향이 커질 수 있어 추정의 효율이 떨어질 수 있다.

표 **2.1.** 모집단, 표본, 총합 및 분할 기호

분할	전수층	표본층	절사층(절사표본층)	전체
모집단	U_C	U_S	U_E	U
모집단크기	N_C	N_S	N_E	N
표본	U_C	S	S_E	S_U
표본크기	N_C	n_S	n_{S_E}	n
표본추출률	1	$0 < \pi_k^S < 1$	$0 < \pi_k^{S_E} < 1$	π
y총합	t_{yC}	t_{yS}	t_{yE}	t_y
x총합	t_{xC}	t_{xS}	t_{xE}	t_x

2.2. 제안된 방법

본 논문에서 제안된 방법은 절사층에서 전혀 조사를 하지 않는 방법이 아니라 절사층에서 소수의 표본을 조사하여 이를 절사층의 총합추정에 사용하는 것이다. 따라서 절사층이라 부르는 것보다는 오히려 절사표본층이라 부르는 것이 타당하다. 이제 기존의 방법과 제안된 방법에서 사용하는 기호를 설명하면 표 2.1과 같다. 기존의 방법과 구별하기 위하여 절사층에서 표본이 추출되지 않았으면 절사층이라 부르고 만약 소수의 표본이 추출되었다면 이를 절사표본층이라 부르겠다.

표 2.1에서 $n_{S_E} = 0$이면 절사층이되고, $n_{S_E} > 0$이면 절사표본층이 된다. 본 연구에서는 $n_{S_E} > 0$인 경우에 사용할 수 있는 새로운 추정법인 복합추정량을 제안하였으며 이 방법과 기존의 방법의 우수성을 비교하였다. 특히 일반적으로 절사표본을 사용하는 주된 이유가 절사층의 표본관리가 어렵고 무응답 비율이 높기 때문이므로 현실적으로 작은 수의 표본이 절사표본층에 배정되는 것이 타당하다.

박인호 (2007)는 Lee 등 (1995)에서 사용한 모의실험 설계를 이용하여 기존의 여러 절사법 총합추정량을 비교하였으며 그 결과 SSW가 우수한 결과를 주는 것을 확인하였다.

이제 SSW를 살펴보자. SSW 방법은 U_C, U_S, U_E 층에서 같은 비를 갖는다는 조건 또는 $R_{yxI} = R_{yx}$라는 조건하에서 얻어진 비추정 결과이다. 즉

$$\hat{t}_y^{SSW} = \frac{t_x}{\hat{t}_{xI}}\hat{t}_{yI} = \frac{t_{xE} + t_{xI}}{\hat{t}_{xI}}\hat{t}_{yI} = \frac{\hat{t}_{yI}}{\hat{t}_{xI}}t_{xE} + \frac{t_{xI}}{\hat{t}_{xI}}\hat{t}_{yI}$$

이다. 이 추정량은 절사층 총합, t_{yE}를 추정하기 위하여 비추정을 사용하였으며 비추정값으로 $\hat{R}_E = \hat{t}_{yI}/\hat{t}_{xI}$을 사용하였다. 이 값은 절사층에서 조사가 전혀 이루어지지 않았기 때문에 표본층과 전수층 자료를 이용하여 비를 추정한 후 이 결과를 이용하여 얻어진다. 즉 $\hat{t}_{yE} = \hat{R}_E t_{xE} = (\hat{t}_{yI}/\hat{t}_{xI})t_{xE}$를 사용하였다. 그러나 본 연구에서는 절사층이 아닌 절사표본층, 즉 절사층에서 일부의 표본을 추출한 경우이므로 절사표본층에서 얻어진 추가 정보를 이용한 추정량을 제안하였다.

본 연구에서는 다음과 같은 추정량을 제안한다.

방법 1: n_{S_E}가 충분히 클 때

절사법은 절사층에서 표본을 구하기 어려운 경우에 사용하는 방법이므로 n_{S_E}가 충분히 큰 경우는 흔하지 않을 것으로 판단된다. 그러나 절사층에서 비를 추정할 정도의 표본이 얻어진 경우에는 다음의 추정량을 사용할 수 있다.

$$\hat{t}_{yE}^{[1]} = \hat{R}_E^{[1]} t_{xE} = \frac{\hat{t}_{yE}}{\hat{t}_{xE}} t_{xE}. \tag{2.8}$$

즉 절사층에서 비를 추정하고 추정된 비를 이용하여 절사층의 총합을 구한다.

표 **3.1.** 모의실험에 사용된 계수

조사변수형태	a	b	c	d	g	평균	표준오차
비례형	0	1.50	0.00	5.13	0.50	71.960	53.900
선형	20	1.50	0.00	13.79	0.25	92.405	54.998
볼록형	0	0.25	0.01	4.91	0.50	43.061	56.609
오목형	0	3.00	−0.01	5.60	0.50	113.591	63.118

방법 2: 복합추정량

만약 절사층의 표본 수가 충분하지 않을 경우에는 위 방법의 효율은 떨어질 수 있다. 따라서 다음의 선형 결합 추정량을 제안할 수 있다.

$$\hat{t}_{yE}^{[2]} = \hat{R}_E^{[2]} t_{xE} = \left(\alpha \frac{\hat{t}_{yE}}{\hat{t}_{xE}} + (1-\alpha) \frac{\hat{t}_{yI}}{\hat{t}_{xI}} \right) t_{xE}. \tag{2.9}$$

이 추정량은 분산이 크고 편향이 작은 추정량과 편향이 크고 분산이 작은 추정량을 선형결합하여 얻는 복합추정량이다. 복합추정량을 사용할 때의 어려운 점은 가중치 α를 추정하는 것이다. 일반적으로 고려할 수 있는 방법은 다음과 같다.

$$\hat{\alpha} = \frac{\mathrm{MSE}\left(R_E^{[SSW]}\right)}{\mathrm{MSE}\left(R_E^{[1]}\right) + \mathrm{MSE}\left(R_E^{[SSW]}\right)} \approx \frac{V\left(R_E^{[SSW]}\right)}{V\left(R_E^{[1]}\right) + V\left(R_E^{[SSW]}\right)}, \tag{2.10}$$

여기서 $R_E^{[SSW]} = \hat{t}_{yI}/\hat{t}_{xI}$이다. 이 추정량은 흔히 소지역 추정법에서 사용되고 있으며 자세한 것은 Rao (2003)를 참조하기 바란다. 만약 $\alpha = 1$이면 방법 1의 추정량이 되고, $\alpha = 0$이면 SSW 방법이 된다. 식 (2.10) 사용에 어려움이 있는 경우에는 간단히 $\alpha = 0.5$를 사용하기도 한다. 본 연구에서는 $\alpha = 0.3, 0.5, 0.7$ 그리고 1인 경우를 살펴보았다. 또한 모의실험에서는 MSE를 추정할 수 있으므로 식 (2.10)을 이용하여 α를 추정할 수 있다. 이 추정값을 이용한 결과도 비교하였다.

3. 모의실험을 통한 모총합 추정량 비교

3.1. 모의실험 설계

2절에서 살펴본 절사표본하에서 모총합 추정량들의 효율성을 비교하기 위해 모의실험을 수행하였다. 모의실험은 Lee 등 (1995)이 사용한 방법을 적용하여 크기 $N = 10{,}000$인 모집단을 다음과 같이 생성하였다. 먼저 보조자료 x_k는 평균 48이고 분산 768을 갖는 감마분포로부터 생성하였다. 주어진 x_k 값에 대하여 모두 네 종류의 조사자료 y_k를 생성하였으며 각각 평균 $\mu(x) = a + bx + cx^2$이고 분산 $\sigma^2(x) = d^2 x^{2g}$을 갖는 감마분포를 가정하였다. 표 3.1은 선택된 상수 a, b, c, d, g의 값과 생성된 모집단 자료의 평균 및 표준편차를 나타낸다. 첫 번째 형태는 관심변수와 보조변수와의 관계가 원점을 지나는 비례 형태(ratio)이고, 두 번째 형태는 양의 절편 값을 갖는 선형관계(regression), 세 번째 형태는 볼록한 형태(convex) 그리고 마지막 형태는 오목한 형태(concave)를 갖도록 하였다. 보조변수와의 상관계수는 약 0.75 정도이다.

다음으로 전수층과 표본층을 분리하기 위해 먼저 보조변수값 x_k를 기준으로 내림차순으로 정리하였다. 전수층은 x_k-누적합이 5% 이상을 차지하는 167개의 개체들로 구성하였고, 절사층은 x_k-누적합이 $100p_x$% 이상을 차지하고 남는 개체들로 구성하였으며, 나머지 개체들은 표본층을 구성하였다. 이때, 절사점으로 다음의 세 기준점들, 80%, 90%, 95% (즉, $p_x = 0.80, 0.90, 0.95$)를 고려하였다. 총표본 크

표 **3.2.** 모의실험을 위한 층별 모집단 크기, 표본 크기 및 추출 비율

절사점기준	N_C	N_S	N_E	n_S	n_{S_E}	n_S/N_S	n_{S_E}/N_E
80%	167	5806	4027	333	0	0.0565	0
				328	5	0.0555	0.0024
				308	25	0.0530	0.0059
				283	50	0.0539	0.0048
90%	167	7287	2550	333	0	0.0457	0
				328	5	0.0450	0.0020
				308	25	0.0423	0.0098
				283	50	0.0388	0.0196
95%	167	8241	1592	333	0	0.0404	0
				328	5	0.0398	0.0031
				308	25	0.0374	0.0157
				283	50	0.0343	0.0314

기는 $n = 500$이고 표본추출률은 $f = n/N = 0.05$를 사용하였다. 또한 새롭게 제안된 추정량의 모의실험을 위하여 절사표본층에서 표본 수 $n_{S_E} = 5, 25, 50$개를 표본추출하였다. 반복은 $R = 3{,}000$번 실시하였으며 각 표본에서 편향(bias), 상대편향(rbias), 제곱근 평균제곱오차(root mean square error; rmse)를 구하여 비교 통계량으로 사용하였다.

$$\text{bias} = \bar{\hat{t}}_y - t_y,$$

$$\text{rbias}(\%) = \frac{100\left(\bar{\hat{t}}_y - t_y\right)}{t_y},$$

$$\text{rmse} = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\left[\hat{t}_y(r) - t_y\right]^2},$$

여기서 $\bar{\hat{t}}_y = \sum_{r=1}^{R}\hat{t}_y(r)/R$이다.

본 논문의 모의실험 결과 중에서 단순절사추정량, BBE, SSW 그리고 BLUP에 관한 결과는 박인호 (2007)의 결과와 일치하여 본 논문에는 수록하지 않았다. 박인호 (2007)의 결과를 요약하면 단순절사추정량은 과소추정량으로 모든 비교 통계량에서 가장 나쁜 결과를 주는 것으로 나타났으며, 비율형인 경우 BLUP이 그리고 그 외의 모든 경우에는 SSW가 가장 우수한 결과를 주는 것으로 나타났다.

3.2. 결과분석

모의실험 결과는 SSW와 $\alpha = 0.3, 0.5, 0.7$ 그리고 1인 경우의 bias, rbias, rmse이다. 그림 3.1과 그림 3.2는 상위 80% 절사점을 사용하였을 때 3,000번의 반복에서 얻어진 표본 추정값의 상자그림이다. 그림 3.1은 절사표본층에서 $n_{S_E} = 5$개의 표본을 추출하여 얻어진 SSW와 $\alpha = 0.3, 0.5, 0.7$ 그리고 1인 복합추정량의 결과이다. 그림을 보면 비율형에서 SSW는 복합추정량과 유사한 편향을 보이고 있으며 분산이 작은 것을 확인할 수 있다. 그러나 다른 형태, 즉 선형, 볼록형 및 오목형에서는 SSW의 편향이 커지는 것을 확인할 수 있다. 물론 모든 형태에서 최소의 분산을 갖고 있음을 알 수 있다.

반면 α가 "1"에 가까워지면서 복합추정량의 분산은 점점 증가하는 것을 확인 할 수 있다. 이와 같이 분산이 커지는 현상은 모든 형태에서 나타나고 있다. 그러나 α가 "1"에 가까워지면서 편향은 줄어들고 있다. 또한 $n_{S_E} = 50$인 경우에는 복합추정량의 분산이 SSW와 모든 α에서 거의 유사하게 나오는 것을

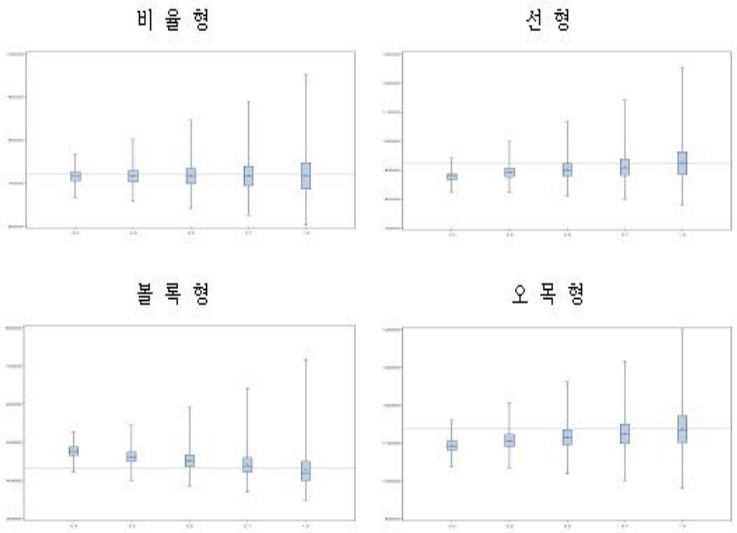

그림 3.1. 가중치 α에 따른 모총합 추정량 상자그림($n_{S_E} = 5$)

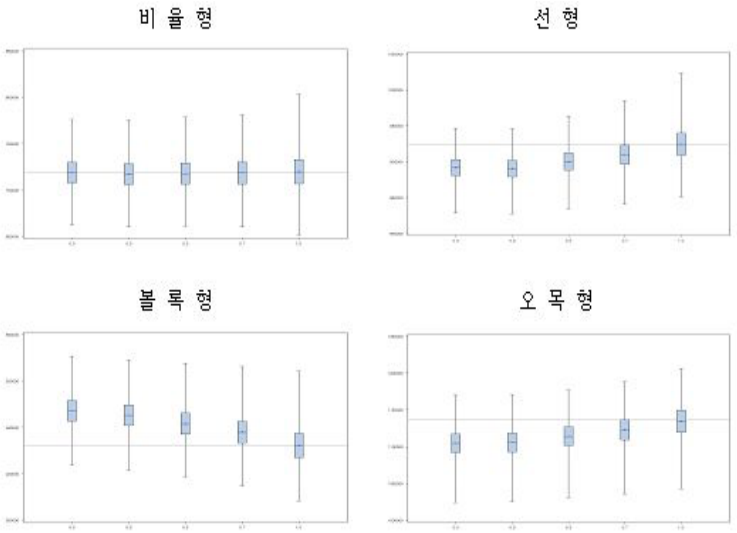

그림 3.2. 가중치 α에 따른 모총합 추정량 상자그림($n_{S_E} = 50$)

확인할 수 있다.

물론 $n_{S_E} = 5$와 $n_{S_E} = 50$인 경우 절사층에서 표본을 추출하지 않았으므로 SSW는 같은 결과를 준다. 그러나 복합추정량의 경우에는 절사표본층의 자료가 커지면서 분산이 줄어드는 것을 보여주고 있다. 다음으로 표 3.3에서 표 3.5에는 비교통계량 결과를 수록하였다.

또한 최적 α를 사용하였을 때의 결과를 확인하기 위한 모의실험을 실시하였다. 복합추정량은 분산이 크고 편향이 작은 추정량과 분산이 작고 편향이 큰 두 추정량을 선형결합하여 얻어진다. 표 3.3에서 표 3.5의 결과에서 절사점이 80%이고 $n_{SE} = 25$인 선형과 오목형의 경우를 제외하면, n_{SE}가 25 이상에서

표 **3.3.** 80% 절사점 하에서 모층합 추정량 비교

n_{S_E}	형태		SSW	$\alpha = 0.3$	$\alpha = 0.5$	$\alpha = 0.7$	$\alpha = 1$
5	비율	bias	−4244.745	−2868.961	−2262.169	−1655.377	−745.189
		rbias	−0.590	−0.399	−0.314	−0.230	−0.104
		rmse	15685.043	20091.793	26617.176	34312.169	46778.286
	선	bias	−48027.139	−32895.534	−22934.549	−12973.563	1967.915
		rbias	−5.197	−3.560	−2.482	−1.404	0.213
		rmse	50355.101	40078.336	39735.790	44938.987	59646.717
	볼록	bias	46527.940	31484.973	21883.498	12282.022	−2120.191
		rbias	10.805	7.312	5.082	2.852	−0.492
		rmse	49303.374	36860.418	32897.028	33441.801	41937.069
	오목	bias	−43178.920	−29448.422	−20472.843	−11497.263	1966.105
		rbias	−3.801	−2.592	−1.802	−1.012	0.173
		rmse	46567.771	37323.907	36613.859	40710.291	53148.902
25	비율	bias	−4244.745	−2730.008	−1877.029	−1024.049	255.420
		rbias	−0.590	−0.379	−0.261	−0.142	0.035
		rmse	15685.043	16570.168	17726.810	19716.534	23786.672
	선	bias	−48027.139	−33725.650	−24060.345	−14395.040	102.917
		rbias	−5.197	−3.650	−2.604	−1.558	0.011
		rmse	50355.101	37507.575	30506.507	26327.270	28005.532
	볼록	bias	46527.940	32191.976	22982.846	13773.716	−39.979
		rbias	10.805	7.476	5.337	3.199	−0.009
		rmse	49303.374	36206.851	28922.799	23658.271	22693.153
	오목	bias	−43178.920	−30138.559	−21606.503	−13074.446	−276.361
		rbias	−3.801	−2.653	−1.902	−1.151	−0.024
		rmse	46567.771	35550.366	29638.113	26096.743	27157.602
50	비율	bias	−4244.745	−2370.496	−1572.204	−773.913	423.524
		rbias	−0.590	−0.329	−0.218	−0.108	0.059
		rmse	15685.043	16629.463	16900.551	17703.367	19751.600
	선	bias	−48027.139	−33649.162	−23931.119	−14213.076	363.988
		rbias	−5.197	−3.641	−2.590	−1.538	0.039
		rmse	50355.101	37507.642	29690.422	23899.296	22554.467
	볼록	bias	46527.940	32482.718	23272.701	14062.685	247.660
		rbias	10.805	7.543	5.405	3.266	0.058
		rmse	49303.374	36524.488	28707.121	22317.551	18811.505
	오목	bias	−43178.920	−30179.045	−21684.763	−13190.482	−1412.393
		rbias	−3.801	−2.657	−1.909	−1.161	−0.124
		rmse	46567.771	35582.041	28901.104	23864.805	21919.424

선형, 오목형, 볼록형인 경우 $\alpha = 1$인 추정량이 가장 작은 rmse 결과를 주고 있다. 또한 모든 비율형에서는 SSW가 가장 작은 rmse 결과를 보여준다. 결국 이러한 경우에는 복합추정량을 사용할 필요가 없게 된다. 따라서 본 모의실험에서는 $n_{S_E} = 5$이고, 선형, 볼록형, 오목형인 경우에 최적 α를 구한 후 각 비교통계량의 결과를 살펴보았다. 이 결과는 표 3.6에 나와있다.

표 3.3의 $n_{S_E} = 5$인 경우를 살펴보자. 그림 3.1과 그림 3.2의 비율형에서 SSW의 편향이 복합추정량과 비슷한 것으로 보였으나 표 3.3을 살펴보면 비율형, 선형, 볼록형, 오목형에서 SSW의 편향이 가장 큰 것을 확인할 수 있다. 다음으로 복합추정량의 경우 $\alpha = 1$에 가까워질수록 편향이 줄어든다. 그러나

표 **3.4.** 90% 절사점 하에서 모총합 추정량 비교

n_{S_E}	형태		SSW	$\alpha = 0.3$	$\alpha = 0.5$	$\alpha = 0.7$	$\alpha = 1$
5	비율	bias	−1011.213	−1023.842	−735.689	−447.536	−15.307
		rbias	−0.141	−0.142	−0.102	−0.062	−0.002
		rmse	16424.619	17030.364	18023.688	19603.718	22796.275
	선	bias	−33805.987	−23367.118	−16310.193	−9253.269	1332.118
		rbias	−3.658	−2.529	−1.765	−1.001	0.144
		rmse	37529.531	29464.038	26072.810	25317.294	29377.907
	볼록	bias	22918.342	16569.627	11972.794	7375.962	480.713
		rbias	5.322	3.848	2.780	1.713	0.112
		rmse	28643.963	24374.930	22240.756	21455.969	23046.850
	오목	bias	−20555.367	−15177.203	−10993.288	−6809.372	−533.500
		rbias	−1.810	−1.336	−0.968	−0.599	−0.047
		rmse	28187.271	25149.379	23904.460	23976.032	26474.866
25	비율	bias	−1011.213	−568.504	−266.678	35.148	487.887
		rbias	−0.141	−0.079	−0.037	0.005	0.068
		rmse	16424.619	17093.765	17317.283	17827.742	19077.594
	선	bias	−33805.987	−23916.094	−17105.141	−10294.188	−77.759
		rbias	−3.658	−2.588	−1.851	−1.114	−0.008
		rmse	37529.531	29324.991	24641.161	21593.670	21556.640
	볼록	bias	22918.342	15693.949	11136.898	6579.847	−255.729
		rbias	5.322	3.645	2.586	1.528	−0.059
		rmse	28643.963	23869.820	21317.644	19738.944	19675.076
	오목	bias	−20555.367	−14873.247	−10670.084	−6466.922	−162.178
		rbias	−1.810	−1.309	−0.939	−0.569	−0.014
		rmse	28187.271	24432.351	22297.641	21100.535	21376.025
50	비율	bias	−1011.213	−862.555	−636.288	−410.022	−70.621
		rbias	−0.141	−0.120	−0.088	−0.057	−0.010
		rmse	16424.619	18023.456	17948.948	18019.754	18393.817
	선	bias	−33805.987	−23543.038	−16688.458	−9833.879	447.989
		rbias	−3.658	−2.548	−1.806	−1.064	0.048
		rmse	37529.531	29432.784	24459.226	20835.074	19561.397
	볼록	bias	22918.342	15705.818	11097.153	6488.488	−424.509
		rbias	5.322	3.647	2.577	1.507	−0.099
		rmse	28643.963	24314.509	21482.241	19459.879	18505.172
	오목	bias	−20555.367	−14702.589	−10445.603	−6188.617	196.861
		rbias	−1.810	−1.294	−0.920	−0.545	0.017
		rmse	28187.271	25174.563	22868.604	21323.587	20784.160

rmse를 비교하면 형태에 따라 다른 결과가 얻어진다. 먼저 비율형을 살펴보면 SSW의 rmse가 가장 작고, $\alpha = 1$에 가까워질수록 커지는 것을 확인 할 수 있다. 그러나 선형, 볼록형, 오목형에서는 SSW의 rmse가 최소가 되지 않는다. 이 형태에서는 $\alpha = 0.5$인 경우에 최소값이 얻어진다. 따라서 형태에 따라 최적의 추정량이 달라지고 있음을 확인할 수 있다. 이러한 결과는 복합추정량에서 얻어지는 일반적인 결과이다.

표 3.3의 $n_{S_E} = 25$인 경우를 살펴보면 비율형인 경우 SSW가 가장 큰 편향을 보인 반면 가장 작은 rmse를 보이고 있다. 그러나 선형, 볼록형, 오목형인 경우 $\alpha = 0.7$ 또는 $\alpha = 1$인 경우에 rmse가 최

표 **3.5.** 95% 절사점 하에서 모총합 추정량 비교

n_{S_E}	형태		SSW	$\alpha = 0.3$	$\alpha = 0.5$	$\alpha = 0.7$	$\alpha = 1$
5	비율	bias	−1667.252	−1157.019	−867.985	−578.952	−145.402
		rbias	−0.232	−0.161	−0.121	−0.080	−0.020
		rmse	17395.020	17402.417	18103.983	19219.428	21521.819
	선	bias	−22510.972	−16031.122	−11531.835	−7032.548	−283.617
		rbias	−2.436	−1.735	−1.248	−0.761	−0.031
		rmse	28433.743	24202.569	22853.798	23071.910	26174.860
	볼록	bias	11514.306	8673.054	6160.231	3647.408	−121.826
		rbias	2.674	2.014	1.431	0.847	−0.028
		rmse	21338.445	20213.743	19772.283	19949.879	21328.494
	오목	bias	−12205.569	−8375.218	−5895.092	−3414.967	305.222
		rbias	−1.075	−0.737	−0.519	−0.301	0.027
		rmse	23288.690	21884.847	21860.106	22585.668	24919.452
25	비율	bias	−1667.252	−1049.231	−728.686	−408.141	72.677
		rbias	−0.232	−0.146	−0.101	−0.057	0.010
		rmse	17395.020	17722.490	17693.861	17760.729	18037.351
	선	bias	−22510.972	−15222.271	−10679.062	−6135.854	678.959
		rbias	−2.436	−1.647	−1.156	−0.664	0.073
		rmse	28433.743	23791.706	21386.751	19914.833	19878.343
	볼록	bias	11514.306	7428.364	5060.059	2691.753	−860.705
		rbias	2.674	1.725	1.175	0.625	−0.200
		rmse	21338.445	20079.208	19265.267	18788.886	18758.580
	오목	bias	−12205.569	−8036.350	−5578.865	−3121.379	564.850
		rbias	−1.075	−0.707	−0.491	−0.275	0.050
		rmse	23288.690	22150.854	21407.444	21022.798	21156.360
50	비율	bias	−1667.252	−1172.764	−901.776	−630.788	−224.306
		rbias	−0.232	−0.163	−0.125	−0.088	−0.031
		rmse	17395.020	18684.247	18556.327	18473.239	18433.901
	선	bias	−22510.972	−16384.648	−11832.478	−7280.308	−452.053
		rbias	−2.436	−1.773	−1.280	−0.788	−0.049
		rmse	28433.743	24660.260	21920.267	19935.969	18879.570
	볼록	bias	11514.306	7491.466	5163.321	2835.177	−657.040
		rbias	2.674	1.740	1.199	0.658	−0.153
		rmse	21338.445	20839.856	20006.228	19453.864	19202.466
	오목	bias	−12205.569	−8007.772	−5578.364	−3148.956	495.155
		rbias	−1.075	−0.705	−0.491	−0.277	0.044
		rmse	23288.690	22740.689	21900.044	21352.775	21128.533

소가 된다. 따라서 모집단의 관심변수와 보조변수의 관계 형태에 따라 SSW 또는 제안한 복합추정량이 각각 우수한 결과를 주고 있음을 확인할 수 있다. 절사점이 90%인 경우와 95 %인 경우의 결과인 표 3.4와 표 3.5에서도 유사한 결과를 확인할 수 있다. 즉 비율형에서는 SSW가 우수하고 그외의 형태에서는 제안한 복합추정량이 우수하다. 그러나 만약 모집단에서 보조변수와 관심변수와의 관계가 비율형이라는 것이 알려져 있다면 SSW 보다는 BLUP이 우수한 것으로 알려져 있으므로 SSW 대신에 BLUP을 사용하는 것이 타당하다. 표 3.6은 $n_{S_E} = 5$인 선형, 볼록형, 오목형에서 최적의 α를 사용하여 구한 결과이다. 최적 α는 약 0.4에서 0.65 사이에서 구해졌다. 이 복합추정량은 SSW에 비해 비교 통계량 기준으로 모두 우수한 것을 확인할 수 있다.

표 **3.6.** $n_{S_E} = 5$일때의 최적 복합추정량 결과

	형태	$\hat{\alpha}$	비교통계량	SSW	최적 복합추정량	$\alpha = 1$
80%	선	0.416	bias	−48027.139	−27118.163	1967.915
			rbias	−5.197	−2.935	0.213
			rmse	50355.101	39158.948	59646.717
	볼록	0.580	bias	46527.940	18042.907	−2120.191
			rbias	10.805	4.190	−0.492
			rmse	49303.374	32544.683	41937.069
	오목	0.434	bias	−43178.920	−23434.784	1966.105
			rbias	−3.801	−2.063	0.173
			rmse	46567.771	36291.577	53148.902
90%	선	0.620	bias	−33805.987	−12592.792	1332.118
			rbias	−3.658	−1.363	0.144
			rmse	37529.531	25330.802	29377.907
	볼록	0.607	bias	22918.342	9350.628	480.713
			rbias	5.322	2.171	0.112
			rmse	28643.963	21383.090	23046.850
	오목	0.531	bias	−20555.367	−8802.805	−533.500
			rbias	−1.810	−0.775	−0.047
			rmse	28187.271	23903.035	26474.866
95%	선	0.541	bias	−22510.972	−10609.481	−283.617
			rbias	−2.436	−1.148	−0.031
			rmse	28433.743	22767.552	26174.860
	볼록	0.500	bias	11514.306	6160.231	−121.826
			rbias	2.674	1.431	−0.028
			rmse	21338.445	19772.283	21328.494
	오목	0.466	bias	−12205.569	−6316.714	305.222
			rbias	−1.075	−0.556	0.027
			rmse	23288.690	21740.474	24919.452

4. 결론

사업체조사에서는 조사의 효율성과 편의성으로 인해 절사표본조사를 많이 사용하고 있다. 박인호 (2007)는 기존의 여러 추정방법들 중 모의실험을 통해 SSW가 우수한 결과를 주는 것을 확인하였다. 물론 비율형만을 고려한다면 SSW보다 BLUP이 우수한 결과를 준다. 그러나 모집단이 어떤 형태인지를 알려져 있지 않으므로 SSW 방법을 사용하는 것은 큰 무리가 없을 것이다.

SSW의 장점은 조사 비용 및 응답 부담을 줄일 수 있다는 것이다. 그러나 소규모 사업체에서 일부를 조사하는 것이 크게 무리가 아니라면 절사층에서 일정 수의 표본을 조사하는 것은 추정의 정확성을 높일 수 있을 것이다.

본 논문에서는 모총합 추정의 정확성을 높이기 위해 절사층에서 일정 표본을 조사하고 모총합 추정에 사용하는 새로운 복합추정량을 제안하였다. 그 결과 조사변수와 보조변수의 관계가 비율형일때는 기존 방법처럼 BLUP이 가장 좋은 결과를 나타낸 반면 그 외의 형태에서는 SSW보다 새롭게 제시한 복합추정량의 결과가 모두 좋게 나왔다. 따라서 조사상황에 맞게 절사층에서 일정 수의 표본을 추출하고 본 논문에서 제시한 복합추정량을 사용한다면 기존보다는 더 우수한 총합 추정을 할 수 있을 것이다.

참고문헌

박인호 (2007). 절사표본조사에 관한 연구, <국민계정>, **30**, 81–106.

신민웅, 이상은 (2001). <표본조사를 위한 표본설계>, 교우사.

Benedetti, R., Bee, M. and Espa, G. (2010). A framework for cut-off sampling in business survey design, *Journal of Official Statistics*, **26**, 651–671.

Clark, K. K. and Kinyon, D. L. (2007). *Can We Continue to Exclude Small Single-Establishment Businesses from Data Collection in the Annual Retail Trade Survey and the Service Annual Survey?*, Presented at the Third International Conference on the Establishment Surveys in Montreal, Quebec, Canada.

Elisson, H. and Elvers, E. (2001). Cut-off sampling and estimation, *Proceedings of Statistics Canada Symposium on Achieving Data Quality in a Statistical Agency: A Methodological Perspective*, 2001.

Hidiroglou, M. A. (1986). The construction of a self-representing stratum of large units in survey design, *The American Statistician*, **4**, 27–31.

Lee, H., Rancourt, E. and Sarndal, C.-E. (1995). Experiment with variance estimation from survey data with imputed value, *Journal of Official Statistics*, **10**, 231–243.

Lee, S. E. (2011). The cut-off point based on MSE in modified cut-off sampling, *Journal of the Korean Official Statistics*, **16**, 82–94.

Lee, S. E. and Shin, K. I. (2011). Alternative determination of cut-off point based on MSE, *Proceedings of ISI 2011*, Dublin.

McDowney, P. (2004). Simulation Result of Probability Proportional Size Sampling for EIA's Monthly Natural Gas Production Survey', a summary note of the Fall 2004 meeting of the American Statistical Association committee on Energy Statistics. http://circa.europa.eu/irc/dsis/nacecpacon/info/data/en/handbook%20part3%20-%20 sampling%20and%20estimation.pdf.

Rao, J. N. K. (2003). *Small Area Estimation*, John Wiley and Sons, New York.

Sarndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*, Springer-Verlag, New York.

A Composite Estimator for the Take-Nothing Stratum of Cut-Off Sampling

Ji-Hak Kim[1] · Key-Il Shin[2]

[1]Department of Statistics, Hankuk University of Foreign Studies
[2]Department of Statistics, Hankuk University of Foreign Studies

(Received August 2011; accepted October 2011)

Abstract

Cut-off sampling that discards a part of the population from the sampling frame, is a widely used method for a highly skewed population like a business survey. Usually to the estimate of population total, we need to estimate the total of the take-nothing stratum. Many estimators have been developed to estimate the total of the take-nothing stratum. In this paper, we suggest a new composite estimator which combines the estimator suggested by Sarndal *et al.* (1992) and a ratio estimator obtained by small samples from the take-nothing stratum. Small simulation studies are performed for the comparison of the estimators and we confirm that the new suggested estimator is superior to the others.

Keywords: Sampling survey, non-sampling error, cut-off sampling, ratio estimator.

This research was supported by the research fund of Hankuk University of Foreign Studies(2011).
[2]Corresponding author: Professor, Department of Statistics, Hankuk University of Foreign Studies, Yongin, Kyonggi 449-791, Korea. E-mail: keyshin@hufs.ac.kr