層化任意抽出에서의 最適標本設計

金 益 贊* • 金 鎭 求**

Optimum Design in Stratified Sampling

Kim Ik－chan • Kim Jin－gu

Abstract

In this thesis, the method that minimizes the variance of mean differences between estimated strata should he studied, in the case of strata L=2, L)2. In addition to this, the optimum design method which, minimizes the variance in case of stated cost, should be introduced, This optimum design method is used for analytical comparison, in a 2×2 contingency table, which decides two factors into each of two classes.

1. 序 論

標本論에서 標本의 精度와 관련된 표본의 크기를 決定하는 문제는 應用統計 分野에 있어서 가장 基本的인 主題라 할 것이다.

L개의 層으로 分類된 母集團에서 각 層마다 最適의 標本크기를 決定하는 문제는

* 제주대학교 사범대학 수학교육과
** 教育大學院 卒業生

L개의 層으로 分類된 母集團에서 각 層마다 最適의 標本크기를 決定하는 문제는 오래전부터 硏究되어 왔다.

특히 標本을 抽出하기 위한 費用函數가 주어졌을때 그 精度를 극대화 하기 위하여 抽出되어야 하는 標本의 크기, 또는 정해진 標本의 精度에 대하여 그 費用을 最小化 하는 標本의 크기를 決定하는 最適割當의 문제는 Stuart(1954), Rao(1973)등 많은 硏究가 있었다.

본 硏究에서는 層化標本抽出에 있어서 費用函數와 관련된 표본크기를 決定하는 方法으로서 母集團 分布 또는 母平均이 미리 알려졌다고 假定된 서로 다른 두 層간의 比較分析 方法을 模索하였다. 즉 2개의 층에서 推定된 층의 平均 사이의 差의 分散을 最小化하기 위한 標本크기를 決定하는 것이다.

이 平均差의 分散을 最小化하는 方法이 L)2 개의 層으로 一般化된다.

한편, 두개의 要因이 다시 두개의 部類로 分類되는 경우의 最適配分을 생각하였다. 즉 두 要因이 각 部類들 사이에 동일한 精度가 요청된 경우에, 주어진 費用을 最小化 하는 標本크기를 決定하는 方法을 2×2 分割表에 의하여 糾明하여 보았다.

크기 N인 母集團의 母平均 $\bar{Y}=\frac{1}{N}\sum_{i=1}^{N}yi$, 標本平均 $\bar{y}=\frac{1}{n}\sum_{i=1}^{n}yi$ 로 母分散을 S^2, 標本分散을 s^2으로, N_h를 h층의 母集團 單位들의 수, n_h를 h층 標本單位數를 나타내는 등 대문자는 母集團 母數를, 小文字는 標本單位로 表記하였다.

2. 一元分類에서의 最適標本設計

서로 다른 層들이 一元分類 형태를 취하고 있을때 層간의 平均比較를 위한 最適設計方案에 대하여 생각하여 보자. 단, 母集團 각층의 크기는 미리 알려져 있다고 假定한다.

만일 2개의 層만 있을 때는 推定된 層平均들 간의 平均差 $(\bar{y}_1-\bar{y}_2)$의 分散을 最小化하는 標本크기 n_1, n_2를 구할 수 있다.

$$V(\bar{y}_1-\bar{y}_2)=V(\bar{y}_1)+(\bar{y}_2)=\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2} \quad \cdots\cdots (2.1)$$

이고

費用函數를 $C=Co+C_1n_1+C_2n_2$ $\cdots\cdots (2.2)$

로 두면,

$$VC' = \left(\frac{S_1^2}{n1} + \frac{S_2^2}{n2}\right)(C_1n_1 + C_2n_2) \quad (\text{단, } C' = C - Co) \quad \cdots\cdots (2.3)$$

한편, Cauchy－Schwarz不等式 $(\sum a_h^2)(\sum b_h^2) \geq (\sum a_h b_h)^2$에 의하여

$$VC' = (\sum a_h^2)(\sum b_h^2) \geq (\sum S_h\sqrt{C_h})^2$$

$$\Rightarrow \frac{b_h}{a_h} = \frac{\sqrt{C_h n_h}}{S_h/\sqrt{n_h}} = \frac{n_h\sqrt{C_h}}{S_h} \equiv k$$

$$\therefore \quad n_h = \frac{kS_h}{\sqrt{C_h}} \cdots\cdots (2.4)$$

그리고 $n = \sum n_h = \frac{kS_1}{\sqrt{C_1}} + \frac{kS_2}{\sqrt{C_2}}$이다.

$$\frac{n_h}{n} = \frac{kS_h/\sqrt{C_h}}{kS_1/\sqrt{C_1} + kS_2/\sqrt{C_2}} = \frac{S_h/\sqrt{C_h}}{\sum(S_h/\sqrt{C_h})}$$

$$n_h = \frac{S_h/\sqrt{C_h}}{\sum(S_h/\sqrt{C_h})}n \quad \cdots\cdots (2.5)$$

$$\therefore \quad n_1 = \frac{nS_1/\sqrt{C_1}}{S_1/\sqrt{C_1} + S_2/\sqrt{C_2}}, \qquad n_2 = \frac{nS_2/\sqrt{C_2}}{S_1/\sqrt{C_1} + S_2/\sqrt{C_2}} \cdots\cdots (2.6)$$

일때 平均差의 分散은 最小가 된다.

한편, L개층이 L> 2인 경우의 最適設計는 서로 다른 층간의 平均 比較를 위하여 要求되는 精度의 量에 따른다.

예를들어 $V(\bar{y}_h - \bar{y}_h) \leq V_{hi}$가 되는 條件을 갖는 $\frac{L(L-1)}{2}$ 개의 集合에 의거해서 費用을 最小化하도록 할 수 있다. 여기서 V_{hi}의 값들은 층 h와 i들간의 충분한 比較를 위하여 要求되는 精度에 의해 選擇된다.

S_h와 C_h가 크게 다르지 않을때 $\frac{L(L-1)}{2}$ 개 쌍들의 層간의 差의 平均 分散을 最小化하기 위하여 다음과 같이 생각하였다.

$$\bar{V} = \frac{1}{LC_2}\{V(\bar{y}_1 - \bar{y}_2) + V(\bar{y}_1 - \bar{y}_3) + \cdots + V(\bar{y}_1 - \bar{y}_L)$$
$$+ V(\bar{y}_2 - \bar{y}_3) + V(\bar{y}_2 - \bar{y}_4) + \cdots + V(\bar{y}_2 - \bar{y}_L)$$
$$+ \cdots + V(\bar{y}_{L-1} - \bar{y}_L)\}$$

$$= \frac{1}{L(L-1)/2}[(L-1)\{V(\bar{y}_1) + V(\bar{y}_2) + \cdots + V(\bar{y}_L)\}]$$

$$= \frac{2}{L}\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2} + \cdots + \frac{S_L^2}{n_L}\right)$$

$$= \frac{2}{L}\sum_{h=1}^{L}\frac{S_h^2}{n_h} \quad \cdots\cdots (2.7)$$

이를 最小化 하기 위하여 費用

$$C = C_0 + \sum C_h n_h \quad \cdots\cdots (2.8)$$

가 정했을때 L=2개의 경우와 마찬가지로 $n_h \propto S_h/\sqrt{C_h}$ 한다는 것을適用하면,

$$n_h = \frac{S_h/\sqrt{C_h}}{\sum_{h=1}^{L}(S_h/\sqrt{C_h})}n \quad \cdots\cdots (2.9)$$

일때 $\bar{V}$는 最小가 된다. 또 式(2.9)를 式(2.8)에 代入함으로서

$$n = \frac{(C-C_0)\sum(S_h/\sqrt{C_h})}{\sum(S_h\sqrt{C_h})} \quad \cdots\cdots (2.10)$$

임을 알 수 있고, 각 층의 표본크기는

$$n_h = \frac{S_h}{\sqrt{C_h}}\frac{C-C_0}{\sum S_h\sqrt{C_h}} \cdots (2.11)$$

이다. 한편 式 (2.11)과 式 (2.7)에 의해 최소분산

$$\bar{V}(\min) = \frac{2}{L}\sum_{h=1}^{L}\frac{S_h^2}{(S_h/\sqrt{C_h})\{(C-C_0)/\sum S_h\sqrt{C_h}\}}$$

$$= \frac{2}{L}\frac{(\sum S_h\sqrt{C_h})^2}{C-C_0} \cdots (2.12)$$

이다.

3. 두가지 要因의 分析調査를 위한 最適設計

대부분의 標本調査에 있어서 根本目的은 有限母集團내의 몇가지 분야들간의 比較이다.

특히 α 와 τ 두가지 要因이 각각 두가지 部類로 나누어져 2x2分割表로 표시되는 경우를 생각한다. 이러한 分割表에서의 두要因 比較 分析調査를 위하여, 주어진 費用에 대해서 分散을 最小化하는 最適設計方案을 본절에서 해결하려고 한다.

이제 α의 i번째 部類의 요소와 τ의 j번째 部類의요소를 (i,j)로 표시하면 각 要因에 대한 두가지 部類는

$$\left.\begin{aligned} D\alpha &= W_{.1}(\bar{Y}_{11}-\bar{Y}_{21}) + W_{.2}(\bar{Y}_{12}-\bar{Y}_{22}) \\ D\tau &= W_{1.}(\bar{Y}_{11}-\bar{Y}_{12}) + W_{2.}(\bar{Y}_{21}-\bar{Y}_{22}) \end{aligned}\right\} \cdots\cdots (3.1)$$

에 의해서 比較될 수 있다.

여기서 N_{ij} : (i,j) 요소의 母集團의 크기

$W_{ij} = \frac{N_{ij}}{N}$, $\quad W_{i.} = \sum_j W_{ij}$, $\quad W_{.j} = \sum_i W_{ij}$

$\bar{Y}_{ij}$: (i,j) 요소내의 母平均을 표시한다.

따라서 式(3.1)의 不偏推定量이

$$\left.\begin{aligned} \hat{D}\alpha &= w_{.1}(\bar{y}_{11}-\bar{y}_{21}) + w_{.2}(\bar{y}_{21}-\bar{y}_{22}) \\ \hat{D}\tau &= w_{1.}(\bar{y}_{11}-\bar{y}_{12}) + w_{2.}(\bar{y}_{21}-\bar{y}_{22}) \end{aligned}\right\} \cdots\cdots (3.2)$$

임을 쉽게 알 수 있다.

이제 $\hat{D}\alpha$와 $\hat{D}\tau$에 대하여 同一한 精度가 要求되는 것으로 한다면 그 目的函數를 두 推定量의 分散의 합의 平均으로서

$$\bar{V} = \frac{1}{2}\{V(\hat{D}\alpha) + V(\hat{D}\tau)\} \cdots\cdots (3.3)$$

으로 定義할 수 있다.

여기서 fpc 를 생략함으로서

$$V(\hat{D}\alpha + V(\hat{D}\tau) = w_{.1}^2\left(\frac{s_{11}^2}{n_{11}} + \frac{s_{21}^2}{n_{21}}\right) + w_{.2}^2\left(\frac{s_{12}^2}{n_{12}} + \frac{s_{22}^2}{n_{22}}\right)$$

$$+ w_{1.}^2\left(\frac{s_{11}^2}{n_{11}} + \frac{s_{12}^2}{n_{12}}\right) + w_{2.}^2\left(\frac{s_{21}^2}{n_{21}} + \frac{s_{22}^2}{n_{22}}\right)$$

$$= \sum_i \sum_j (w.j^2 + wi.^2) \frac{s_{ij}^2}{n_{ij}}$$

따라서,

$$\overline{V} = \sum_i \sum_j \frac{g_{ij}^2}{n_{ij}} \quad \cdots\cdots (3.4)$$

(단, $2g_{ij} = (w._j^2 + w_{i.}^2)s_{ij}^2$)

한편, 費用函數를

$$C = C0 + \sum_i \sum_j C_{ij}\, n_{ij} \quad \cdots\cdots (3.5)$$

로 定義하면 $C' = C - C0 = \sum_i \sum_j C_{ij} n_{ij}$로 두었을 때 最適設計配分 方式이 適用될 수 있다.

즉, $\overline{V}C' = (\Sigma\Sigma \frac{g_{ij}^2}{n_{ij}})\ (\Sigma\Sigma C_{ij}\, n_{ij})$

$= \{\Sigma\Sigma(g_{ij}/\sqrt{n_{ij}})^2\}\ \{\Sigma\Sigma(\sqrt{C_{ij}\, n_{ij}})^2\}$

$\geq (\Sigma\Sigma g_{ij}\sqrt{C_{ij}})^2$

$$\frac{\sqrt{C_{ij}\, n_{ij}}}{g_{ij}/\sqrt{n_{ij}}} = k \Rightarrow n_{ij} = \frac{g_{ij}}{\sqrt{C_{ij}}} k$$

그리고

$$n = \Sigma\Sigma n_{ij} = \Sigma\Sigma(k g_{ij}/\sqrt{C_{ij}})$$

따라서

$$\frac{n_{ij}}{n} = \frac{k\, g_{ij}/\sqrt{C_{ij}}}{\Sigma\Sigma(k\, g_{ij}/\sqrt{C_{ij}})}$$

$$\Rightarrow n_{ij} = \frac{g_{ij}/\sqrt{C_{ij}}}{\Sigma\Sigma\ (g_{ij}/\sqrt{C_{ij}})} n \quad \cdots\cdots (3.6)$$

式(3.6)을 式(3.5)에 大入하면,

$$C - C0 = \Sigma\Sigma C_{ij} \frac{g_{ij}/\sqrt{C_{ij}}}{\Sigma\Sigma(g_{ij}/\sqrt{C_{ij}})} n \quad \text{이 되어}$$

$$n = \frac{(C - C0)\Sigma\Sigma(g_{ij}/\sqrt{C_{ij}})}{\Sigma\Sigma g_{ij}\sqrt{C_{ij}}} \quad \cdots\cdots (3.7)$$

이 된다.

또 式(3.7)를 式(3.6)에 大入하면,

$$n_{ij} = \frac{g_{ij}}{\sqrt{C_{ij}}} (C - C_0) / \sum\sum g_{ij} C_{ij} \cdots\cdots (3.8)$$

이 되고, 이때의 最小分散 $\overline{V}_{min}$는

$$\overline{V}_{min} = \frac{(\sum\sum g_{ij}\sqrt{C_{ij}})^2}{C - C_0} \cdots\cdots (3.9)$$

이다.

한편, $2g_{ij}^2 = (w_{.j}^2 + w_{i.}^2)s_{ij}^2 = \frac{n_{.j}^2 + n_{i.}^2}{n^2} s_{ij}^2$에서 $\frac{n_{.j}^2 + n_{i.}^2}{n^2} = 1$

로 두면, $g_{ij}^2 = \frac{s_{ij}^2}{2}$이고 式(3.7)와 式(3.8)은 각각 다음과 같이

같이 유도된다.

$$n = \frac{(C - C_0)\sum\sum s_{ij}/\sqrt{C_{ij}}}{\sum\sum s_{ij}\sqrt{C_{ij}}} \cdots\cdots (3.10)$$

$$n_{ij} = \frac{s_{ij}}{\sqrt{C_{ij}}} (C - C_0) / \sum\sum s_{ij} C_{ij} \cdots\cdots (3.11)$$

또 이 경우의 $\overline{V}_{min}$은

$$\overline{V}_{min} = \sum_i \sum_j \frac{\frac{s_{ij}^2}{2}}{\frac{s_{ij}}{\sqrt{C_{ij}}} \frac{(C - C_0)}{\sum\sum s_{ij}\sqrt{C_{ij}}}}$$

$$= \frac{1}{2} (\sum i \sum j s_{ij} C_{ij})^2 / C - C_0 \cdots\cdots (3.12)$$

이다.

參 考 文 獻

〔1〕 Cochran, W.G. Sampling Techniques. New York: John Wiley & Sons, Inc., 1977

〔2〕 Kish, L. Survey Sampling. New York: John Wiley & Sons, Inc., 1965.

〔3〕 Rao, J.N.K. On Double Sampling for Stratification and Analytical Surveys. Biometrika., 60, 1973.

〔4〕 Stuart, A. A Simple Presentation of Optimum Sampling results. Jour. Stat.Soc., B, 1954.

〔5〕 Kim, I.C.: Bayes and Minimax Procedures in Double Sampling. Doctoral thesis, 1988.