연구논문

왜도가 심한 모집단의 절사층 추정*

Estimation of Cut-off Stratum in the Highly Skewed Population

한 근 식**

Geunshik Han

사업체조사에서 절사법이 흔히 이용되고 있다. 이 경우, 절사되는 사업체가 모집단에서 차지하는 비중이 모집단 전체에 비하여 매우 작다. 따라서 절사층에 속한 사업체는 목표모집단에서 제외되고, 나머지 부분에 대한 추정만이 이루어지고 있다. 이는 목표모집단 총계 추정에 영향을 작게 미치는 사업체 정보를 활용하지 않겠다는 의미이며, 더불어 작은 사업체에 응답 부담을 덜어 주겠다는 의도로 볼 수 있다. 그러나 예산의 부담을 덜기 위해서 절사층의 크기를 증가시키는 것은 목표모집단의 모수 추정에 상당한 편의를 제공한다는 점을 간과해서는 안 된다. 본 연구에서는 사례를 이용하여 모집단을 절사층, 표본층, 전수층으로 구분하고, 보조변수를 이용하여 절사층을 추정하는 방법을 보였다.

주제어: 전수층, 표본층, 절사층

In business survey, cut-off sampling is usual. The contribution from cut-off part of the population is at least small in comparison with the remaining population. In this case, part of the target population is excluded from the selection and parameter estimations are only based on Take-all and Take-some stratum. It may be tempting not to use resources on enterprises that contribute little to the overall results of the survey. And this reduces the response burden for these small enterprises. But, the size of cut-off stratum has been increased as a way to manage reduced budgets. This leads to additional bias. In this study, the population have been separated as three stratum, cut -off, take-some, take-all, and we will estimate cut-off part using auxiliary variable.

key words: take-all stratum, take-some stratum, cut-off stratum

* 본 논문은 2002년도 한신대학교 교내특별연구비 지원에 의해 연구되었음.

** 교신저자(corresponding author): 한신대학교 정보과학대학 교수 한근식.
E-mail: gshan@hs.ac.kr

I. 서론

사업체관련 표본조사에서 총계의 추정이 주 목적인 경우가 흔하다. 그러나 모집단 분포의 왜도가 심한 경우, 예를 들어, 반도체 산업체의 총 매출액, 철강업계의 총 종사자 수 등은 상위 몇 개의 사업체가 모집단의 총 매출액, 총 종사자 수 등에서 차지하는 비중이 80% ~90%에 이른다(한근식 외 1인 1996). 이러한 모집단을 대상으로 표본을 추출하여 총계를 추정하고자 할 때 주로 이용하는 표본설계방법이 절사법(cut-off method)이다. 이 방법은 층화추출의 특수한 경우로도 생각할 수 있는데, 모집단의 분포가 심한 왜도를 보이거나 소수의 모집단 요소들이 모집단 총계의 대부분을 차지하는 경우 네이만(Neyman)의 최적할당방법은 h번째 층에 할당된 표본의 크기, n_h가 해당 층의 모집단 크기, N_h보다 크게 되는 경우가 흔하다. 즉 $N_h < n_h$ 되는 경우가 많다. 이런 경우 h번째 층에 실제로 할당할 수 있는 표본의 크기는 $n_h = N_h$가 된다(Cochran 1977). 만약 층이 두 개이고 $N_1 < n_1$, $N_2 > n_2$ 이라면 첫 번째 층에 할당된 표본의 크기는 $n_1 = N_1$ 이 되므로 첫 번째 층은 전수조사가 이루어지며, 두 번째 층은 표본조사가 이루어진다. 이런 의미에서 첫 번째 층을 전수층(take-all stratum), 두 번째 층을 표본층(take-some stratum)이라고 부른다.

히디로글로우(Hidiroglou 1986)는 주어진 허용오차 하에서 모집단을 전수층과 표본층으로 구분하는 알고리즘을 제시하였으며, 라벨리와 히디로글로우(Lavallee & Hidiroglou 1988)는 모집단을 전수층과 2개 이상의 표본층으로 구분하여 추정하는 알고리즘을 개발하였다.

사업체 조사에서 절사법을 이용하는 경우 표본층의 크기는 전수층의 크기에 비해 상대적으로 큰 반면에 표본층이 총계에서 차지하는 비중은 상대적으로 작다. 즉, 모집단을 구성하는 사업체들을 관심변수의 크기순으로 정렬하면 하위그룹에 속하는 사업체들의 숫자는 많으나 그

들이 차지하는 총계에서의 비중은 미미하기 때문에 이 부분에 속하는 사업체들 중 하위 사업체들을 목표모집단(target population)에서 제외하고 표본설계를 한다. 이때 제외되는 사업체들이 속한 부분을 절사층이라 한다. 절사층에 속하는 사업체들을 표본조사에 포함하는 경우 응답자의 부담을 가중시켜 무응답을 유발하거나 불성실한 응답이 제공되어 이를 토대로 모집단의 총계를 추정할 때 비표본오차로 인해 과대 혹은 과소추정치를 제공하는 결과를 가져온다. 따라서 Canadian Monthly Survey of Manufacturing의 경우, 각 주(province)에서 생산액의 2%에 해당하는 최하위 사업체들을 절사한 나머지 사업체들을 목표모집단으로 설정하고 있다. 스웨덴에서 실시한 Mining, Quarrying and Manufacturing 실태조사의 경우, 종업원 수가 10인 이하인 사업체는 원 모집단에서 절사한 후 나머지 사업체들을 목표모집단으로 설정하여 조사하고 있다. 한편 통계청에서 시행하고 있는 건설업 통계조사(2002), 운수업 통계조사(2002) 등은 절사법을 이용하여 추정하고 있다. 위와 같은 조사는 히디로글로우(Hidiroglou 1988)나 라벨리와 히디로글로우(Lavallee and Hidiroglou 1988)가 제시한 방법과는 달리 모집단 전체를 추정하지 않고 표본층의 일부인 절사층에 속하는 사업체를 버리고 총계를 추정하고 있다.

II. 모집단과 절사법을 이용한 표본의 크기

본 연구의 대상모집단은 2002년 통계청에서 조사한 광업, 제조업 조사자료 중에서 한국표준산업분류상 기계관련 사업체들이며 모집단의 크기는 N=9,211개 사업체이다. 이 중 마이너스 생산액을 나타내는 4개 사업체를 제외한 9,207개 사업체가 본 연구에 이용되었다. 모집단 자료가 포함하는 이용가능한 정보는 사업체 코드, 생산액(단위 1,000원), 종업원 수(단위: 명) 등이다. 한편 해당 표본조사의 목적은

2003년도 기계관련 사업체들의 생산액을 추정하는 것이며, 9,207개 사업체의 총 생산액은 91,429,079,000원이다. <표 1>은 허용오차를 5%로 주었을 때 원 모집단에서 제외되는 부분의 크기에 따른 표본과 목표모집단의 크기 변화를 보여주고 있다.

이 모집단에서 절사하는 사업체가 없이 95%신뢰구간에서 허용오차를 5%로 하는 표본설계를 하는 경우 총 표본의 크기는 n = 348개 사업체였다. 이 중 전수층은 222개 사업체이며, 8,985개로 구성된 표본층에서 126개의 사업체가 표본층의 표본크기[1]로 결정되었다.

<표 1>에서 보는 바와 같이 생산액이 50만원 이하인 사업체를 절사한 후의 목표모집단은 원 모집단의 99.11%를 반영하고 있어 절사법을 이용한다 하더라도 목표모집단에서의 손실되는 부분은 미미하다. 그러나 목표모집단의 크기는 9,207개 사업체에서 5,193개 사업체로 약 45%가량 줄어들었다. 생산액이 작은 4,018개의 사업체를 절사함으로써 불성실한 응답과 무응답 등 표본조사에서의 비표본오차를 상당부분 줄일 수 있다.

<표 1> 허용오차 5%에서 절사점의 이동에 따른 표본의 크기

절사점 (1000원)	목표모집단 크기 (감소비율,%)	표본크기		모집단 총계 (생산액)	목표모집단의 총계 (감소비율, %)
		총계	전수층/ 표본층		
0	9,207*	348	222/126	91,429,079	
500	5,193(43.6)	273	182/91	90,618,278	0.89
1,000	3,592(60.9)	235	159/76	89,464,981	2.15
2,000	2,367(74.3)	197	128/69	87,738,319	4.04
3,000	1,826(80.2)	175	114/61	86,408,260	5.49
4,000	1,493(83.8)	159	104/55	85,250,835	6.76
5,000	1,265(86.3)	148	101/47	84,233,513	7.87
6,000	1,093(88.1)	138	96/42	83,291,202	8.91

1) 응용통계연구 9권 2호 163쪽, 표본의 크기 결정식 참조

이제 생산액이 100만원 이하인 사업체를 절사한 후 표본조사를 시행한다고 생각하자. 이때 목표모집단의 총계는 원 모집단 총계에서 2.15%가 감소한 값을 반영하게 된다. 그리고 목표모집단의 크기는 3,592개 사업체로 원 모집단 크기에 대해 60.9% 감소함을 볼 수 있다. 그 만큼 비표본오차도 줄어들 것이다. 이와 같이 절사점을 증가시켜 600만원 이하 사업체를 절사하는 경우, 목표모집단의 크기는 9,207에서 1,093개 사업체로 줄어들게 되어 표본의 추출을 용이하게 할 뿐만 아니라 비용의 절감, 비표본오차의 감소 등의 효과를 기대할 수 있다. 그러나 목표모집단 총계는 원 모집단 총계에서 8.91%가 제외된 것이다. 따라서 절사점이 큰 경우, 제외된 부분에 대한 추정을 고려할 필요가 있다.

III. 절사층의 추정

표본조사에서 가장 큰 문제 중의 하나는 무응답을 줄이는 것이다. 표본설계자는 무응답을 줄이기 위해 답례품의 질을 높이고, 조사원의 수를 증가시키며 그들에 대한 교육을 강화하고 있다. 이와 같이 응답자의 부담을 덜고 조사원의 교육을 위해 투입되는 비용은 표본의 크기가 줄어든 것 이상으로 지출되곤 한다. 그러나 증가된 비용만큼 양질의 응답이 제공되지는 않는다.

같은 방법으로, 조사연구자의 부담을 줄이는 방법은 없을까? 만약 <표 1>에서 보는 바와 같이 600만원 이하를 절사한 후 목표모집단에 대해 표본조사를 시행하고 600만원 이하의 부분을 추정할 수는 없을까? 절사된 부분 전체에 대한 추정이 어렵다면 일부분에 대한 추정은 가능한가? 이 질문에 대한 부분적인 답변이 본 연구의 목적이다.

본 연구에서는 활용가능한 적절한 보조변수를 이용하여 절사된 부분의 상당부분이 추정가능함을 보이고자 한다.

〈표 2〉는 절사점의 이동에 따른 목표모집단 내에서의 생산액과 종업원 수 간의 상관계수를 보여주고 있다.

〈표 2〉 절사점의 이동에 따른 생산액과 종업원 수 간의 상관계수

절사점	0	500	1,000	2,000	3,000	4,000	5,000	6,000
상관계수	0.500	0.495	0.496	0.488	0.485	0.478	0.472	0.465

〈표 2〉에서 보는 바와 같이 원 모집단에서 생산액과 종업원 수 간의 상관계수는 0.500이며, 생산액이 50만원 이하인 사업체를 절사한 후 목표모집단에서의 생산액과 종업원 수와의 상관계수는 0.495, 생산액이 100만원 이하인 사업체를 제외한 목표모집단에서의 생산액과 종업원 수와의 상관계수는 0.496 등이다.

절사(cut-off)층	표본(take-some)층	전수(take-all)층

〈그림 1〉 절사층을 포함한 모집단의 표현

〈그림 1〉에서 보는 바와 같이 표본층과 전수층의 생산액에 대한 추정은 절사법으로 가능하다. 이제 절사층의 추정은 〈표 2〉의 상관관계를 이용하여 비(ratio) 추정, 회귀추정 등의 방법을 이용하여 추정할 수 있다. 여기에서는 보조변수로서 각 사업체의 종업원 수를 이용하여 비추정을 하였으며, 비추정을 위해 절사층의 크기에 5%, 3%, 1% 등에 해당하는 표본을 단순임의추출하여 생산액과 종업원 수의 평균을 추정하고 총계와 표준편차를 추정하였다. 〈표 4〉부터 〈표 6〉까지는 이 과정을 5,000회 반복한 결과를 요약한 것이다. 여기에서 총계의 추정식 (Scheaffer, Mendenhakk & Ott 1990)은 $\hat{\tau_y} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i} \tau_x = r\tau_x$을 이용

하였으며, 총계의 추정치에 대한 분산의 추정식은 다음과 같다.

$$\hat{V}(\hat{\tau}_y) = \tau_x^2\left(\frac{N-n}{nN}\right)\left(\frac{1}{\mu_x^2}\right)\frac{\sum_{i=1}^{n}(y_i - rx_i)^2}{n-1}$$

〈표 4〉는 절사층을 500 이하, 1,000 이하, 2,000 이하, 3,000 이하, 4,000 이하, 5,000 이하, 6,000 이하로 결정하였을 경우, 각각의 절사층의 총계를 추정한 결과를 보여주고 있다.

〈표 4〉 절사층의 5%를 추출하는 경우 각 층별 비추정치

절사층 범위	N	n	절사층 총계	CV
0 - 499	4,016	201	840,870	0.14
0 - 999	5,602	281	2,030,303	0.13
0 - 1999	6,812	343	4,151,116	0.08
0 - 2999	7,339	370	5,283,737	0.11
0 - 3999	7,663	386	6,266,464	0.09
0 - 4999	7,878	398	7,080,196	0.10
0 - 5999	8,042	406	8,427,040	0.08

〈표 5〉 절사층의 3%를 추출하는 경우 각 층별 비추정치

절사층 범위	N	n	절사층 총계	CV
0 - 499	4,016	121	973,073	0.14
0 - 999	5,602	169	1,972,598	0.16
0 - 1999	6,812	206	3,562,633	0.19
0 - 2999	7,339	222	5,868,470	0.10
0 - 3999	7,663	232	6,694,903	0.11
0 - 4999	7,878	239	7,681,390	0.10
0 - 5999	8,042	244	8,850,129	0.10

〈표 6〉 절사층의 1%를 추출하는 경우 각 층별 비추정치

절사층 범위	N	n	절사층 총계	CV
0 - 499	4,016	41	915,371	0.23
0 - 999	5,602	57	2,151,065	0.19
0 - 1999	6,812	69	3,380,359	0.25
0 - 2999	7,339	74	5,627,426	0.20
0 - 3999	7,663	78	6,895,240	0.17
0 - 4999	7,878	80	6,656,054	0.27
0 - 5999	8,042	82	8,238,821	0.16

IV. 결론 및 제언

〈표 4〉부터 〈표 6〉에서 보는 바와 같이 모집단을 전수층, 표본층, 그리고 절사층으로 구분하여 절사층의 총계를 보조변수를 이용하여 추정하였다. 각 층의 크기에 따라 차이는 있지만 절사층 크기의 5%정도의 표본추출이 가능하고 적절한 상관관계를 갖는 보조변수의 활용이 가능하다면 절사층의 총계추정이 가능하다는 것을 알 수 있다. 5%와 3% 표본 추출의 경우 3,000 이하를 절사층에 포함시켰을 때 CV값이 0.1정도였으나 1% 표본 추출의 경우 CV값이 0.2정도로, 조사의 목적과 요구되는 정도(precision)에 따라 절사층의 표본크기를 결정할 필요가 있다.

한국표준산업 분류상, 대분류에 속하는 중분류들 각각을 절사법을 이용하여 추정할 때, 절사층을 추정하지 않고 버리게 되면 중분류들의 합으로 이루어진 대분류의 총계는 실제 모집단을 지나치게 과소추정(under estimation)하게 된다. 이러한 경우 본 논문에서 제시한 방법을 이용하면 과소추정을 극복할 수 있을 것으로 기대된다.

위와 같은 방법으로 절사층을 추정하는 과정에서 보조변수의 활용이 가장 중요한 것으로 판단되었다. 본 연구의 경우, 각 사업체의 총 종업원 중 해당 품목 생산에 종사하는 종업원 수가 파악되지 않아 표본조사 후 3주 간에 걸쳐 전화조사하여 보조변수로 활용하였다. 보조변수 활용에서 또 다른 문제는 극단값이 존재하는 경우인데, 이는 추정치를 크게 왜곡시키는 요인이 된다. 이 문제 해결을 위해서는 훈련된 조사원은 물론이고 응답자의 성실한 응답이 추정치의 정도를 높이는 관건이기도 하다.

참고문헌

한근식 · 김용철. 1996. "왜도가 심한 모집단에서의 절사법 효과에 관한 연구." «응용통계연구» 9(2): 161-169.

통계청. 2002. «광업, 제조업 통계조사보고서». 통계청.

통계청. 2002. «기준 건설업 통계조사». 통계청.

통계청. 2002. «운수업 통계조사». 통계청.

Cochran, W. G. 1977. *Sampling Techniques*(2nd ed.). Wiley & Sons

Hidiroglou, M. A. 1986. "The construction of a self representing stratum of large units in survey design." *The American Statistician* 40: 27-31.

Lavallee, P. and Hidiroglou, M. A. 1988. "On the Stratification of Skewed Populations." *Survey methodology* 14: 33-43.

Scheaffer, Mendenhall, Ott. 1990. *Elementary Survey Sampling*(3rd ed.). Duxbury Press

Statistics Canada. 2001. "Monthly Survey of Manufacturing." *Statistical Data Documentation System*, Statistics Canada.

Statistics Sweden. 2000. "Swedish Monthly Survey of Mining, Quarrying and Manufacturing." *Statistics Sweden*.