Proceedings of Joint Conference of Korean Data And Information Science Society and The Korean Data Analysis Society, April 28 - 29, 2006, pp. 181-187

A comparison of alternative estimators in view of the Rao-Hartley-Cochran sampling scheme

홍기학[1] · 이기성[2] · 손창균[3]

요 약

대규모 표본조사와 관련해서 관심변수와 보조변수간의 약한 상관관계를 고려한 Amahia et al.(1989)의 대체추정방법을 Rao-Hartley-Cocharn추출방법에 적용해서 Rao추정량과 효율성을 비교하였다.

주요용어 : Rao-Hartley-Cochran 추출방법, 상관계수, 초모집단모형, 예감분산.

1. 서론

다단계 표본설계에서 1차추출단위(primary sampling unit : psu)를 관심변수 y와 매우 상관관계가 높은 보조변수 x에 비례하도록 추출함으로서 모집단 총합 또는 모집단 평균에 대한 비편향추정량의 효율성을 향상시킬 수 있다는 것은 널리 알려진 사실이다.

일반적으로 대규모 표본조사에서는 여러 개의 속성(관심변수)에 대한 모집단 총합이나 평균 등을 동시에 추정해야하는 경우가 많다. 이러한 경우에도 조사자는 보통 하나의 보조변수에 대한 정보에 근거해서 psu를 추출한다. 따라서 보조변수와 매우 높은 상관관계를 갖는 속성들의 모집단 총합이나 평균 등에 대한 추정량의 효율은 매우 높을 수 있지만, 그렇지 못한 속성들에 대한 추정량은 비편향임에도 불구하고 큰 분산값으로 인해 추정의 효율은 매우 떨어지는 문제점이 있다.

Rao(1966)는 위와 같은 문제점을 해결하기 위한 한 방법으로 편향추정량이지만 평균제곱오차가 기존의 비편향추정량의 분산보다 작고, 관심변수와 보조변수간의 관계가 전혀 없을 경우에도 편향의 크기가 기존의 비편향추정량의 표준오차에 비해 상대적으로 크지 않은 대체추정방법을 PPS추출에 적용하였다. Rao가 가정한 관심변수와 보조변수간의 무상관은 실제 조사모집단내에서는 거의 일어날 수 없는 것으로서 현실성에 문제가 있었다. 이에 대한 대안으로 Bansal과 Singh(1985)는 관심변수와 보조변수간의 약한 상관관계를 고려한 대체추정량을 제시하고, 그 효율성을 Rao 추정량과 비교하였다. 그들은 또한 자신들의 추정량을 Rao-Hartley-Cochran 추출방법에 적용해서 Rao 추정

[1]520-714 전남 나주시 대호동 252, 동신대학교 컴퓨터학과 교수. E-mail : khhong@dsu.ac.kr
[2]565-701 전북 완주군 삼례읍 후정리 490, 우석대학교 e-정보공학과 교수. E-mail : gisung@woosuk.ac.kr
[3]122-705 서울시 은평구 불광동 산 42-14, 한국보건사회연구원 책임연구원. E-mail : cksons85@naver.com

량과 비교하였다. Amahia et al.(1989)은 관심변수와 보조변수간의 상관관계를 고려한 대체추정방법을 제시하였다. Rao이후 많은 연구자들이 다양한 pps추출법에 적용한 대체추정방법을 발표하였다.

본 논문에서는 Amahia et al.의 방법을 Rao-Hartley-Cochran 추출방법에 적용해서 Rao추정량과 효율성을 비교분석 하고자 한다.

2. Rao-Hartley-Cochran 추출방법에의 적용

2.1 Rao-Hartley-Cochran 추출방법

Rao-Hartley-Cochran 추출방법은 비등확률(unequal probability)을 이용해서 psu를 뽑는 방법이다. 크기 n의 표본을 얻기 위하여 모집단을 n개의 랜덤그룹 $(N_1, N_2, \cdots, N_n)$으로 분할한 후 각 그룹으로부터 독립적으로 하나의 psu를 pps방법으로 뽑는다. psu i가 그룹 g에 속한다면, psu i의 추출확률은 $\frac{p_i}{P_i}$,$(P_i = \sum_g p_i)$가 된다. 모집단 총합 Y에 대한 추정량 및 분산은 다음과 같다.

$$\hat{Y}_{RHC} = \sum_{i=1}^{n} \frac{y_i}{p_i} P_i . \tag{2.1}$$

식(2.1)은 모집단 총합 Y에 대한 비편향추정량이다. $\hat{Y}_{RHC}$의 분산은 각 그룹의 크기가 같을 때 최소가 된다. 즉, $N_1 = N_2 = \cdots = N_n = \frac{N}{n}$ 일 때, 다음과 같은 최소분산을 갖는다.

$$V_{\min}(\hat{Y}_{RHC}) = \frac{N-n}{(N-1)n}\left(\sum_{i=1}^{N} \frac{y^2{}_i}{p_i} - (\sum_{i=1}^{N} y_i)^2\right). \tag{2.2}$$

2.2 Rao추정량

Rao-Hartley-Cochran 추출방법에 적용한 Rao(1966)의 대체추정량은 식(2.1)에서 y_i를 Ny_ip_i로 대체함으로서 다음과 같이 얻어진다.

$$\hat{Y}_R = N\sum_{i=1}^{n} y_i P_i . \tag{2.3}$$

$\hat{Y}_R$의 분산은 식(2.2)에서 y_i대신에 Ny_ip_i를 대입함으로서 다음과 같이 얻어진다.

$$V(\hat{Y}_R) = \frac{(N-n)N^2}{(N-1)n}\left(\sum_{i=1}^{N} y^2{}_i p_i - (\sum_{i=1}^{N} y_i p_i)^2\right). \tag{2.4}$$

2.3 대체추정량

Amahia et al.(1989)의 대체추정량을 Rao-Hartley-Cochran 추출방법에 적용한 대체추정량은 식(2.1)에서 y_i를 $\frac{y_i p_i}{p^w{}_i}$, ($p^w{}_i = (1-\rho)\frac{1}{N} + \rho p_i$)로 대체함으로서 다음과 같이 얻을 수 있다. 이 때, ρ는 관심변수 y_i와 추출확률 p_i(보조변수)간의 모집단 상관계수이다.

$$\hat{Y}_H = \sum_{i=1}^{n} \frac{y_i}{p^w{}_i} P_i . \tag{2.5}$$

$\hat{Y}_H$의 분산은 식(2.2)에서 y_i대신에 $\frac{y_i p_i}{p^w{}_i}$를 대입함으로서 다음과 같이 얻어진다.

$$V(\hat{Y}_H) = \frac{N-n}{(N-1)n}\left(\sum_{i=1}^{N} \frac{y_i^2 p_i}{p_i^{w2}} - (\sum_{i=1}^{N} \frac{y_i p_i}{p_i^w})^2 \right). \tag{2.6}$$

3. 효율성 비교

3.1 편향

Rao-Hartley-Cochran 추출방법에 적용한 두 추정량 $\hat{Y}_R$과 $\hat{Y}_H$는 모두 모집단 총합 Y에 대한 편향추정량이다. 이들 편향들은 각각 다음과 같이 표현된다.

$$\begin{aligned} B(\hat{Y}_R) &= E(\hat{Y}_R) - Y = N\sum_{i=1}^{N} y_i p_i - Y \\ &= \sum_{i=1}^{N} y_i (N p_i - 1), \end{aligned} \tag{3.1}$$

$$\begin{aligned} B(\hat{Y}_H) &= E(\hat{Y}_H) - Y = \sum_{i=1}^{N} y_i p_i / p^w{}_i - Y \\ &= \sum_{i=1}^{N} y_i (p_i / p^w{}_i - 1). \end{aligned} \tag{3.2}$$

식(3.1)과 (3.2)로부터

$$B(\hat{Y}_R) - B(\hat{Y}_H) = \sum_{i=1}^{N} y_i p_i (N - \frac{1}{p^w{}_i}) \geq 0 \tag{3.3}$$

이 되고, $p^w{}_i = (1-\rho)\frac{1}{N} + \rho p_i$이므로 관심변수 y_i와 추출확률 $p_i (= x_i / \sum_{i=1}^{N} x_i)$간의 상관계수 ρ가 $\rho \geq 0$이면 식(3.3)은 항상 0보다 크거나 같다.

3.2 분산

Rao-Hartley-Cochran 추출방법에 적용한 두 추정량의 효율성을 비교하기 위하여 다음과 같은 초모집단모형을 가정한다.

$$\begin{aligned} &y_i = \beta p_i + e_i, \\ &E(e_i \mid p_i) = 0, \\ &E(e_i e_j \mid p_i p_j) = 0, \quad i \neq j, \\ &E(e_i^2 \mid p_i) = a p_i^g, \quad a > 0, \ g \geq 0. \end{aligned} \tag{3.4}$$

식(3.4)의 모형 하에서 식(2.4)의 예감분산(anticipated variance : AV)은 다음과 같이 구해진다.

$$\begin{aligned} AV_R = \ & E_M[\, V(\, \hat{Y}_R)] = E_M\left[\frac{(N-n)N^2}{(N-1)n} \left(\sum_{i=1}^{N} y^2{}_i p_i - \left(\sum_{i=1}^{N} y_i p_i \right)^2 \right) \right] \\ = \ & \frac{(N-n)N^2}{(N-1)n} \left[a \sum_{i=1}^{N} p_i^{g+1}(1-p_i) + \beta^2 \left(\sum_{i=1}^{N} p_i^3 - \left(\sum_{i=1}^{N} p_i^2 \right)^2 \right) \right]. \end{aligned} \tag{3.5}$$

같은 방법으로 식(2.6)의 예감분산은 다음과 같이 구할 수 있다.

$$\begin{aligned} AV_H = \ & E_M[\, V(\, \hat{Y}_H)] = E_M\left[\frac{N-n}{(N-1)n} \left(\sum_{i=1}^{N} \frac{y^2{}_i p_i}{p_i^{w2}} - \left(\sum_{i=1}^{N} \frac{y_i p_i}{p^w{}_i} \right)^2 \right) \right] \\ = \ & \frac{N-n}{(N-1)n} \left[a \sum_{i=1}^{N} \frac{p_i^{g+1}}{p_i^{w2}} (1-p_i) + \beta^2 \left(\sum_{i=1}^{N} \frac{p_i^3}{p^w{}_i} - \left(\sum_{i=1}^{N} \frac{p_i^2}{p^w{}_i} \right)^2 \right) \right]. \end{aligned} \tag{3.6}$$

<정리 3.1> 식(3.4)와 같이 주어진 초모집단 하에서 다음 조건을 만족하면 $\hat{Y}_H$는 $\hat{Y}_R$ 보다 예감분산 측면에서 더 효율적이다.

$$\rho^2 < \frac{1}{(1+\delta)} .$$

위 식에서

$$\delta = \frac{\left[\sum_{i=1}^{N} p_i^3 (N^2 - 1/p_i^{w2}) - (N \sum_{i=1}^{N} p_i^2)^2 + (\sum_{i=1}^{N} p_i^2 / p_i^w)^2 \right] (\sum_{i=1}^{N} p_i^g)}{N \sigma_p^2 \cdot \sum_{i=1}^{N} p_i^{g+1}(1-p_i)(N^2 - 1/p_i^{w2})}$$

이다.

<증명> $AV_R - AV_H > 0$ 이 되는 조건을 다음과 같은 관계식들을 이용해서 구한다.

$$\sigma_p^2 = \frac{1}{N} \left[\sum_{i=1}^{N} p_i^2 - (\sum_{i=1}^{N} p_i)^2 / N \right],$$

$$\frac{a}{N}\sum_{i=1}^{N} p_i^g = \sigma_y^2(1-\rho^2),$$

$$\beta^2 = \rho^2(\sigma_y^2/\sigma_p^2) = \frac{\rho^2}{(1-\rho^2)} \cdot a(\sum_{i=1}^{N} p_i^g)/N\sigma_p^2.$$

<참고 1> <정리 3.1>에서 $\frac{a}{N}\sum_{i=1}^{N} p_i^g = \sigma_y^2(1-\rho^2)$의 가정을 하지 않을 경우

$$\begin{aligned}
&\frac{(N-1)n}{(N-n)}(AV_R - AV_H) \\
&= a\left(\sum_{i=1}^{N} p_i^{g+1}(1-p_i)(N^2 - \frac{1}{p_i^{w2}})\right) \\
&\quad + \beta^2\left[N^2\{\sum_{i=1}^{N} p_i^3 - (\sum_{i=1}^{N} p_i^2)^2\} - \{\sum_{i=1}^{N} \frac{p_i^3}{p_i^{w2}} - (\sum_{i=1}^{N} \frac{p_i^2}{p_i^w})^2\}\right] \\
&= aC + \beta^2 D.
\end{aligned} \tag{3.7}$$

으로 나타낼 수 있다. 위 식에서

$$C = \left(\sum_{i=1}^{N} p_i^{g+1}\ (1-p_i)(N^2 - \frac{1}{p_i^{w2}})\right)$$

$$\begin{aligned}
D &= \left[N^2\{\sum_{i=1}^{N} p_i^3 - (\sum_{i=1}^{N} p_i^2)^2\} - \{\sum_{i=1}^{N} \frac{p_i^3}{p_i^{w2}} - (\sum_{i=1}^{N} \frac{p_i^2}{p_i^w})^2\}\right] \\
&= V\left(p_i/\frac{1}{N}\right) - V(p_i/p_i^w)
\end{aligned}$$

이다. 그런데 C는 항상 0보다 크거나 같으므로, $\hat{Y}_H$의 효율성은 D값의 크기에 의존한다. 따라서 $D>0$ 이면, $\hat{Y}_H$이 $\hat{Y}_R$ 에 비하여 예감분산 측면에서 더 효율적이라는 것을 알 수 있다.

<참고 2> $\sum_{i=1}^{N} p_i = 1$이고, $p^w{}_i = (1-\rho)\frac{1}{N} + \rho p_i$, $(0<\rho<1)$에서

$$N\left|p_i - \sum_{i=1}^{N} p_i^2\right| - \left|\frac{p_i}{p_i^w} - \sum_{i=1}^{N} \frac{p_i^2}{p_i^w}\right| \geq 0$$

이면

$$D = V\left(p_i/\frac{1}{N}\right) - V(p_i/p_i^w) \geq 0$$

이 된다.

<증명>

$$V\left(p_i/\frac{1}{N}\right) = E\left(p_i/\frac{1}{N} - E(p_i/\frac{1}{N})\right)^2 = E(Np_i - \sum_{i=1}^{N} Np_i^2)^2$$

이고

$$V(p_i/p_i^w) = E\left(\frac{p_i}{p_i^w} - E(\frac{p_i}{p_i^w})\right)^2 = E\left(\frac{p_i}{p_i^w} - \sum_{i=1}^{N}\frac{p_i^2}{p_i^w}\right)^2$$

이므로

$$N\left|p_i - \sum_{i=1}^{N} p_i^2\right| - \left|\frac{p_i}{p_i^w} - \sum_{i=1}^{N}\frac{p_i^2}{p_i^w}\right| \geq 0$$

이면

$$D = V\left(p_i/\frac{1}{N}\right) - V(p_i/p_i^w) \geq 0$$

임을 보일 수 있다.

참고문헌

[1] Amahia, N., Chaubey, Y. P. and Rao, T. J. (1989). Efficiency of a New Estimator in PPS Sampling for Multiple Characteristics, *Journal of Statistical Planning and Inference,* Vol. 21, 75-84.

[2] Bansal, M. L., and Singh, R. (1985). An Alternative Estimator for Multiple Characteristics in PPS Sampling, *Journal of Statistical Planning and Inference,* Vol. 11, 313-320.

[3] Bansal, M. L. and Singh, R. (1990). An Alternative Estimator for Multiple Characteristics in Rao-Hartley-Cochran Scheme, *Communication Statistics -Theory & Methods-,* Vol. 19, 1777-1784.

[4] Lee, G. S., Hong, K. H. and Son, C. K. (2001). 중복시행 무관질문모형의 효율성, *Journal of the Korean Data Analysis Society,* Vol. 3, 297-304.

[5] Lee, G. S. and Hong, K. H. (2003). 3단계 집락추출법에 의한 양적속성의 무관질문모형, *Journal of the Korean Data Analysis Society,* Vol. 5, 85-99.

[6] Rao, J. N. K. (1966). Alternative Estimators in PPS Sampling for Multiple Characteristics, *Sankhya* A, Vol. 28, 47-60.

[7] Valliant, R., Dorfman, A. H. and Royal, R. M. (2000). *Finite Population Sampling and Inference -A Prediction Approach-,* Wiley, U.S.A.

On the relative efficiency of alternative estimators in Rao-Hartley-Cochran sampling scheme

Ki-Hak Hong[1], *Gi-Sung Lee*[2], *Chang-Kyoon Son*[3]

Abstract

In this paper we suggest a new alternative estimator for the characteristics that are pooly correlated with the selection probabilities by applying the Amahia et al.(1989)'s estimator to Rao-Hartley-Cochran sampling scheme and compare it with that of Rao(1966)'s under a super-population model.

Keywords : Rao-Hartley-Cochran sampling scheme, correlation coefficient, super-population model anticipated variance.

[1]Professor, Department of Computer Science, Dongshin University, 252 Daeho-Dong, Naju, Chonnam 520-714, Korea. E-mail : khhong@dsu.ac.kr

[2]Professor, Department of e-Information Engineering, Woosuk University, 490 Hujeong-ri, Wanju-gun, Jeonbuk 565-701, Korea. E-mail : gisung@woosuk.ac.kr

[3]Research fellowt, Korea Institute for Health and Social Affairs, San 42-14 Bulgwangdong Eunpyeonggu, Seoul 122-705, Korea. E-mail : ckson85@naver.com