응용통계연구 제 8권 1호
1995년 3월 pp. 1-7.

낮은 모비율 추정을 위한 표본추출방법[1)]

김 지 현[2)]

요 약

표본조사에서 추정하고자 하는 모비율이 낮으면 많은 표본을 추출해야 한다. 이 때 단위 표본의 비용이 높아 추출할 수 있는 표본의 크기가 제한된다면 추정량의 분산을 줄일 수 있는 방법을 신중히 고려해야 할 것이다. 본 논문에서는 모집단에 대한 사전 정보가 있어 이를 이용하여 층화추출하는 경우에 층의 표본크기의 새로운 배분방법을 제안하고 기존의 배분방법들과 비교해 본다.

1. 문제의 배경과 새로운 표본추출방법의 제안

유한 모집단에서 어떤 특성을 갖는 원소들의 비율을 추정하는 문제에서 그 비율이 매우 낮을 것이 (예를 들어 <0.05) 예상될 때, 모비율의 신뢰구간이 의미를 갖기 위해서는 많은 표본을 추출해야 한다. 이 때 단위 표본의 비용이 높은 경우, 필요한 정도(precision)를 보장하는 크기 만큼의 표본을 추출할 수 없다는 현실적 어려움에 직면하게 된다. 하지만 모집단에 대한 사전 정보가 있다면 이를 이용하여 층화를 하면 추정량의 정도를 높일 수 있다. 본 연구에서는 이러한 낮은 모비율의 추정을 위한 층화추출의 경우에, 최적배분과 비례배분 등의 표본크기 배분 방법보다 더 효과적인 표본크기 배분 방법을 제안하고 다른 방법들과 비교해 보고자 한다.

주어진 문제를 구체적인 예를 들어 설명해 보면 다음과 같다. ㄱ폐기물 매립장에는 트럭 단위로 폐기물이 반입된다. 규정에 따라 한 트럭에는 한 배출업소에서 나온 동일한 종류의 폐기물만 적재하도록 되어 있다. 따라서 모집단은 일정 기간 반입된 모든 폐기물이 될 것이며, 모집단의 기본단위는 한 트럭에 적재된 폐기물로 정의할 수 있다. 정해진 기간에 (예를 들어 93년 3월 한 달) 반입되는 전체 폐기물 중에 중금속의 함유량이 일정 기준치를 넘는 폐기물의 비율이 얼마인가를 추정하는 문제를 생각해 보자. 반입되는 폐기물에 관한 기초 정보(배출업소, 운반업체, 폐기물의 종류 등)는 트럭 단위의 모든 폐기물에 대해 갖고 있다. 그리고 ㄱ폐기물 매립장 관리 사업소는 92년 6월 매립장을 개장한 이후부터, 반입되는 산업 폐기물에 대해서 검사를 실시해 왔는데, 처음으로 반입되는 배출업소인 경우, 검사를 연속 2회 내지 3회 실시하고 그 외에 의심이 갈 때나 필요할 때 일부 검사를 해 왔으며 그 결과를 보관하고 있다. 여기서 1회 검사란 한 트럭에 실려있는 폐기물을 대표할 수 있는 시료를 채취해, 수은, 납, 시안, 구리, 카드뮴, 크롬, 비소 등 7가지 중금속의 함유량을 측정하는 것을 말한다. 자세한 시료 채취 및 검사 방법은 환경처 고시 제 91-85호에 정해져 있으며, 이 7가지 중금속 중에서 한 가지라도 기준치를 초과하면 특정 폐기물로 분류된다. 하지만 이 검사된 표본은 첫째, 확률 표본이 아니고 둘째, 검사의 객관성에 이의가 제기되었기 때문에, 모비율의 추정을 위해서 다시 표본을 추출하

1) 본 연구는 1994년도 숭실대학교 연구비 지원을 받아 수행되었습니다. 건설적인 조언을 주신 심사위원께 감사드립니다.

2) (156-743) 서울시 동작구 상도5동 1-1, 숭실대학교 통계학과.

여 검사하기로 하였다. 그 동안의 검사 결과를 보면 그 객관성의 결여를 감안하더라도 모비율은 매우 낮을 것이 예상된다. 단순확률 복원추출을 가정하고 적절한 정도를 보장하기 위한 표본의 크기를 계산하여 보았다. 모비율 p의 신뢰구간이 0을 포함하지 않으려면, 식 $\hat{p}\pm 2\sqrt{(\hat{p}\hat{q})/n}$에서, $\hat{p}=0.02$인 경우와 $\hat{p}=0.05$인 경우 표본의 크기가 각각 196과 76이상일 것이 요구된다. 따라서 단위 표본의 검사 비용이 높고 예산이 한정되어 100개 이상의 표본은 불가능하다고 할 때 추정량의 정도를 높일 수 있는 방법이 절실히 요구된다.

일반적으로 단순확률추출보다는 층화추출을 할 경우 추정량의 정도를 높일 수 있음이 알려져 있다. 매립지의 개장 이후 신규로 반입된 모든 배출업소의 폐기물에 대한 검사 결과가 적어도 2개에서 많게는 수십 개까지 있으므로, 지금까지 반입된 적이 없는 새로운 배출업소를 제외한 모든 배출업소에 대한 사전 정보가 있는 셈이며, 이를 이용하면 중금속의 기준치 초과 위험에 대해 안전한 업소와 그렇지 못한 업소로 분류할 수 있을 것이다. 즉, 93년 6월 한 달 동안 트럭 단위로 반입되는 폐기물을 배출업소에 따라 안전층과 비안전층으로 나누어 층화추출을 고려할 수 있을 것이다. 이와 같이 두 개의 층으로 나눌 수 있었다고 할 때, 각 층의 표본 크기 배분 방법에 따라 추정량의 정도가 달라진다. 층의 크기와 분산을 동시에 고려하는 최적배분(Neyman배분)이 가장 높은 정도를 가져 온다고 알려져 있으나, 이것은 아래 식 (2.1)과 같이 추정량의 형태에 대해 제한을 두었을 때에 성립하는 사실이다. 이 예에서 만약 안전층에서 추출된 표본의 중금속 함유량이 기준치를 넘을 확률을 거의 무시할 수 있다면 (비록 과거의 검사 결과의 객관성에 문제가 있다고 하더라도 함유량의 많고 적음의 상대적 크기는 믿을 수 있다고 여겨지고 또한 배출업소와 폐기물의 종류로부터 폐기물이 안전하다고 판단되기도 하므로 안전층의 모비율은 거의 0에 가까울 것이다) 이 안전층에서 표본을 추출하는 것은, 비싼 표본의 낭비라고 생각된다. 따라서 예산의 범위 내에서 가능한 모든 표본을 비안전층에서만 추출한다면 최적배분에 의한 추정량보다 정도가 더 높은 추정량을 얻을 수 있으리라 판단되어, 새로운 표본추출방법으로 제안한다. 다음 두 절에서는 새로이 제안된 표본추출방법에 의한 추정량이 최적배분에 의한 추정량에 비해 어떤 장단점이 있는지 알아본다.

2. 표본추출방법

앞 절에서 제기된 문제를 요약해 보면, 낮은 값의 모비율 추정 문제에서 모집단을 두 개의 층으로 나눌 수 있고 한 층의 모비율이 0에 가깝다면 상대적으로 높은 비율을 가진 층에서만 표본을 추출할 것을 제안하였다. 새로이 제안한 표본추출방법은 층화추출이면서 표본을 한 층에만 일방적으로 배분하므로 **일방배분**이라고 부르기로 한다. 이 방법을 다른 방법들과 비교해 보기 위해 주어진 문제를 모형화 해 보자.

모집단의 두 개의 층을 S_1, S_2, 각 층의 크기를 N_1, N_2, 그리고 각 층의 모비율을 p_1, p_2로 표시한다. 층의 크기는 알려져 있고 $N=N_1+N_2$라고 할 때 추정하고자 하는 모수는 모비율

$$p=\frac{N_1}{N}p_1+\frac{N_2}{N}p_2$$

이다. n은 총 표본의 크기이고 n_1, n_2는 각 층에서 취한 표본의 크기이며 $n=n_1+n_2$ 이다. 수

학적 편의를 위해 복원추출을 가정하는데, 이 가정은 비복원추출이더라도 모집단의 크기가 표본의 크기보다 충분히 크다면 근사적으로 만족된다. 표본추출방법의 비교를 위해 크게는 단순확률추출과 층화추출로 구분하고 층화추출에서 비례배분, Neyman배분, 일방배분으로 다시 구분하여, 총 네 가지 추출방법에 대해 각 추정량의 형태와 기대값, 분산을 구해 보자.

단순확률추출

층의 구분없이 전체 모집단에서 n개를 단순확률추출하여 그 중 중금속 함유량이 기준치를 넘는 표본의 비율을 $\hat{p}_R$라고 하면, $\hat{p}_R$은 모수 p의 불편추정량이고 분산은

$$V(\hat{p}_R) = \frac{p(1-p)}{n}$$

이다.

층화추출; 비례배분

층화추출의 경우, X_i를 층 i에서 추출한 n_i개의 표본 중에서 중금속 함유량이 기준치를 넘는 표본의 수라고 할 때, 복원추출의 가정하에 X_i의 분포는 모수가 n_i, p_i 인 이항분포이다. 또한 층화추출에서 $n_1>0$, $n_2>0$ 이면 표본크기의 배분 방법에 상관없이

$$\hat{p} = \frac{N_1}{N}\frac{X_1}{n_1} + \frac{N_2}{N}\frac{X_2}{n_2} \tag{2.1}$$

으로 나타낼 수 있다. 이 추정량은 불편추정량이며 분산은

$$V(\hat{p}) = \left(\frac{N_1}{N}\right)^2 \frac{p_1(1-p_1)}{n_1} + \left(\frac{N_2}{N}\right)^2 \frac{p_2(1-p_2)}{n_2} \tag{2.2}$$

이다. 만약 각 층에서의 표본크기를 층의 크기에 비례하도록 하여 층화추출했을 때의 추정량을 $\hat{p}_P$라고 하면, $n_1 = nN_1/N$, $n_2 = nN_2/N$ 이므로 식 (2.2)로부터

$$V(\hat{p}_P) = \frac{1}{n}\left(\frac{N_1}{N}p_1(1-p_1) + \frac{N_2}{N}p_2(1-p_2)\right)$$

이다.

층화추출; Neyman배분

Neyman배분은 층의 크기와 분산을 같이 고려하여 표본크기를 배분하는 방법으로서 각 층에서의 표본크기는

$$n_i = n\frac{N_i\sqrt{p_i(1-p_i)}}{\sum_{i=1}^{2} N_i\sqrt{p_i(1-p_i)}}, \quad i=1, 2 \tag{2.3}$$

이다 (Cochran(1977) 식 (5.60) 참조). Neyman배분에 의한 추정량을 $\hat{p}_N$이라고 하면 분산은 식 (2.2)와 (2.3)으로부터

$$V(\hat{p}_N) = \left(\frac{N_1}{N}\sqrt{\frac{p_1(1-p_1)}{n}} + \frac{N_2}{N}\sqrt{\frac{p_2(1-p_2)}{n}} \right)^2 \qquad (2.4)$$

임을 보일 수 있다.

<u>층화추출;일방배분</u>

일방배분이란 ㄱ폐기물 매립장의 예와 같이 $p_1 < p_2$ 이고 $p_1 \approx 0$ 일 때 S_2에서만 표본을 추출하는 방법으로서 $n_1=0$, $n_2=n$ 이다. 따라서 일방배분에 의한 추정량을 $\hat{p}_O$라고 하면

$$\hat{p}_O = \frac{N_2}{N}\frac{X_2}{n} \qquad (2.5)$$

으로서 편의(bias)가 $p_1(N_1/N)$ 인 편의추정량임을 알 수 있다. 한편 분산은

$$V(\hat{p}_O) = \left(\frac{N_2}{N}\right)^2 \frac{p_2(1-p_2)}{n} \qquad (2.6)$$

이다.

3. 표본추출방법의 비교 및 결론

Neyman배분에 의해 표본크기를 배분하려면, 알려져 있지 않은 모수 p_1과 p_2 또는 층의 분산이 필요한데 이 값이 주어졌을 때 (현실적으로는 근사한 추정값이 있을 때) Neyman배분에 의한 추정량은 단순확률추출이나 비례배분에 의한 추정량보다 작은 분산을 갖는다. (한편, Stevens Jr. 등은(1991) 잘못된 모수의 값으로 Neyman배분을 했을 때 그 추정량의 정도가 비례배분에 의한 추정량의 정도보다 나빠질 수 있다는 것을 구체적 예를 들어 지적하였다.) 따라서 $\hat{p}_R$, $\hat{p}_P$를 다음 정리의 비교 대상에서 제외하였다. 편의추정량인 $\hat{p}_O$와 불편추정량인 $\hat{p}_N$의 정도를 비교하기 위해서 평균제곱오차(MSE, Mean Squared Error)를 판단기준으로 하였다.

정리 $p_1 < \frac{1-p_1}{n}$ 이면 $\mathrm{MSE}(\hat{p}_O) < \mathrm{MSE}(\hat{p}_N)$ 이다.

증명 식 (2.4)로부터

$$\mathrm{MSE}(\hat{p}_N) = V(\hat{p}_N) = \left(\frac{N_2}{N}\right)^2 \frac{p_2(1-p_2)}{n} + \left(\frac{N_1}{N}\right)^2 \frac{p_1(1-p_1)}{n} + 2\frac{N_1N_2}{nN^2}\sqrt{p_1(1-p_1)}\sqrt{p_2(1-p_2)}.$$

식 (2.5)와 (2.6)으로부터

$$\mathrm{MSE}(\hat{p}_O) = V(\hat{p}_O) + (\mathrm{E}(\hat{p}_O) - p)^2$$
$$= \left(\frac{N_2}{N}\right)^2 \frac{p_2(1-p_2)}{n} + \left(\frac{N_1}{N}\right)^2 p_1^2.$$

$$\mathrm{MSE}(\hat{p}_N) - \mathrm{MSE}(\hat{p}_O) = \left(\frac{N_1}{N}\right)^2 p_1\left(\frac{1-p_1}{n} - p_1\right) + 2\frac{N_1 N_2}{nN^2}\sqrt{p_1(1-p_1)}\sqrt{p_2(1-p_2)}.$$

따라서 조건 $p_1 < (1-p_1)/n$은 $\mathrm{MSE}(\hat{p}_O) < \mathrm{MSE}(\hat{p}_N)$ 이기 위한 충분조건이다. ■

MSE의 기준에서 $\hat{p}_O$가 $\hat{p}_N$보다 좋아지기 위한 조건 $p_1 < (1-p_1)/n$ 은 표본의 크기 n이 그다지 크지 않고 $p_1 \approx 0$ 인 경우 쉽게 만족되는 조건이다. $N_1 = 2{,}000$, $N_2 = 1{,}500$ 이고 $n = 100$ 인 경우 (ㄱ폐기물 매립장에서의 예에서 적용되는 값이다), 몇 가지 가능한 값의 p_1, p_2 에 대해서 두 추정량의 MSE를 비교해보면 <표 3.1>과 같다.

<표 3.1> $\mathrm{MSE}(\hat{p}_O)$의 상대적 감소
($N_1 = 2{,}000$, $N_2 = 1{,}500$, $n = 100$)

p_1	p_2	$\frac{\mathrm{MSE}(\hat{p}_N) - \mathrm{MSE}(\hat{p}_O)}{\mathrm{MSE}(\hat{p}_N)}$
.005	.05	.466
.01	.05	.469
.02	.05	.275
.005	.10	.392
.01	.10	.424
.02	.10	.320

<표 3.1>에서 알 수 있는 바와 같이 조건 $p_1 < (1-p_1)/n$ 이 만족되는 경우는 물론, 만족되지 않는 일부 경우에도 $\mathrm{MSE}(\hat{p}_O)$는 $\mathrm{MSE}(\hat{p}_N)$에 비해 약 30-40% 의 상대적 감소를 나타낸다.

Neyman배분을 위해서는 알려져 있지 않은 모수 $p_1(>0)$, $p_2(>0)$의 추정값이 필요한데 ㄱ폐기물 매립장의 예에서와 같이 층 S_1의 특성상 사전정보에 의한 추정량 $\hat{p}_1$은 0인 것이 보통이다. 따라서 Neyman배분을 위해서는 $\hat{p}_2$와의 상대적 크기를 고려하여 주관적으로 결정한 p_1의 값을 사용하여야 할 것이다. <표 3.2>와 <표 3.3>에서는 Neyman배분에서 층의 비율 p_1, p_2 의 값을 잘못 가정하여 표본 배분을 하였을 때, 여러 가지 표본추출방법에 따른 추정량들의 표준오차를 비교해 보았다.

<표 3.2> Neyman배분에서 가정된 p_1, p_2 의 값이 각각 0.005, 0.05 일 때 실제 p_1, p_2 값의 변화에 따른 여러 추정량의 비교 (N_1=2,000, N_2=1,500, n=100)

			단순확률추출	비례배분	Neyman배분*	일방배분	
p_1	p_2	p	S.E.($\hat{p}_R$)	S.E.($\hat{p}_P$)	S.E.($\hat{p}_N$)	S.E.($\hat{p}_O$)	bias($\hat{p}_O$)
.0005	.05	.0217	.0146	.0144	.0114	.0093	-.0003
.0010	.05	.0220	.0147	.0145	.0116	.0093	-.0006
.0050	.05	.0243	.0154	.0152	.0134	.0093	-.0029
.0100	.05	.0271	.0162	.0161	.0152	.0093	-.0057

* Neyman배분에 의한 n_1, n_2의 값은 각각 30, 70 이다.

<표 3.3> Neyman배분에서 가정된 p_1, p_2 의 값이 각각 0.005, 0.10 일 때 실제 p_1, p_2 값의 변화에 따른 여러 추정량의 비교 (N_1=2,000, N_2=1,500, n=100)

			단순확률추출	비례배분	Neyman배분*	일방배분	
p_1	p_2	p	S.E.($\hat{p}_R$)	S.E.($\hat{p}_P$)	S.E.($\hat{p}_N$)	S.E.($\hat{p}_O$)	bias($\hat{p}_O$)
.0005	.10	.0431	.0203	.0197	.0150	.0129	-.0003
.0010	.10	.0434	.0204	.0198	.0152	.0129	-.0006
.0050	.10	.0457	.0209	.0204	.0169	.0129	-.0029
.0100	.10	.0486	.0215	.0210	.0188	.0129	-.0057

* Neyman배분에 의한 n_1, n_2의 값은 각각 24, 76 이다.

<표 3.1>과 <표 3.2>에서 알 수 있는 바와 같이 일방배분에 의한 추정량 $\hat{p}_O$는 편의추정량이지만 p_1의 값이 0에 가까울수록 편의의 크기는 무시될 수 있고 표준오차가 다른 추정량들에 비해 상대적으로 크게 줄어든다.

이상의 비교 결과 다음과 같이 결론 지을 수 있을 것이다. 낮은 모비율의 추정을 위한 표본조사에서 사전정보를 이용하여 두 층으로 층화를 할 때 안전층 S_1의 위험확률 p_1을 3절의 정리에서와 같이 충분히 낮게 할 수만 있다면 (ㄱ폐기물 매립장의 예에서 이는 현실적으로 충분히 가능하다) 상대적으로 위험이 높은 층 S_2에서만 표본을 추출하는 일방배분의 표본추출방법이 권장되어야 한다.

참고문헌

[1] Cochran, W. G. (1977). *Sampling Techniques*, 3rd edition, John Wiley & Sons, New York.

[2] Stevens, Jr. D. L. and Olsen, A. R. (1991). Statistical issues in environment monitoring and assessment, *1991 Proceedings of the section on statistics and the environment*, American Statistical Association.

A Sampling Scheme for the Estimation of Low Proportion

Ji-Hyun Kim[3)]

Abstract

In sample survey for the estimation of low proportion, usually a large size of sample is required for a meaningful estimator. If the cost of a sample unit is high, we have to make every effort to improve the precision of the estimator. In this study, a new efficient allocation method of sample size in stratified sampling is proposed provided we have some prior information for the stratification.

3) Department of Statistics, Soong Sil University, Sangdo 5 Dong 1-1, Dongjakku, Seoul, 156-743, KOREA.