The Korean Communications
in Statistics Vol. 6, No. 3, 1999
pp. 749-753

Optimal Allocations in Two-Stage Cluster Sampling[1)]

BongSung Koh[2)]

Abstract

The cost is known to be proportional to the size of sample. We consider a cost function of the form, $Cost = c_1 n^p + c_2 n^p m^q$, where c_1, c_2, p, and q are all positive constants. This cost function is to be used in finding an optimal allocation in two-stage cluster sampling. The optimal allocations of n and m gives the properties of uniqueness under some conditions and of monotonicity with $p>0$, when $q=1$.

1. 서론

표본의 최적배분(optimal allocation)은 책정된 예산 또는 표본분산하에 모집단의 특성을 가장 잘 나타내는 최적의 표본 크기를 결정하는 것이다. 이중 이단집락추출법의 최적배분에 관한 문제는 Cochran(1977), Jessen(1978), Snedecor와 Cochran(1980), 그리고 Sokal와 Rohlf(1981)등에 의해 잘 알려져 있으며, 이들은 선형비용함수를 이용한 최적배분을 생각하였다. 이와는 달리 Hansen, Hurwitz와 Madow(1973), Beardwood, Halton 그리고 Hammersley(1959)등은 표본추출단위들간에 여행비용을 포함된 비용함수를 생각하였으며, Somerville(1970)은 일차추출단위 표본크기의 α승에 비례하는 비용함수를 사용하였다.

본 연구에서는 이들 비용함수를 보다 일반화한 비용함수를 고려하였다. 일차추출단위와 이차추출단위의 표본추출에 여행비용이 존재하는 경우, 일차추출단위나 이차추출단위의 크기가 크거나 작게 증가함에 따라 비용이 빠르게 또는 천천히 증가하는 비용함수를 고려하였다.

$$C = c_1 n^p + c_2 n^p m^q, \tag{1}$$

여기서 c_1, c_2, p 와 q 는 양수이며, n과 m은 이단집락추출법에서 일차추출단위와 이차추출단위의 크기이다.

이 비용함수 식 (1)을 살펴보면 일차추출단위와 이차추출단위의 승수 p, q에 따라 표본추출비용이 빠르게 또는 느리게 변하는 것을 알 수 있다. 즉 일차추출단위의 승수 p가 $0<p<1$과 $0<q<1$사이에 있을 때 비용함수는 일차추출단위와 이차추출단위의 크기가 증가함에 따라 천천히 증가한다. $p>1$, $q>1$인 경우에 비용은 일차와 이차추출단위의 크기에 따라 급격히 증가하는 함수이다. 그리고 $0<p<1$과 $q>1$인 경우에 비용함수는 일차추출단위의 표본크기는 천천히 증가하고 이차추출단위의 표본크기는 빠르게 증가하게 된다.

본 논문은 식 (1)과 같이 정의된 비용함수를 이용하여 이단집락추출법의 최적배분의 문제를 생

1) This paper was supported by Jeonju University Research Fund, 1999.
2) Assistant Professor, Department of Information Statistics, Jeonju University, Chonju, 560-759, Korea

각하고 있다. 즉 식 (1)의 비용함수를 고려하였을 때 어떤 조건하에 최적배분이 존재하는가, 또한 $q=1$일 때 기대되는 일차추출단위의 크기 p의 최적배분을 살펴보고 있다.

2. 최적배분

Cochran(1977, p277)은 이단집락추출법의 전체표본평균 $\bar{\bar{y}}$의 분산은 일차추출단위 평균간의 모분산 S_1^2과 일차추출단위내에서 요소간의 모분산 S_2^2의 함수형태로 다음과 같음을 보였다.

$$V(\bar{\bar{y}}) = \frac{(1-f_1)\,S_1^2}{n} + \frac{(1-f_2)\,S_2^2}{nm} \quad , \tag{2}$$

여기서 f_1과 f_2는 일차추출단위와 이차추출단위의 표본추출율로, 만약 f_1과 f_2가 작다면 전체표본평균의 분산은 다음과 같이 같단한 형태가 된다.

$$V(\bar{\bar{y}}) = \frac{S_1^2}{n} + \frac{S_2^2}{nm}, \tag{3}$$

본 논문에서는 식 (3)의 분산을 이용하고 있다.

최적배분은 요구분산 V_0를 고정시켰을 때 비용함수를 최소화하는 것으로 n과 m의 조합으로 나타난다. 여기서 비용함수 (1)을 이용한 최적배분은 라그랑즈 승수법을 이용하여 다음과 같은 함수를 정의할 수 있다.

$$f(n,m) = c_1 n^p + c_2 n^p m^q + \lambda\left(\frac{S_1^2}{n} + \frac{S_2^2}{nm} - V_o\right) \quad ,$$

여기서 λ는 상수이다.

이차추출단위의 최적배분 m^*를 구하기 위해 위 식을 m에 대해 정리하면 다음과 같다.

$$qc_2\,S_1^2 m^{q+1} + (q-p)\,c_2 S_2^2 m^q - pc_1 S_2^2 = 0 \quad . \tag{4}$$

즉, 본 논문에서 제시한 비용함수를 이용한 경우의 이단집락추출법의 최적배분은 다음과 같은 결과하에 존재하게 된다.

결과 1.

만약 $p \le q$이면 일정한 최적배분이 존재한다.

증명)

식 (4)를 m에 대해 일차미분을 하면 우측부분은 비음이며, m^*는 임의의 양수이다. 그리고 식 (3)에 m^*를 대입하면 일차추출단위의 최적배분 n^*을 얻을 수 있다.

식 (4)를 m에 대해 일차 미분을 하면 다음과 같다.

$$q(q+1)c_2\,S_1^2 m^q + q(q-p)c_2\,S_2^2 m^{q-1}$$

$$= qc_2m^{q-1}[(q+1)\ S_1^{\ 2}m+(q-p)\ S_2^{\ 2}] \tag{5}$$

$$\geq 0\ ,\ \text{if}\ q\geq p$$

그리고 결과 1에 조건을 추가한다면, 다음과 같은 결론들을 내릴 수 있다.

결과 2.

만약 $S_1^2\geq S_2^2$이고 $2q\geq p-1$이라면 여기에는 일정한 최적배분이 존재한다.

증명)

$S_1^2\geq S_2^2$이기 때문에 식 (5)는 다음 식보다 크거나 같게 되며

$$qc_2S_2^2m^{q-1}[2q-p+1]\ ,$$

만약 $2q\geq p-1$이라면 위 식은 비음이다.

한편 본 논문에서 제시한 비용함수에 $q=1$라 하면 비용함수는 다음과 같다.

$$Cost=c_1n^p+c_2n^pm$$

이 비용함수는 m보다는 n의 비용변화율(승수)에 밀접한 관계가 있게 된다. 만약 비용과 m의 관계처럼 하나의 기준을 잡는다면, n으로 인한 비용의 변화는 $0<p<1$ 이거나 $p>1$ 일 때 더 작거나 커지게 된다. 이러한 제약은 식 (4)에서 m의 최적배분을 다음과 같이 쉽게 찾을 수 있다.

$$m^*=\frac{c_2\ S_2^{\ 2}(p-1)+\sqrt{\ c_2^{\ 2}\ S_2^{\ 4}\ (p-1)^2+4c_1\ S_1^{\ 2}c_2\ S_2^{\ 2}p}}{2c_2\ S_1^{\ 2}}\ . \tag{6}$$

여기서 일차추출단위의 최적배분은 다음과 같다.

$$n^*=(\ S_1^{\ 2}+\frac{S_2^{\ 2}}{m^*})/V_0\ .$$

결과 3.

비용함수 $C=c_1n^p+c_2n^pm$ 에서 이차추출단위의 최적배분은 일차추출단위의 비용변화율 p의 증가함수로, 다음과 같은 조건을 따른다.

1. $p\geq 1$
2. $c_2S_2^2>c_1S_1^2$

증명)

식 (6)을 p에 대해 미분을 하면 다음과 같다.

$$\frac{S_2^2[\,c_2^2\ S_2{}^4(p-1)^2+4c_1\ S_1{}^2c_2\ S_2{}^2p]^{1/2}+c_2S_2^4(p-1)+2c_1S_1^2S_2^2}{2S_1^2[\ c_2{}^2\ S_2{}^4\ (p-1)^2+4c_1\ S_1{}^2c_2\ S_2{}^2p]^{1/2}} \tag{7}$$

식 (7)은 비음이므로 정리를 하면

$$\frac{S_2^2c_1(c_2S_2^2-c_1S_1^2)}{S_1^2c_2[\,c_2S_2^2(p-1)^2+4c_1S_1^2p]} \geq 0 \tag{8}$$

이 된다. 즉 식 (8)은 $p \geq 1$ 이고 $c_2S_2^2 > c_1S_1^2$일 때 양수가 된다.

3. 결론 및 논의

본 논문에서 제시한 비용함수에서 p와 q가 모두 1이라면, 이차추출단위의 최적배분은 다음과 같으며, 표본의 크기가 양수임은 이미 잘 알려져 있다.

$$m^*_{(p=1)}=\sqrt{\frac{c_1\ S_2{}^2}{c_2\ S_1{}^2}}\ ,$$

여기서 $c_1S_2^2 \geq c_2S_1^2$ 이므로 다음과 같은 관계를 얻을 수 있다.

$$\frac{c_1}{c_2} > \frac{S_1{}^2}{S_2^2}\ .$$

이는 집락내의 표본추출단위가 동질적일때($S_1{}^2 \geq S_2{}^2$) 이차추출단위의 크기는 작게하고 일차추출단위의 크기는 크게 하여, 정도를 향상시키는 Yamane(1967)등의 연구 결과와 일치한다.

그러나 비용함수에서 $q=1$ 이고 $p>1$ 이라면, 이차추출단위의 최적배분 m^* 는 일차추출단위 p 의 증가함수이며, 이는 이차추출단위 보다는 일차추출단위로 인한 비용의 변화가 크다는 것을 알 수 있다. 따라서 이와 같은 상황은 이차추출단위의 비용이 일차추출단위의 비용보다 더 큰 경우의 최적배분으로, 선형인 비용함수 $Cost=c_1n+c_2nm$ 와 다른 다음의 결과를 갖는다.

$$\frac{c_2}{c_1} > \frac{S_1{}^2}{S_2^2}\ .$$

한편 본 논문에서 사용한 비용함수의 p를 2, 1, 1/2로 생각하면 다음과 같다.

$$Z_1=c_1n^2+c_2n^2m\ .$$

$$Z_2=c_1n+c_2nm\ .$$

$$Z_3=c_1\sqrt{n}+c_2\sqrt{n}m\ .$$

여기서 Z_1은 이차추출단위 보다는 일차추출단위에 의해 비용이 급격히 증가하는 비용함수이며, Z_2는 일차추출단위와 이차추출단위가 선형인 비용함수, 그리고 Z_3은 일차추출단위의 크기에 따라 비용이 완만히 증가하는 비용함수로 생각할 수 있다.

이들 비용함수를 이용한 이단집락추출에서 이차추출단위의 최적배분은 라그랑즈 승수법을 이용하면 다음과 같다.

$$m_1^* = \frac{c_2 S_2^2 + \sqrt{c_2^2 S_2^4 + 8 c_1 S_2^2 c_2 S_1^2}}{2 c_2 S_1^2} .$$

$$m_2^* = \sqrt{\frac{c_1 S_2^2}{c_2 S_1^2}} .$$

$$m_3^* = \frac{-c_2 S_2^2 + \sqrt{c_2^2 S_2^4 + 8 c_2 S_1^2 c_1 S_2^2}}{4 c_2 S_1^2} .$$

이때 이차추출단위의 최적배분을 비교하여 보면, 다음과 같다.

$$m_1^* > m_2^* > m_3^* .$$

즉, 이 결과는 $Cost = c_1 n^p + c_2 n^p m$을 이용한 이차추출단위의 최적배분은 일차추출단위의 승수 p의 증가함수로 p가 커질수록 이차추출단위의 최적배분의 크기도 커지며, 반대로 일차추출단위의 최적배분은 역의 관계가 된다는 것을 의미한다.

참고문헌

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

[2] Hansen, M. H., Hurwitz, N. W, and Madow, W. G.(1973), *Sample Survey Methods and Theory, I : Methods and Applications,* New York : John Wiley.

[3] Jessen, R. J.(1978), *Statistical Survey Techniques*, New York: John Wiley & Sons.

[4] Snedecor, G. W. and Cochran, W. G.(1980), *Statistical Methods*(7th ed.), Ames. IA: Iowa State University Press.

[5] Sokal, R. R. and Rohlf, F. J.(1981), *Biometry*(2nd ed.), San Francisco: W. H. Freeman.

[6] Somerville, P. N.(1970), Optimum sample size for a problem in choosing the population with the largest mean, *Journal of the American Statistical Association*, V. 65, No. 330, 763-775.

[7] Yamane, T.(1967), *Elementary Sampling Theory*, Prentice-Hall, Inc., Englewood Cliffs, N.J.