# Minimum Variance Estimation for the Power Allocation in Stratified Sampling

손창균[1] 홍기학[2] 이기성[3]


## 요약

본 논문에서는 초 모집단 모형 하에서 HT 추정량의 분산의 하한에 관계된 층화추정량의 효율성에 대해 다루었다. 특별히 Dalenius-Hodges 층화와 표본배분방법 중 멱배분(power allocation)을 적용했을 때 최소분산 성질에 대해 살펴보았다.


주요 용어 : 초모집단 모형, Dalenius-Hodges 층화, 최소분산추정, 층화추정, HT 추정량.

## 1. 서론

일반적으로 층화 표본설계는 추정량의 정도를 높이고, 지역, 또는 연령과 같은 층화 변수를 사용함으로서 표본 설계가 용이하다는 장점을 가진다. 층화 추출설계에서 사용된 층의 수가 적절할 때, HT 추정량의 분산에 대한 최대 하한과 그와 연관된 추정량들의 효율성은 Bethel(1989)에 의해 논의된 바 있다. 이 과정에서 초모집단 모형을 적용하며, 추정량으로는 HT 추정량을 사용하고, 이용 가능한 보조정보로는 단변량 보조정보를 가정하였다. 그러나 모집단이 다양한 크기를 가진 부차 모집단들로 구성되어 있는 경우 기존의 표본 배분 방법을 적용하게 되면 부차 모집단의 특성들을 제대로 반영하지 못하는 문제점이 발생한다. 즉, 층화 표본설계에서 표본 배분을 위해 층별 총합과 전체 모집단 총합을 필요로 하며, 특별히 네이만 배분의 경우 전체 모집단과 큰 층에 대해 필요이상의 정도를 제공하지만, 작은 층의 경우에는 충분하지 않다.

이러한 문제점을 적절히 해결하기 위한 표본 배분 방법중의 하나가 멱 배분이며, 이 방법은 모집단 전체를 대상으로 하는 표본 배분 방법보다는 전체 모집단을 구성하고 있는 다양한 크기를 가진 부차 모집단에 대한 조사의 경우 표본 배분의 문제에 초점을 맞추고 있다.

따라서 본 논문에서는 다양한 크기로 구성된 부차모집단으로 구성된 모집단 조사에서 표본배분방법을 멱 배분으로 적용한 경우 층 경계에 대해 Dalenius-Hodges(D-H)(1959)가 제시한 방법을 적용했을 때 분산의 하한을 도출하며, 기존의 표본 배분 방법과의 효율성을 비교하고자 한다.

본 논문의 구성은 2절에서 초모집단 모형 하에서 일반적인 HT 추정량의 분산의 하한을 소개하고, 3절에서는 D-H 층화 방법에 멱 배분을 적용했을 때의 분산의 하한을 도출하여 여러 가지 배분방법과 본 논문에서 제안한 방법들간의 효율성을 비교하였고, 끝으로 4절에서는 결론과 추후 연구 과제들을 다루었다.


1) (520-714) 전남 나주시 대호동 동신대학교 컴퓨터학과 전임강사
2) (520-714) 전남 나주시 대호동 동신대학교 컴퓨터학과 부교수
3) (565-701) 전북 완주군 삼례읍 후정리 우석대학교 전산통계학과 부교수

## 2. 모형 가정과 최소 분산

추출설계 $D$를 이용하여 유한모집단 $U=\{(y_1,x_1),(y_2,x_2),\cdots,(y_N,x_N)\}$을 조사한다고 하면, $(y_k,x_k)$는 결합분포를 가진 확률변수 $(Y,X)$의 실현치이고, $(Y,X)$는 다음과 같은 모형을 만족한다고 하자.

$$Y=\alpha+\beta X+\gamma(X)\varepsilon \tag{2.1}$$

여기서 $\alpha$, $\beta$는 상수이고, $X$와 $\varepsilon$는 서로 독립이며, $E(\varepsilon)=0$, $V(\varepsilon)=\sigma^2$, $\gamma(X)>0$이다.

$y_k$의 모평균 $\bar{Y}=1/N\sum^N y_k$의 추정량으로 HT 추정량인 $\hat{\bar{Y}}=1/N\sum^n y_k/\pi_k$를 적용하자. 층화변수 $X$가 기지이므로 최적기준은 다음의 예측 분산을 최소로 한다.

$$\begin{aligned} v(\hat{\bar{Y}}-\bar{Y}) &= E_M[v_D(\hat{\bar{Y}}-\bar{Y})|X]+v_M[E_D(\hat{\bar{Y}}-\bar{Y})|X] \\ &= E_M[v_D(\hat{\bar{Y}}-\bar{Y})|X] \\ &= E_M[v_D(\hat{\bar{Y}})|X] \end{aligned} \tag{2.2}$$

$\hat{Y}=\alpha+\beta X$라 하면 예측 분산은 다음과 같다.

$$\begin{aligned} E_M[v_D(\hat{\bar{Y}})|X] &= 1/N^2 v_D(\hat{\bar{Y}})+\sigma_\varepsilon^2/N^2\sum^N\gamma(X_k)^2(1/\pi_k-1) \\ &\geq \sigma_\varepsilon^2/N^2\sum^N\gamma(X_k)^2(1/\pi_k-1) \\ &\geq \sigma_\varepsilon^2/N^2\Big[\sum^N(\gamma(X_k))^2-\sum^N\gamma(X_k)^2\Big]/n \\ &= \frac{\sigma_\varepsilon^2}{n}\Big[[E(\gamma(X))]^2-\frac{n}{N}E(\gamma(X)^2)\Big] \end{aligned} \tag{2.3}$$

식(2.3)은 분산의 하한으로 이용될 수 있으며, Godambe와 Joshi(1965)는 설계 비편향 예측 추정량의 기대 분산의 하한으로 이 값을 도출하였고, Cassel(1977)은 일반화 차의 추정량의 기대분산으로 도출한바 있다.

특별히 표본설계 $D$가 층화인 경우 $E_h(\cdot)$와 $S_h^2(\cdot)$를 각각 $h$층의 기대값과 분산이라 하자. 그러면 층화 표본 설계 하에서 $\hat{\bar{Y}}$의 예측 분산은 다음과 같다.

$$\begin{aligned} E_M[v_D(\hat{\bar{Y}})|X] &= E_M\Big[\sum_h^H W_h^2 S_h^2(Y)\Big(\frac{1}{n_h}-\frac{1}{N_h}\Big)|X\Big] \\ &= \frac{(1-f_h)}{n_h}\sum_h^H W_h^2[\beta^2 S_h^2(X)+\sigma_\varepsilon^2 E_h(\gamma(X)^2)] \end{aligned} \tag{2.4}$$

여기서 $f_h=n_h/N_h$이다.

위 식으로부터 $n_h$가 $N_h$에 비해 상대적으로 작다면 $f_h\approx 0$으로 무시될 수 있으며, 따라서 식(2.4)는 다음과 같다.

$$E_M[v_D(\hat{\bar{Y}})|X]\approx\frac{1}{n_h}\sum_h^H W_h^2[\beta^2 S_h^2(X)+\sigma_\varepsilon^2 E_h(\gamma(X)^2)] \tag{2.5}$$

## 3. 먹 배분 하에서의 분산 추정

층화 단순임의 표본을 추출한다고 할 때, $N_h$는 $h=1,2,\cdots,H$ 층의 모집단 크기이다. 층화 모집단으로부터 $\sum_h n_h = n$의 조건하에서 다음의 손실함수를 최소로 하는 $h$층의 표본수 $n_h$를 결정하고자 한다.

$$L = \sum_h (X_h^q \, cv(\hat{Y}_h))^2 \tag{3.1}$$

식(3.1)에서 $cv^2(\hat{Y}_h) = v(\hat{Y}_h)/Y_h^2$이며, $X_h$는 $h$에 대한 층화 변수이고, $q$는 $0 \le q \le 1$인 범위를 갖는 상수이다. 또한, $\bar{\hat{Y}}_h = \sum_i^{n_h} y_{hi}/n_h$로서 $\bar{Y}_h$의 추정량이다.

그러면 식(3.1)의 손실함수 $L$은 $n_h$가 다음과 같을 때 최소가 된다.

$$n_h = n \frac{S_h(Y) X_h^q / \bar{Y}_h}{\sum_h S_h(Y) X_h^q / \bar{Y}_h} \tag{3.2}$$

이 때 $v(\bar{\hat{Y}}_h) = (1/n_h - 1/N_h) S_h^2(Y)$ 이며, $S_h^2(Y) = 1/(N_h-1) \sum^{N_h} (y_{hi} - \bar{Y}_h)^2$, $\bar{Y}_h = Y_h/N_h$ 이다.

따라서 $q(0<q<1)$의 선택 값에 따라 특정한 표본 배분이 될 수 있으며, 특별히 $q=1$와 $X_h = Y_h$ 일 경우 Neyman 배분과 일치한다.

만일 층별로 $cv(\hat{Y}_h)$가 일정하다면, 식(3.2)의 먹 배분은 다음과 같이 축소된다.

$$n_h \propto (X_h)^q \tag{3.3}$$

이 때, $q$의 값은 일반적으로 1/2 또는 1/3이 적절한 것으로 알려져 있다.

식(2.5)에 식(3.2)를 대입하여 정리하면 다음과 같다.

$$\begin{aligned} E_M[v_{DP}(\bar{\hat{Y}})|X] &\approx \frac{1}{n_h} \sum_h^H W_h^2 [\beta^2 S_h^2(X) + \sigma_\varepsilon^2 E_h(\gamma(X)^2)] \\ &\approx \frac{1}{n} \left( \frac{\bar{Y}_h}{X_h^q} \right) \left[ \sum_h^H W_h \left( \frac{X_h^q}{\bar{Y}_h} \right)^{1/2} [\beta^2 S_h^2(X) + \sigma_\varepsilon^2 E_h(\gamma(X)^2)]^{1/2} \right]^2 \end{aligned} \tag{3.4}$$

**Remark 3.1]** 만일 $q=1$과 $X_h = Y_h$인 네이만 배분을 고려하면, 식(3.4)는 다음과 같이 축소되며, Bethel(1989)과 같다.

$$E_M[v_{DP}(\bar{\hat{Y}})|X] \approx \frac{1}{n} \left[ \sum_h^H W_h [\beta^2 S_h^2(X) + \sigma_\varepsilon^2 E_h(\gamma(X)^2)]^{1/2} \right]^2 \tag{3.5}$$

식(3.4)으로부터 분산의 하한을 결정하기 위해 층수 $H$가 증가함에 따라 식(3.4)의 인수들을 테일러 전개하여 정리하면 다음과 같은 식으로 유도할 수 있다.

$$\begin{aligned} &\sum^H W_h \left( \frac{X_h^q}{\bar{Y}_h} \right)^{1/2} [\beta^2 S_h^2(X) + \sigma_\varepsilon^2 E_h(\gamma(X)^2)]^{1/2} \\ &= O(1) \sum_{h=1}^H W_h S_h^2(X) \left( \frac{X_h^q}{\bar{Y}_h} \right)^{1/2} + \sum_{h=1}^H W_h \left[ \sigma_\varepsilon^2 E_h(\gamma(X)^2) \frac{X_h^q}{\bar{Y}_h} \right]^{1/2} \end{aligned} \tag{3.6}$$

그런데. 추정량의 분산이 층의 경계에 좌우되므로, 적절하게 층 경계가 결정된다면, 분산의 하한을 결정할 수 있을 것이다. 따라서 다음과 같이 D-H의 층화방법을 적용하기 위해 $a_h$와 $a_{h-1}$를 층경계의 상한과 하한이라 하면, D-H 층화방법은 다음을 만족하는 층을 생성하게 된다.

$$\int_{a_{h-1}}^{a_h} f_x(t)\,dt = c/H$$

여기서, $c = \int f_x(t)^{1/2}dt$이다.

$M_h = (a_h + a_{h-1})/2$라 하고, 층수 $H$이 크면, 다음과 같은 근사식이 성립한다.

$$S_h^2(X) \approx \frac{(a_h - a_{h-1})^2}{12}$$

$$c/H \approx f_x(M_h)^{1/2}(a_h - a_{h-1})$$

$$W_h = \int_{a_{h-1}}^{a_h} f_x(t)\,dt \approx f_x(M_h)(a_h - a_{h-1}) \approx \left(\frac{c}{H}\right)^2 (a_h - a_{h-1})$$

이러한 성질로부터 식(3.6)의 오른쪽 첫 항은 근사적으로 다음과 같다.

$$\sum_{h=1}^{H} W_h S_h^2(X)\left(\frac{X_h^q}{\overline{Y}_h}\right)^{1/2} \approx \sum_{h=1}^{H} \frac{(c/H)^2}{a_h - a_{h-1}} \frac{(a_h - a_{h-1})^2}{12}\left(\frac{X_h^q}{\overline{Y}_h}\right)^{1/2}$$

$$= O(H^{-2})\left(\frac{X_h^q}{\overline{Y}_h}\right)^{1/2}$$

이와 유사한 방법으로 식(3.6)의 오른쪽 두 번째 항은 다음과 같이 표현할 수 있다.

$$\sum_{h=1}^{H} W_h\left[\frac{X_h^q}{\overline{Y}_h} E_h(\gamma(X))^2\right]^{1/2} \approx \sum_{h=1}^{H} W_h E_h(\gamma(X))\left(\frac{X_h^q}{\overline{Y}_h}\right)^{1/2} = E(\gamma(X))\left(\frac{X_h^q}{\overline{Y}_h}\right)^{1/2}$$

결과적으로 식(3.4)는 근사적으로 다음과 같은 식으로 표현이 가능하다.

$$E_M[v_{DP}(\hat{\overline{Y}})|X] \approx \frac{1}{n}\left(\frac{\overline{Y}_h}{X_h^q}\right)\left[O(H^{-2})\left(\frac{X_h^q}{\overline{Y}_h}\right)^{1/2} + \sigma_\varepsilon E(\gamma(X))\left(\frac{X_h^q}{\overline{Y}_h}\right)^{1/2}\right]^2$$

$$= O(H^{-2}) + \frac{1}{n}\sigma_\varepsilon^2[E(\gamma(X))]^2 \qquad (3.7)$$

$$\to \frac{1}{n}\sigma_\varepsilon^2[E(\gamma(X))]^2$$

따라서 멱 배분의 경우 층의 수가 적절히 크면, 식(3.7)과 같이 분산의 하한을 도출할 수 있으며, 이 결과는 Bethel(1989)의 결과와 일치한다.

## 4. 결론

본 논문에서는 층화 표본 설계의 경우 D-H 층 경계 결정 방법과 멱배분을 적용했을 때, 분산의 하한을 구해 보고, 이를 다른 표본 배분 방법과 비교하였다. 특별히 멱 배분에서 $q = 1$과 $X_h = Y_h$인 경우 네이만 배분과 동일하며, 분산의 하한 또한 동일하게 유도되었다.

추가적으로 다양한 층 경계 결정 방법을 멱배분과 연계하여 분산의 하한을 도출해 보고, 이들간의 효율성에 대해서는 추후 연구과제로 남긴다.

## 참고문헌

[1] 손창균, 홍기학, 이기성(2000), "Minimum Variance Estimation in Stratified Sampling by an Extended Ekman Rule", 한국통계학회 호남제주지회 발표 논문.

[2] Bankier, M. D.(1988), "Power Allocations : Determining Sample Sizes for Subnational Area", *The American Statistician*, Vol, 42, No. 3, pp. 174-177.

[3] Bethel, J.(1989), "Minimum Variance Estimation in Stratified Sampling", *Journal of American Statistics Association*, Vol. 84, pp. 260-265.

[4] Cassel, C. M., Särndal C. E., and Wretman, J. H.(1977), *Foundations of Inference in Survey Sampling*, New York : John Weily.

[5] Dalenius, T.(1950), "The Problem of Optimum Stratification," *Skandinavisk Aktuarietidskrift*.

[6] Dalenius, T., and Hodges, J. L.(1959), "Minimum Variance of Stratification," *Journal of American Statistics Association*, Vol. 54, pp. 88-101.

[7] Godambe, V. P., and Joshi, V. M.(1965), "Adimissibility and Bayes Estimation in Sampling Finite Populations I", *Annals of Mathematical Statistics*, Vol. 36, pp. 1701-1722.

[8] Särndal, C. E.(1980), "On $\pi$-Inverse Weighting Versus Best Linear Unbiased Weighting in Probability Sampling", *Biometrika*, Vol. 67, pp. 639-650.

[9] Wirght, R. L.(1983a), "Finite Population Sampling With Multivariate Auxiliary Information," *Journal of the American Statistical Association*, Vol. 78, pp. 879-884.