確率抽出에 의한 層別샘플링의 經濟性에 關한 硏究
—A Study on economically optimal Determination of the Parameters of the Stratified Random Sampling—

黃 義 徹*
李 榮 植**

Abstract

In stratified random sampling a simple random sample must be taken in each stratum to reduce the maximum gain in precision given the minimum cost. The purpose of this paper is to deal with the pro-pertics of the estimates and variances and obtain the economic design of stratified random sampling through the optimum allocation of the sample sizes. In addition, the between stratum variation and the within stratum variation is stratifying the population are described.

Ⅰ. 序 論

母集團이 분명히 異質成分으로 구성되어 있다고 생각할 때는 이것들을 몇개의 로트로 層別하고 각 層으로부터 샘플을 取해야 된다. Lot를 여러개로 層別하여 서브로트(層)로 나누고 그 各層으로 부터 샘플을 取하는 方法을 層別샘플링 方法이라 한다. 이때 標本은 最小限의 時間과 經費로 주어진 正確度를 갖도록 抽出되어야 한다. 여기서 먼저 層이란 subpopulation이라 말할 수 있는데 母集團 內에서 同一한 성격을 갖는 단위들의 集合을 말한다. 또한 母集團을 各層으로 구분하는 것을 層別 또는 層化(stratification)라고 하며 이 層化된 sub population에서 그 層의 크기에 비례하여 單純任意標本을 抽出하는 方式이 層別抽出法이 될수 있다. 여기서 各測定單位는 抽出單位가 되고 갭이 없다고 가정할 때 抽出構造가 대상의 母集團이될 수 있다. 그리고 層化된 實驗計劃의 블럭化(blocking)와 유사한 것으로 母集團이 보다 효과적으로 層化될수록 標本의 精度를 올릴 수 있고 標本으로 부터 보다 유용한 정보를 얻을 수 있다. 따라서 層別抽出法은 抽出作業의 운영상 어떤 종류의 層別샘플링이나 單純랜덤 샘플링보다 편리하고 作業의 정확을 기하기가 쉬운경우가 있다. 母集團의 特定部分이 어떤 目的에서 보아 硏究領域이 그 연구영역에 관한 推定에 대하여 특정의 目標精度가 지정되어 있는 경우에는 그와같은 연구영역 자체를 層으로 하도록 설계함으로써 소기의 目標精度에 알맞는 샘플配分을 계획할 수 있다.

이상과 같은 점들을 감안하여 샘플을 위한 推定量과 分散, 標本크기의 割當方法, 各種配分法의 비고등으로 層別샘플링의 經濟性을 논하고자 한다.

Ⅱ 샘플을 위한 推定量과 分散

지금 크기 N의 모집단이 각각 크기 N_1, N_2, N_3, …, N_i으로 구성된 여러개의 층으로 나누어져 있을 때 各層의 平均과 分散을 각각 μ_i와 σ_i^2이라 하자. 또한 각 층으로부터 크기가 각각 n_1, n_2, n_3, …, n_h인 표본을 단순확률 추출하기로 하고 i번째 층에서 j번째 관측치를 X_{ij}로 표시하면, 각 층의 표본 평균 $\bar{X}_i$는 $E(\bar{X}_i)=\mu_i$이며 분산은

$$V_{ar}(\bar{X}_i)=\frac{\sigma_i^2}{n_i}\left(1-\frac{n_i}{N_i}\right) \quad \cdots\cdots (1)$$

* 漢陽大工大 産業工學科 敎授
**安養專門大學 工業經營科 副敎授
접수 1990년 4월 25일

	층(strata)			
	1	2	·····	h
모집단 : 크 기	N_1	N_2	·····	N_h
평 균	μ_1	μ_2	·····	μ_h
분 산	σ_1^2	σ_2^2	·····	σ_h^2
표본 : 크 기	n_1	n_2	·····	n_h
표본평균	$\bar{X}_1$	$\bar{X}_2$	·····	$\bar{X}_h$
표본분산	s_1^2	s_2	·····	s_h^2

$N=\sum_{i=1}^{h} N_i$, 母平均$=\frac{1}{N}\sum_{i=1}^{h} N_i\mu_i$

$\bar{X}_i=\frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij}$, $s_i^2\frac{1}{n_i-1}\sum_{j=1}^{n_i}(X_{ij}-\bar{X}_i)^2$

(그림 1) 모집단의 구성과 통계량

전체 모집단의 평균은 각층의 평균에 대한 가중평균(weighted average)으로서

$$\mu=\frac{N_1}{N}\mu_1+\frac{N_2}{N}\mu_2+\cdots\cdots+\frac{N_i}{N}\mu_i \quad \cdots\cdots (2)$$

가 되므로 μ에 대한 불편 추정량은

$$\bar{X}_{st}=\frac{N_1}{N}\bar{X}_1+\frac{N_2}{N}\bar{X}_2+\cdots\cdots+\frac{N_i}{N}\bar{X}_i \quad \cdots\cdots (3)$$

로 된다(여기서 첨자 st는 층화표본(stratified sample)에 대한 표본평균 이란 뜻으로 쓰인 것이다.).

각 층에서 추출된 표본은 서로 독립이므로 $\bar{X}_{st}$의 분산은

$$\sum_{i=1}^{k} V_{ar}\left(\frac{N_i}{N}\bar{X}_i\right)=\sum_{i=1}^{k}\frac{N_i^2}{N^2}V_{ar}(\bar{X}_i)=\sum_{i=1}^{k}\frac{N_i^2\,\sigma_i^2}{N^2\,n_i}\left(1-\frac{n_i}{N_i}\right)$$

$$=\sum_{i=1}^{k}\frac{N_i^2\,\sigma_i^2}{N^2\,n_i}(N_i-n_i)$$

로 주어진다. 또한 실제로 層化抽出法이 單純抽出法보다도 精密度가 높다는 것을 層化抽出에 의한 μ의 推定의 式(그림 2)에 의해 증명할 수 있다.

點推定量 : $\bar{X}_{st}=\frac{N_1}{N}\bar{X}_1+\frac{N_2}{N}X_2+\cdots\cdots+\frac{N_h}{N}\bar{X}_h=\frac{1}{N}\sum_{i=1}^{h} N_i\bar{X}_i$

$E(\bar{X}_{st})=\mu$

$Var(\bar{X}_{st})=\frac{1}{N^2}\left[N_1(N_1-n_1)\frac{\sigma_1^2}{n_1}+N_2(N_2-n_2)\frac{\sigma_2^2}{n_2}+\cdots\cdots+N_h(N_h-n_h)\frac{\sigma_h^2}{n_h}\right]$

$=\frac{1}{N^2}\sum_{i=1}^{h} N_i(N_i-n_i)\frac{\sigma_i^2}{n_i}$

95%近似信賴區間 : $\bar{X}_{st}\pm\frac{2}{N}\sqrt{\sum_{i=1}^{h} N_i(N_i-n_i)\frac{s_t^2}{n_i}}$

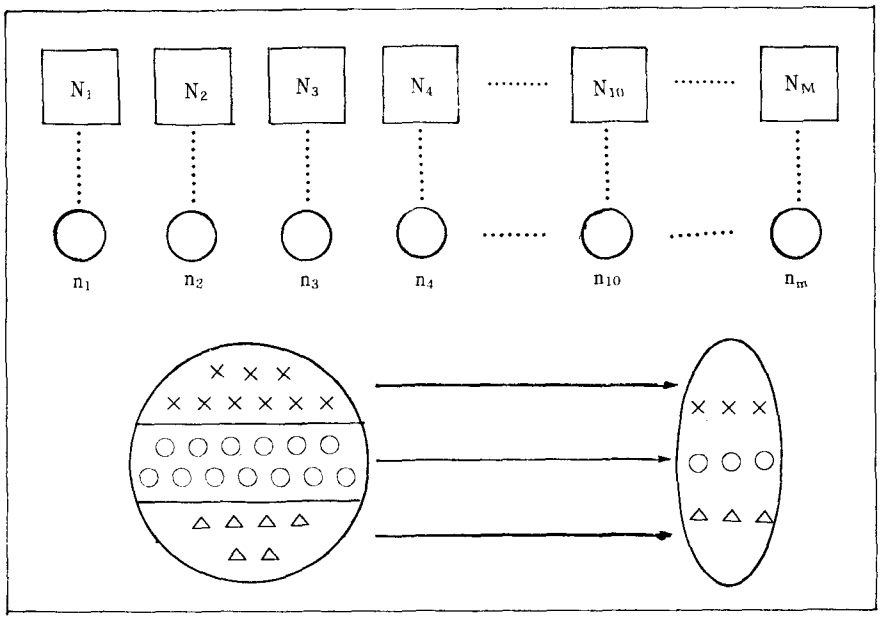
(그림 2) 層化抽出에 의한 μ의 推定

Ⅲ. 標本의 크기를 割當하는 方法

各層으로부터 샘플링을 하는 方法에는 各層의 크기에 比例하여 샘플링하는 比例샘플링, 各層의 크기와 標準偏差에 比例하여 샘플링하는 네이만샘플링 및 各層으로부터 샘플링하는 費用까지도 고려하는

데밍샘플링이 있다.

총표본의 크기 n은 費用, 時間 등의 제약에 의해서 결정되나 추정량의 분산을 작게하도록 n개를 各層에 할당하는 것이 바람직할 것이다.



(그림 3) 層別 샘플링

우선 각층의 크기에 비례해서 각층의 표본의 크기를 결정하는 比例割當을 생각할 수 있다. 즉 각층의 표본의 크기를

$$n_i = n\left(\frac{N_i}{N}\right),\ i=1,\ 2,\ 3,\ \cdots\cdots,\ h$$

로 정하여 추출하는 방법을 比例層化抽出(proportional stratified sampling)이라 한다. 이 方法에 의하면 各層이 그층의 크기에 따라 가중되는 장점이 있으며, 실제로 이 비례층화 추출법이 많이 사용된다.

다음은 추정량 $\bar{X}_{st}$의 분산을 최소로하는 할당법으로 즉 i번째 층의 분산이 σ_i^2이면 i번째 층의 표본의 크기를 $N_i\sigma_i$에 비례하도록 하는 할당방법을 最適割當(optimal allocation)이라 한다.

즉, $n_i = n\dfrac{N_i\sigma_i}{\sum_{j=1}^{h} N_j\sigma_j}$

가 된다.

1) 比例配分法(Proportional allocation)

各層의 크기에 比例해서 즉, 일정한 抽出 n/N에 의해 各層으로 부터 샘플링하는 方法이다. 즉 各層에서의 配分 n_i를

$$\frac{n_1}{N_1}=\frac{n_2}{N_2}\cdots\cdots=\frac{n_i}{N_i}\cdots\cdots=\frac{n}{N} \quad \cdots\cdots (4)$$

으로 정하는 방법을 比例配分法 또는 層別比例샘플링이라 한다.

또한, $W_i=\dfrac{N_i}{N}=\dfrac{n_i}{n}$

가 되어

$$\hat{\mu}_{st}=\frac{1}{n}\sum_{i=1}^{L} n_i\bar{x}_i=\frac{1}{n}\sum_{i=1}^{L}\sum_{j=1}^{n_i} x_{ij}=\bar{x} \quad \cdots\cdots (5)$$

가 되고, n의 不偏推定量 $\hat{\mu}_{st}$는 샘플의 單純平均 $\bar{x}$와 일치한다. 따라서 比例配分法은 不偏推定 $\hat{\mu}_{st}$의 계산은 매우 간단하게 된다. 이 配分法의 경우 $\hat{X}_{st}=N\bar{x}$ 가 되어 올바른 推定의 계산을 단순화하는 점은 比例配分法의 가장 중요한 이점이다.

$$V_p(\hat{\mu}_{st})=V(\bar{x})=\left(1-\frac{n}{N}\right)\frac{\sigma_w^2}{n}=\left(\frac{1}{n}-\frac{1}{N}\right)\sigma_w^2 \quad \cdots\cdots (6)$$

$$V_p(\bar{X}_{st})=N^2\left(1-\frac{n}{N}\right)\frac{\sigma_w^2}{n}=N^2\left(\frac{1}{n}-\frac{1}{N}\right)\sigma_w^2 \quad \cdots\cdots (7)$$

$\frac{n}{N}<0.1$로 有限修正이 무시되는 경우

$$V(\hat{\mu}_{st})=V(\bar{x})=\frac{\sigma_w^2}{n} \quad \cdots\cdots (8)$$

$V(\hat{X}_{st})=N^2\frac{\sigma_w^2}{n}$으로 해도 된다 $\cdots\cdots (9)$

單純랜덤 샘플링일 때의 μ의 不偏推定 $\bar{x}$의 分散式과의 비교는 분명히 σ_w^2와 σ^2 사이에는

$$\sigma^2=\sigma_w^2+\sigma_B^2 \quad \cdots\cdots (10)$$

여기서 $\sigma_B^2=\sum_{i=1}^{L} W_i(\mu_i-\mu)^2$ $\cdots\cdots (11)$

σ_B^2은 各層의 W_i를 무게로 한 加重層間分散인데, 이것을 간단히 層間分散(between stuatavariance)이라 한다. 또 σ^2을 全分散(total variance)이라 하고 σ_B^2의 값은 모든 μ_i가 일치하여 μ와 같은 경우에 한해서 0이 되지만 실제문제에서는 항상 陽의 값을 취한다고 보아도 된다. 일반적으로 서로 다른 두개의 샘플링方式의 相對精度(relative precision)는 샘플사이즈를 같게 했을 때의 兩者의 分散의 역비로 정의되는데 이것에 따르면 層別比例샘플링의 單純랜덤 샘플링에 대한 相對精度는

$$\frac{V_R}{V_P}=1+\frac{\sigma_B^2}{\sigma_W^2} \quad \cdots\cdots (12)$$

으로 표시된다. 여기서 V_R, V_P는 각각 샘플 사이즈를 모두 n으로 했을 때의 單純랜덤 샘플링, 층별비례샘플링에 의한 母平均의 不偏推定量 $\bar{x}$, $\bar{\mu}_{st}$의 分散을 나타낸다고 한다. 될수록 층내의 散布 σ_w^2이 작아지게 끔 層別設計를 하면 σ_B^2은 당연히 커져서 위식의 값은 커진다.

2) 네이만配分法(Neyman allocation)

전체로서의 샘플사이즈 n을 일정한 값으로 유지할 때 각층에서의 n_i 配分을 어떻게하면 推定精度를 가장높게, 推定量의 분산을 가장 작게 할 수 있는가를 알아보기 위해 推定量으로서

$$\hat{\mu}_{st}=\sum_{i=1}^{L}\frac{N_i}{N}\bar{x}_i=\sum_{i=1}^{L}W_i\bar{x}_i \quad \cdots\cdots (13)$$

의 $\hat{\mu}_{st}$를 채택하여 $V(\hat{\mu}_{st})$를 n이 일정하다는 조건하에서 최소로 하는 것을 생각해 본다. $\hat{\mu}_{st}$ 대신에 $\hat{X}_{st}$에 대하여 생각해 보면 $V(\hat{\mu}_{st})$는 일반적으로

$$V(\hat{\mu}_{st})=\sum_{i=1}^{L}W_i^2\frac{\sigma_i^2}{n_i}-\sum_{i=1}^{L}\frac{W_i^2}{N_i}\sigma_i^2=\sum_{i=1}^{L}W_i^2\frac{\sigma_i^2}{n_i}-\frac{\sigma_w^2}{N} \quad \cdots\cdots (14)$$

으로 표시되나 이식의 우변의 제2항은 n_i의 配分에는 관계없는 값이므로 $V(\hat{\mu}_{st})$의 식에서 有限修正을 무시한 式

$$V(\hat{\mu}_{st})=\sum_{i=1}^{L}W_i^2\frac{\sigma_i^2}{n_i} \quad \cdots\cdots (15)$$

을 n=일정이라는 조건하에서 최소로 하는 것을 생각하면 된다. (15)식의 우변항은 $n_i(i=1, 2, \cdots, L)$라는 L개의 변수의 함수이며 이것을 $\sum_{i=1}^{L}n_i=n=$일정 이하는 조건하에서 최소로 하는 순수한 수학문제가 된다. 이와 같은 일정한 等式의 條件下에서 多變量函數의 最小, 혹은 最大로 논할 때의 가장 일반적인 방법은 승수(multiplien)을 사용하는 方法으로 不定의 승수 λ를 도입하여

$$f(n_i)=\sum_{i=1}^{L}W^2{}_i\frac{\sigma_i^2}{n_i}+\lambda\left(\sum_{i=1}^{L}n_i-n\right) \quad \cdots\cdots (16)$$

라는 $n_i(1, 2, 3, \cdots, L)$의 함수를 생각하고 이것을 各 n_i로 편미분한 식을 0으로 놓은 L개의 방정식과 $\sum_{i=1}^{L} n = n$=일정을 합친 (L+1)개의 방정식에서 (L+1)개의 미지수 $n_i(i=1, 2, \cdots, L)$ 및 λ로 구하는 것이다.

식 (16)에서 $\frac{\partial f}{\partial n_i} = -\frac{W_i^2 \sigma_i^2}{n_i^2} + \lambda$ 가 되므로

$$-\frac{W_i^2 \sigma_i^2}{n_i^2} + \lambda = 0 \quad (i=1, 2, \cdots, L) \cdots\cdots (17)$$

(17)식과 $\sum_{i=1}^{L} n_i = n$을 연립방정식으로 해서 n_i λ에 대하여 풀면 식 (17)에서

$$\frac{n_i}{N_i} \propto \sigma_i \cdots\cdots (18)$$

가 얻어진다.

여기서 $\bar{\sigma}_w = \sum_{i=1}^{L} w_i \sigma_i$는 w_i를 무게로 하는 各層의 層內標準差異 σ_i의 加重平均이며 平均層內 標準偏差라 한다. 식 (18)과 $\sqrt{\lambda} = \frac{\bar{\sigma}_w}{n}$에 의해 $n_i = \frac{w_i \sigma_i}{\bar{\sigma}_w} n$가 되므로 이것을 식 (14)에 대입하면 위에 조건을 만족시키는 配分에 대응하는 $\bar{\mu}_{st}$의 분산은

$$V_N(\bar{\mu}_{st}) = \frac{\bar{\sigma}_w^2}{n} - \frac{\sigma_w^2}{N} \cdots\cdots (19)$$

으로 된다.

한편, 有限修正을 무시할 수 있을때에는 ($n/N<0.1$일 때)

$$V_N(\hat{\mu}_{st}) = \frac{\bar{\sigma}_w^2}{n} \cdots\cdots (20)$$

와 같이 된다.

n=일정 이라는 조건하에서 $\hat{\mu}_{st}$의 분산을 최소로하는 配分 즉 (18)식에서와 같이 各層에서 抽出率 n_i/N_i가 σ_i에 비례하고 $\hat{\mu}_{st}$의 분산이 (19)식 또는 (20)식으로 표시되는 配分法이 네이만配分法(Neyman allocation)이다.

3) 最適配分法(Optimun allocation)

하나의 샘플링單位를 조사하는 비용이 層에 따라서 다른 경우에 조사비용은 일정하다는 조건하에서 $V(\bar{\mu}_{st})$를 최소로 하는 配分法을 말한다.

지금 첫째 1층으로부터 하나의 샘플링단위를 조사하는 비용을 K_i라 하면 全調査費用 중의 變動部分(n_i의 配分에 영향되는 부분)은

$$K = \sum_{i=1}^{L} K_i n_i \cdots\cdots (21)$$

으로 표시된다.

最適配分은 K가 일정하다는 조건하에서 $\sum_{i=1}^{L} W_i^2 \frac{\sigma_i^2}{n_i}$을 최소로하는 配分이 된다.

$$\left(\sum_{i=1}^{n} a_i^2\right)\left(\sum_{i=1}^{n} b_i^2\right) \geq \left(\sum_{i=1}^{n} a_i b_i\right)^2 \cdots\cdots (22)$$

의 不等式을 사용하여 이 配分을 유도하기 위해 (22)식에서

$a_i = \sqrt{K_i n_i}$, $b_i = \frac{W_i a_i}{\sqrt{n_i}}$, $n=L$라 하면 (22)식은

$$\left(\sum_{i=1}^{L} K_i n_i\right)\left(\sum_{i=1}^{L} W_i^2 \frac{\sigma_i^2}{n_i}\right) \geq \left(\sum_{i=1}^{L} W_i \sigma_i \sqrt{K_i}\right)^2$$

즉 $K\left(\sum_{i=1}^{L} W_i^2 \frac{\sigma_i^2}{n_i}\right) \geq \left(\sum_{i=1}^{L} W_i \sigma_i \sqrt{K_i}\right)^2 \cdots\cdots (23)$

와 같이 된다. 여기서 등호가 성립하려면 $\frac{a_1}{b_1} = \frac{a_2}{b_2} = \cdots\cdots = \frac{a_L}{b_L}$이어야 한다.

따라서

$$\frac{n_i}{N_i} \propto \frac{\sigma_i}{\sqrt{K_i}} \quad \cdots\cdots (24)$$

가 만족하는 경우가 되어 이 비례관계가 最適配分을 주게된다. 이때에 이루어지는 $\sum_{i=1}^{L} W_i^2 \frac{\sigma_i^2}{n_i}$의 최소값은 (23)식에서

$$V_0(\hat{\mu}_{st}) = \frac{1}{k}\left(\sum_{i=1}^{L} W_i \sigma_i \sqrt{k_i}\right)^2 - \frac{\sigma_w^2}{N} \quad \cdots\cdots (25)$$

로 되고 有限修正을 무시할 수 있는 경우는

$$V_0(\hat{\mu}_{st}) = \frac{1}{K}\left(\sum_{i=1}^{L} W_i \sigma_i \sqrt{k_i}\right)^2 \quad \cdots\cdots (26)$$

가 되어 k=일정일 때 $V_N \leq V_0$가 된다.

Ⅳ. 各種配分法의 比較

層別샘플링에서 各層을 集計단위로 해서 집계하고 그 결과에 층마다의 무게를 부과한 加重平均을 사용하였다. 그러므로 비례배분법 이외의 다른 배분법을 사용하는 것은 이런 의미에서 集計와 計算, 作業量의 증가를 가져오게 한다. 네이만 배분법이나 最適配分法을 사용할 때에는 그것에 의해 달성되는 精度의 向上分($\hat{\mu}_{st}$의 分散 감소분)이 作業증가분을 충분히 보상할 수 있느냐가 검토되어야 한다.

먼저 比例配分法과 네이만配分法을 비교하기 위해 샘플사이즈 n이 같다고 할 경우 V_P와 V_N의 差를 구하면

$$V_P - V_N = \frac{1}{n}(\sigma_w^2 - \bar{\sigma}_w^2) = \frac{1}{n}\sum_{i=1}^{L} w_i(\sigma_i - \bar{\sigma}_w)^2 \quad \cdots\cdots (27)$$

이 등식의 우변의 $\sum_{i=1}^{L} W_i(\sigma_i - \bar{\sigma}_W)^2$은 $\bar{\sigma}_w = \sum_{i=1}^{L} W_i \sigma_i$식에 의해 W_i를 무게로 하는 σ_i의 가중 분산을 표시하고 있다. 따라서 이와같은 σ_i의 산포가 클수록 네이만配分은 비례 배분에 비해 높은 精度가 얻어지게 된다.

σ_i의 分布가 작으면 比例配分法의 집계 계산상의 편리함이라는 이점을 버리고 네이만配分法을 채용하는 것은 문제가 된다. 다음에 동일한 費用 k하에서 네이만 配分法과 最適配分法을비교하기 위해 V_N과 V_0와의 差로 취하면

$$V_N - V_0 = \frac{\bar{\sigma}_w^2}{k}\sum_{i=1}^{L} W_i'(\sqrt{k_i} - \sqrt{k})^2 \quad \cdots\cdots (28)$$

이 얻어진다. 여기서 $W_i' = \frac{W_i \sigma_i}{\bar{\sigma}_w}$ 따라서 $\sum_{i=1}^{L} W_i' = 1$ 또는 $\sqrt{k} = \sum_{i=1}^{L} W_i' \sqrt{k_i}$이다.

즉, $\sqrt{k}$는 W_i를 무게로 하는 $\sqrt{k_i}$의 加重平均이며 위식의 우변의 $\sum_{i=1}^{L} W_i'(\sqrt{k} - \sqrt{k})^2$은 W_i'를 무게로 하는 $\sqrt{k_i}$의 加重分散이다.

그러므로 위식으로 표시되는 $\sqrt{k_i}$의 分布가 클수록 일정 비용하에서 最適配分은 네이만配分에 비해서 높은 精度를 얻게된다.

1) 最適配分과 네이만 샘플과의 격차

네이만배분식이나 최적배분식은 다함께 各層의 층내표준편차 σ_i의 값을 예상하고 있으나 현실적으로 σ_i의 값은 계획단계에서는 정확히 알지 못한다. 다만 과거의 同種의 調査의 경험이나 각층내의 분포형에 대한 가정에 의해 σ_i의 近似値를 알 수 있을 뿐이다. 따라서 그와같은 σ_i의 근사치를 사용해서 네이만 또는 최적배분의 결과를 얻어낼 수 있다.

이 조사의 정도는 보다 구체적으로 말해 배분의 근사도에 비해서 결과의 근사도 즉 推定量의 분산의 증가도로서 i層에 대해 진정한 네이만 배분을 n_iN, 실제로 사용되는 근사네이만 배분을 n_iN'라 하면당연히 $\sum_{i=1}^{L} n_iN = \sum_{i=1}^{L} n_iN' = n$이다. 네이만 배분의 결과로서 얻어진 $\hat{\mu}_{st}$의 분산을 V_N'라 하면 (19)식에서

$$V_N = \frac{\bar{\sigma}_w^2}{n} - \frac{\sigma_w^2}{L} \quad \cdots\cdots (29)$$

와 같이 된다.

위의 近似네이만 배분의 결과를 얻어지는 $\hat{\mu}_{st}$의 분산을 V_N'라 하면 (13)식에서

$$V_N' = \sum_{i=1}^{L} W_i^2 \frac{\sigma_i^2}{n_i N'} - \frac{\sigma_w^2}{N} \quad \cdots\cdots (30)$$

이 되므로

$$V_N' - V_N = \sum_{i=1}^{L} W_i^2 \frac{\sigma_i^2}{n_i N'} - \frac{\sigma_w^2}{n} \quad \cdots\cdots (31)$$

이 된다. 진정한 네이만 배분 $n_i N$에 대해서 $W_i \sigma_i = \frac{n_i N}{n} \sigma_w$가 되는 것을 사용하고 또한 유한 수정을 무시하고 (31)식을 변형하면

$$\frac{V_N' - V_N}{V_N} = \frac{1}{n} \sum_{i=1}^{L} \frac{(n_i N' - n_i N)^2}{n_i N'} \quad \cdots\cdots (32)$$

가 얻어진다. 또 동일한 방식으로 제 i층에 대해 진정한 최적배분포 n_{i0}, 실제로 사용되는 근사최적배분을 n_{i0}'라 하면 당연히 $\sum_{i=1}^{L} K_i n_{i0} = \sum_{i=1}^{L} K_i n_{i0}' = K$이다.

진정한 최적배분의 결과로서 얻어지는 $\hat{\mu}_{st}$의 분산을 V_0, 위의 근사 최적 배분의 결과로서 얻어지는 $\hat{\mu}_{st}$의 분산을 V_0'라 하면

$$V_0' - V_0 = \sum_{i=1}^{L} W_i^2 - \frac{1}{K}\left(\sum_{i=1}^{L} W_i \sigma_i \sqrt{k_i}\right)^2 \quad \cdots\cdots (33)$$

으로 된다.

진정한 최적배분 n_{i0}에 대해서

$w_i \sigma_i = \frac{n_{i0}\sqrt{K_i}}{K} \sum_{i=1}^{L} W_i \sigma_i \sqrt{K_i}$가 되는것을 이용하고 또한 유한수정을 무시하고 (33)식을 변형하면

$$\frac{V_0' - V_0}{V_0} = \frac{1}{k} \sum_{i=1}^{L} \frac{K_i (n_{i0}' - n_{i0})^2}{n_{i0}'} \quad \cdots\cdots (34)$$

이 된다.

近似네이만 配分, 近似最適配分의 近似度를 측정하기 위해 각각 $\frac{n_i N' - n_i N}{n_i N'}$, $\frac{n_{i0}' - n_{i0}}{n_{i0}}$를 택하여 이것들의 절대치의 상한을 g라면

$$\left|\frac{n_i N' - n_i N}{n_i N'}\right| < g, \quad \left|\frac{n_{i0}' - n_{i0}}{n_{i0}'}\right| < g \quad \cdots\cdots (35)$$

이 된다.

이들 부등식을 (33), (34)식에 대입하면

$$\frac{V_N' - V_N}{V_N} < g^2, \quad \frac{V_0' - V_0}{V_0} < g^2 \quad \cdots\cdots (36)$$

와 같이 된다.

즉, σ_i의 정확한 값을 알지 못하기 때문에 그 근사치를 사용해서 근사적인 네이만배분이나 최적배분을 하여도 그 결과는 精度면에서 진정한 네이만배분이나 최적배분의 경우와 그다지 차이가 없다.

2) 2단계샘플링에서의 最適配分

2단계 샘플링에서의 최적배분은 조사비용 m과 $\bar{n}$와의 함수로서 주어진 경우에 그 費用函數를 일정값으로 유지하면서 $V(\hat{\mu}_t)$의 값이 최소가 되게끔 m, $\bar{n}$의 값을 정하는 것이다.

비용함수의 전형적인 식은 다음과 같다.

$$K = K_1 m + K_2 m \bar{n} \quad \cdots\cdots (37)$$

여기서 k_1은 관찰해야 할 要素의 갯수에는 관계없으며 하나의 1차단위를 샘플에 포함시키는데 필요한 費用이고 k_2는 한개의 要素를 관찰하는데 필요한 費用이다. 그러므로 문제는 (37)식에서 k를 일정하게 유지하면서 $V(\hat{\mu}_t)$의 값을 최소로 만드는 m, $\bar{n}$를 구하는 문제가 된다. (22)식에서와 마찬가지로

$$a_1 = \sqrt{k_1 m}, \quad a_2 = \sqrt{k_2 m \bar{n}}, \quad b_1 = \frac{\sigma_B}{\sqrt{m}}, \quad b_2 = \frac{\sigma_w}{\sqrt{m\bar{n}}}$$

라 놓으면

$$a_1^2+a_2^2=k_1m+k_2m\bar{n},\quad b_1^2+b_2^2=\frac{\sigma_B^2}{m}+\frac{\sigma_W^2}{m\bar{n}},\quad a_1b_1+a_2b_2=\sqrt{k_1}\,\sigma_B+\sqrt{k_2}\,\sigma_W$$

이므로

$$(k_1m+k_2m\bar{n})\left(\frac{\sigma_B^2}{m}+\frac{\sigma_B^2}{mn}\right)\geq(\sqrt{k_1}\,\sigma_B+\sqrt{k_2}\,\sigma_W)^2$$

즉 $k\left(\frac{\sigma_{B2}}{m}+\frac{\sigma_w^2}{m\bar{n}}\right)>(\sqrt{k_1}\,\sigma_B+\sqrt{k_2}\,\sigma_B)^2$

부호가 성립하는 것은 $\frac{a_1}{b_1}=\frac{a_2}{b_2}$인 경우로

$$\bar{n}=\sqrt{\frac{k_1}{k_2}}\,\frac{\sigma_W}{\sigma_B} \quad \cdots\cdots (38)$$

의 경우이고 $V(\hat{\mu}_{st})$의 최소값은

$$Vmin(\hat{\mu}_t)=\frac{1}{k}(\sqrt{k_1}\,\sigma_B+\sqrt{k_2}\,\sigma_B)^2 \quad \cdots\cdots (39)$$

으로 주어진다. (39)식을 (38)식에 대입하면 최적조건을 만족하는 m은

$$m=\frac{k}{k_1+\sqrt{k_1k_2}\,\frac{\sigma_W}{\sigma_B}}$$

로 주어진다.

분산 공식에서 유한수정을 무시할 수 없을 때에는 식 (37)의 k=일정의 조건하에서 費用이 최소가 되도록 m, $\bar{n}$를 정하면 된다.

$$\bar{n}=\frac{k_1}{k_2}\,\frac{1}{\frac{\sigma_B^2}{\sigma_W^2}\,\frac{1}{N}},\quad m=\frac{k}{k_1+\sqrt{k_1k_2}\cdot\frac{1}{\sqrt{\frac{\sigma_B^2}{\sigma_W^2}-\frac{1}{N}}}} \quad \cdots\cdots (40)$$

이때에 달성되는 최소분산은

$$Vmin(\hat{\mu}_t)=\frac{\sigma_W^2}{K}\left(\sqrt{k_1}\,\sqrt{\frac{\sigma_B^2}{\sigma_B^2}-\frac{1}{N}}+\sqrt{k_2}\right)^2-\frac{\sigma_B^2}{M} \quad \cdots\cdots (41)$$

이다.

3) 기타 샘플링의 精度의 比較

① 層別과 比例 샘플링의 비교

$$\alpha=\frac{V_S(\bar{x})}{V_R(\bar{x})}=\frac{\sigma_W^2/m\bar{n}}{\sigma^2/n}=\frac{\sigma_W^2}{\sigma^2} \quad \cdots\cdots (42)$$

그런데 $\sigma^2=\sigma_B^2+\sigma_W^2$이므로 층별을 매우 서투르게 하여 $\sigma_b^2=0$일때 비로서 $\alpha=1$로 되고 $\sigma_b^2>0$이면 $\alpha<1$이 된다.

일반적으로 層別·比例 샘플링은 추정의 정밀도가 좋고 또 샘플링의 조작도 쉬우므로 권장할만한 샘플링방법이 된다.

② 聚落샘플링과의 비교

$$\alpha=\frac{V_C(\bar{x})}{V_R(\bar{x})}=\frac{\sigma_{B2}/m}{\sigma^2/n}=\frac{\bar{N}\sigma_b^2}{\sigma^2} \quad \cdots\cdots (43)$$

따라서 $\sigma_b^2>\frac{\sigma^2}{\bar{N}}$이면 $\alpha>1$

$\sigma_b^2\fallingdotseq\frac{\sigma^2}{N}$이면 $\alpha=1$

$\sigma_b^2<\frac{\sigma^2}{N}$이면 $\alpha<1$

즉 聚落을 잘 만들어서 σ_b가 상당히 작아질수만 있다면 취락샘플링 쪽이 추정의 정밀도는 좋으나 취락을 만드는 것이 서툴러서 σ_b가 크게되면 推定의 精密度는 랜덤샘플링보다 못하게 된다. 그러나 샘플링의 수고나 費用을 고려하면 취락샘플링이 有利할 때도 있다.

V. 結 論

層別 샘플링의 취급에 있어서 지금까지 各層의 크기 N_i, 혹은 상대적 크기 $W_i=\frac{N_i}{N}$의 값을 정확히 알고 있다고 보고 이론을 전개하여 왔다. 그러나 실제 문제에 있어서는 N_i, W_i의 값은 정확하게는 알지 못하고 그 近似値를 알고있는데 불과할 때가 많다. 이와같이 층의 크기의 값이 오차를 포함하고 있을때 지금까지 말한 결과가 어떤 영향을 미치는가는 W_i의 근사치로서 사용되는 값을 W_i'라 하면 당연히 $\sum_{i=1}^{L} W_i'=1$이 된다. 이때 μ의 推定量으로서

$$\hat{\mu}_{st}=\sum_{i=1}^{L}\frac{N_i}{N}\bar{X}_i=\sum_{i=1}^{L}W_i\bar{x}_i$$

의 식에서 $\sum_{i=1}^{L}W_i\bar{x}_i$ 대신에 다음식을 사용할 수 있다.

$$\tilde{\mu}_{st}=\sum_{i=1}^{L}W_i'\bar{x}_i \quad \cdots\cdots (44)$$

이 값의 기대치를 취하면

$$E(\tilde{\mu}_{st})=\sum_{i=1}^{L}W_i'\mu_i \quad \cdots\cdots (45)$$

그런데 $\mu=\sum_{i=1}^{L}W_i\mu_i$이므로 $\tilde{\mu}_{st}$는 μ의 추정량으로서 偏倚를 갖고 있다.

이 偏倚 $B(\tilde{\mu}_{st})$는 다음 식으로 표시된다.

$$B(\tilde{\mu}_{st})=\sum_{i=1}^{L}(W_i'-W_i)\mu_i \quad \cdots\cdots (46)$$

이 편기는 위 식에서 분명하듯 샘플사이즈 n의 크기와는 관계없는 定誤差라는 점이 매우 중요하다. 즉 n을 아무리 크게 하여도 이 편기를 감소시킬 수는 없다. 다음에 $\tilde{\mu}_{st}$의 분산은

$$V(\tilde{\mu}_{st})=\sum_{i=1}^{L}W_i^2\frac{N_i-n_i}{N_i-1}\frac{\sigma_i^2}{n_i} \quad \cdots\cdots (47)$$

그런데 $\tilde{\mu}_{st}$는 偏倚가 있는 推定量이므로 그 誤差의 정도는 分散과 偏倚를 綜合한 平均제곱오차(MSE)로 측정되어야 한다. MSE는 分散과 偏倚의 제곱과의 합으로 표시되므로

$$MSE(\tilde{\mu}_{st})=\sum_{i=1}^{L}W_i^2\frac{N_i-n_i}{N_i-1}\frac{\sigma_i^2}{n_i}+\{\sum_{i=1}^{L}(W_i'-W_i)\mu_i\}^2 \quad \cdots\cdots (48)$$

또 식 $V(\tilde{\mu}_{st})=\sum_{i=1}^{L}W_i'^2\left(\frac{1}{n_i}-\frac{1}{N_i}\right)s_i^2$ ····· (49)

은 분명히 $V(\tilde{\mu}_{st})$의 不偏推定値이므로 $MSE(\tilde{\mu}_{st})$의 推定으로서는 (48)식의 우변의 제2항분 만큼 과소평가의 편기를 갖는다. 단순랜덤샘플의경우 母平均 μ의 推定量 $\bar{x}$의 분산은 $V(\bar{x})=\frac{N-n}{N-1}\cdot\frac{\sigma^2}{n}$으로 표시되므로 이것은 n을 크게함에 따라 차차 작아지고 n을 충분히 크게 하면 분산값은 얼마든지 0에 접근시킬 수 있다.

그러나 층별샘플링에서 W_i대신 W_i'로 사용되는 推定量 MSE는 (48)식에서 분명하듯 $\{\sum_{i=1}^{L}(W_i'-W_i)\mu_i\}^2$보다 작아질 수는 없다.

그러므로 n을 충분히 크게하면 單純랜덤샘플링이 層別샘플링보다 精度의 점에서 좋게된다. 일반적으로 W_i의 誤差때문에 발생하는 精度의 손실은 層別의 효과가 심한 경우일수록 커진다는 결론이다. 母集團을 同質的인 層으로 나눌 수 있을 때 層化抽出法을 사용하면 推定量의 分散을 줄일 수 있는 長點이 있다.

層을 나누는 기본 原理는 層間의 變異性(variability)을 크게하고 層內部에서의 變異性은 작게하는 것이다. 따라서 各層에서는 상대적으로 작은 標本을 갖고도 정확한 推定을 할 수 있어야 하며 標本은 最小限의 時間과 經費로 주어진 正確度를 갖도록 抽出되어야 한다.

參 考 文 獻

1. 金宇哲, 朴聖炫, "現代統計學", 英志文化社, 1981, pp. 361~368.
2. 金成寅, "샘플링檢査" 博英社, 1986, pp. 280~304.
3. 黃義徹, "最新品質管理", 博英社, 1985, pp. 280~304.
4. 品質管理便覽集委員會, "品質管理便覽", 日本規格協會, 1962, pp. 254~275.
5. Taro Yamane, "Elementary Sampling Theory" Prentice-Hall, INC., Englewood Cliffs, N. J. pp. 102~157.
6. William G. Cochran, "Samgling Techniques", third edition John Wiley & Sons Inc., pp. 89~149.
7. Douglas C. Montgomery, "Statistical Quality Control", John Wiley & Sons Inc., 1985, pp. 171~220.
8. Duncan, A. T., "Quality Codntrol and Industrial Statistics", 4th ed, Homewood Ill Richard Irwin Inc., 1974.
9. Eugene L. Grant, "Statistical Quality Control" McGraw-Hill Book Company, 1988.
10. Juran, J. M "Quality Control Handbook" McGraw-Hill New York 3th ed, 1974, 25A.
11. W. K. Chiu, "Minimum Cost Control Schemes Using np Charts" Int, J. Prod, Res, Vol. 13, No. 4, pp. 341~349.
12. ———, "Economic Design of Attribute Control Charts" Technometrics, Vol. 17, No. 1, pp. 81~87.
13. _____, "A Sensivity Study of Minimum Cost np-charts" Int, J, Prod, Res Vol. 15, No. 3, pp. 237~242.