光州保健專門大學 論文集 第18輯(1993)
The Journal of Kwangju Health
Junior College. Vol. XVIII.

層化抽出에서 比例配定과 最適配定에 關한 硏究

임상병리과
전임강사 송 정

Ⅰ. 序 論

標本調査는 母集團으로부터 적당한 任意標本을 抽出한 다음, 그 標本에 대하여 必要한 情報를 조사하고 그 情報를 기초로하여 母集團에 대한 結論을 推定하는 것이다. 이러한 標本調査는 標本抽出에서부터 이루어진다.

本 論文에서는 標本抽出法중의 하나인 層化抽出에서 比例配定과 最適配定을 比較硏究코자 한다. 標本의 配定은 層化抽出에 있어서 標本의 크기 n 을 각 層에 합리적으로 配定하는 方法으로 각 層의 크기와 層內變動에 따라서 決定되며, 이상적인 配定은 最小의 費用으로 最高의 精度를 얻는 것이다. 推定의 精度는 分散의 크기에 의하여 決定되기 때문에 比例配定法과 最適配定法의 分散의 크기를 比較해서 어느 抽出法이 더 效率的인지 알아보는 것이 本 論文의 目的이다.

Ⅱ. 記 號

本 論文에서 必要한 다음 몇가지 記號를 定한다.

N : 母集團의 크기
N_h : h 層의 母集團의 크기
n : 標本의 크기
n_h : h 層의 標本의 크기
$f_h = \frac{n_h}{N_h}$: h 層의 抽出率
S_h^2 : h 層의 分散
$\overline{X_{st}}$: 層化抽出에서 母平均 μ의 推定量
$V(\overline{X_{st}})$: 層化抽出에서 推定量 $\overline{X_{st}}$ 의 分散
$V(\overline{X_{prop}})$: 比例配定에서 推定量 $\overline{X_{st}}$ 의 分散

$V(\overline{X_{opt}})$: 最適配定에서 推定量 $\overline{X_{st}}$ 의 分散
C : 標本調査의 總費用
C_o : 固定費用
C_h : h 層의 標本抽出 單位當 費用 (變動費用)

Ⅲ. 比例配定과 最適配定

1. 比例配定

各 層에 標本을 配分하는 方法으로서 가장 단순하고 자주 이용되는 方法은 標本의 크기를 層의 크기에 比例的으로 配定하는 方法이다. 즉 比例配定은 層化抽出에 있어서 각 層의 크기 N_h 는 알 수 있으나 層內變動에 관하여는 전혀 알 수 없는 경우, 標本 크기 n 을 각 層의 크기 N_h 에 比例하여 配定하는 方法으로 Bowley (1926) 가 제시하였다. 比例配定은

$$\frac{n_1}{N_1} = \frac{n_2}{N_2} = \cdots = \frac{n_k}{N_k} = \frac{n}{N} = f \text{ (抽出率)}$$

로 부터 h 層의 標本의 크기 n_h 를 母集團의 加重値 W_h 에 比例的으로 配分하는 方法이다. $(W_h = \frac{N_h}{N})$

그러므로 각 層의 標本의 크기는

$$n_h = n\frac{N_h}{N} \qquad (h = 1, 2, \ldots, k)$$

으로 나타내며 작은 層은 큰 層보다 層內變動이 작다는 가정을 전제로 하고 있다. 이와같은 比例的 配定에 의한 層化抽出 方法은 각 層의 抽出率이 일정하므로 計算上에 利點이 있고 시간이 절약된다. 그 이유는 比例的 配定이 스스로 加重化하는 標本이 되기 때문이다.

< 정 리 1 > 比例配定이 쓰여질때 母平均 μ 의 推定量은

$$\overline{X_{st}} = \frac{1}{n}\sum_{h=1}^{k}\sum_{i=1}^{n_h} X_{hi} \quad \cdots\cdots\cdots\cdots\cdots\cdots (1)$$

이다.[5]

(증 명) 層化抽出에서 標本平均

$$\overline{X_{st}} = \frac{\sum_{h}^{k} N_h X_h}{\sum_{h}^{k} N_h}$$

에 比例配定의 조건

$$N_h = \frac{n_h N}{n}$$

을 대입하여 μ 의 推定量을 구한다.

< 정 리 2 > 比例配定에 있어서 推定量 $\overline{X_{st}}$ 의 分散은

$$V(\overline{X_{prop}}) = \frac{N-n}{N}\sum^{k}\frac{N_h}{N}\frac{S_h^2}{n} \quad (fpc \neq 1)$$

$$= \sum^{k}\frac{N_h}{N}\frac{S_h^2}{n} \quad (fpc = 1) \quad \ldots (2)$$

이다.[5] 여기서 $\frac{N-n}{N}$ 은 分散의 有限母集團修正係數이고 fpc 로 나타낸다.

(증 명) 層化抽出에서 推定量 $\overline{X_{st}}$ 의 分散 [8]

$$V(\overline{X_{st}}) = \frac{1}{N^2}\sum^{k}N_h^2\frac{N_h-n_h}{N_h}\frac{S_h^2}{n_h}$$

에 比例配定의 조건

$$n_h = n\frac{N_h}{N}$$

을 대입하여 구한다.

比例配定法은 費用과 관계가 없고 단지 層의 크기만 알면 되기 때문에 작업이 간단하여 실용성이 높다. 또한 각 層의 分散이 같은 값이면 比例配定은 다음에 말하는 最適配定과 같은 效果를 같게 된다.

2. 最適配定

最適配定은 標本의 크기 n 을 각 層에 配定하는데 있어서 一定한 費用下에서 標本分散 $V(\overline{X_{st}})$ 을 最小로 하는 原則, 또는 一定한 標本分散 $V(\overline{X_{st}})$ 下에서 費用을 最小로 하는 原則을 가지고 각 層의 標本 크기 n_h 를 결정하는 方法으로 Neyman (1934)에 의하여 처음 제시 되었다. 그 후 Mahalanobis (1944)가 費用函數를 도입하였고, Stuart (1954)는 Cauchy 부등식을 이용하여 最適配定을 유도 하였다.

層化抽出에서 標本調査의 總費用을 C, 固定費用을 C_o, 變動費用으로서 h 번째 層에서 抽出한 標本單位에 필요한 調査費用을 C_h 로 나타낼때 標本調査의 總費用은

$$C = C_o + \sum^{k}C_h n_h \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3)$$

로 나타내며, 이를 費用函數라 한다. 위 식 (3)에서 分散을 最小化하는데 固定費用 C_o 는 관계가 없으므로 $C - C_o$ 를 C' 으로 표시하면

$$C' = \sum^{k}C_h n_h \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (4)$$

이다. 식 (4) 의 費用函數에서 固定된 豫算 C', 標本의 크기 n 을 $n_1, n_2 \ldots, n_k$ 로 配分함으로써 標本分散을 最小化 시키는 方法, 즉 $V(\overline{X_{st}})$ 를 最小化하기 위하여 層에 標本의 크기 n 을 配定하는 方法을 最適配定이라 한다.

< 정 리 3 > 最適配定에서 식 (4)와 같은 費用이 주어졌을때 $\overline{X_{st}}$ 의 分散을 最小로 하는 각 層의 標本의 크기 n_h 의 값은

$$n_h = n \frac{\frac{N_h S_h}{\sqrt{C_h}}}{\sum^{k} \frac{N_h S_h}{\sqrt{C_h}}} \quad \cdots\cdots\cdots\cdots (5)$$

이다. [6]

(증 명) 식 (4)가 一定하다는 條件下에서 $V(\overline{X_{st}})$ 이 最小되게 각 層의 標本의 크기 n_h 를 구하면 된다.

$$\Psi = \frac{1}{N^2} \sum^{k} N_h^2 \frac{N_h - n_h}{N_h} \frac{S_h^2}{n_h} + \lambda \left(\sum^{k} C_h n_h - C \right)$$

에서 Lagrange 乘數法을 사용하여

$$\frac{\partial \Psi}{\partial n_h} = 0$$

을 구하면

$$- \frac{N_h^2}{N^2} \frac{S_h^2}{n_h^2} + \lambda C_h = 0$$

을 얻는다. 그러므로

$$n_h = \frac{1}{\sqrt{\lambda}} \frac{N_h S_h}{N \sqrt{C_h}}$$

이다. λ 를 구하기 위해 n_h 의 양변에 Σ 를 취하면

$$\sum^{k} n_h = n = \frac{1}{\sqrt{\lambda}} \sum^{k} \frac{N_h S_h}{N \sqrt{C_h}}$$

$$\sqrt{\lambda} = \frac{1}{n} \frac{1}{N} \sum^{k} \frac{N_h S_h}{\sqrt{C_h}}$$

$\sqrt{\lambda}$ 를 n_h 에 대입하면 h 層의 標本의 크기는

$$n_h = n \frac{\frac{N_h S_h}{\sqrt{C_h}}}{\sum^{k} \frac{N_h S_h}{\sqrt{C_h}}}$$

을 얻는다.

위식 (5)에서 h 層의 標本의 크기 n_h 는 $N_h S_h$ 의 크기에 比例하여 N_h 나 S_h 가 크면, h 層의 標本의 크기 n_h 는 커진다. 또한 費用要素인 $\sqrt{C_h}$ 는 層의 費用 C_h 가 작을수록 層으로부터 많은 標本數를 택하게 됨을 알 수 있다.

< 정 리 4 > 最適配定에 있어서 推定量 $\overline{X_{st}}$ 의 分散은

$$V(\overline{X_{opt}}) = \frac{1}{N^2} \frac{1}{n} \sum^{k} N_h S_h \sqrt{C_h} \sum^{k} \frac{N_h S_h}{\sqrt{C_h}} - \frac{1}{N^2} \sum^{k} N_h S_h^2 \quad \cdots (6)$$

이다. [5]

(증 명) 層化抽出에서 推定量 $\overline{X_{st}}$ 의 分散 [8]

$$V(\overline{X_{st}}) = \frac{1}{N^2}\sum^{k} N_h^2 \frac{N_h - n_h}{N_h}\frac{S_h^2}{n_h}$$

에 (5)식

$$n_h = n\frac{\frac{N_hS_h}{\sqrt{C_h}}}{\sum^{k}\frac{N_hS_h}{\sqrt{C_h}}}$$

을 대입하여 구한다.

層化抽出에서 모든 層의 單位費用 C_h 가 一定하다면 즉, $C_h = C_f$ (固定)로 두면 費用函數로서 식 (3)은

$$C = C_o + \sum^{k} C_h n_h = C_o + C_f\sum^{k} n_h = C_o + C_f\, n$$

이므로

$$n = \frac{C - C_o}{C_f}$$

와 같이 나타낼 수 있으며 이처럼 n 이 固定되었을때 分散 $V(\overline{X_{st}})$ 을 最小로 하는 각 層의 標本의 크기 n_h 의 값은

$$n_h = n\frac{N_hS_h}{\sum^{k} N_hS_h} \quad \cdots\cdots\cdots\cdots\cdots\cdots \quad (7)$$

이다. 위 식 (7)에서 標本의 크기 n 은 N_hS_h 에 比例的으로 配定되고 있음을 알 수 있다. 즉 層內의 變動이 크면 더욱 큰 n_h 값을 갖게 된다. 이러한 方法은 最適配定의 특수한 경우이며 Neyman 配定 이라고도 한다.

< 정 리 5 > Neyman 配定에 의하여 얻어지는 $\overline{X_{st}}$ 의 分散은

$$V(\overline{X_{opt}}) = \frac{1}{N^2}\frac{1}{n}\left(\sum^{k} N_hS_h\right)^2 - \frac{1}{N^2}\sum^{k} N_hS_h^2 \quad \cdots \quad (8)$$

이다.

(증 명) 層化抽出에서 推定量 $\overline{X_{st}}$ 의 分散 [8]

$$V(\overline{X_{st}}) = \frac{1}{N^2}\sum^{k} N_h^2 \frac{N_h - n_h}{N_h}\frac{S_h^2}{n_h}$$

에 (7)식

$$n_h = n\frac{N_hS_h}{\sum^{k} N_hS_h}$$

을 대입하여 구한다.

Neyman 配定은 層이 이질적이거나 크기가 큰 層에 표본수를 많이 割當하는 것을 나타낸다. 따라서 각 層의 分散이 모두 같은 경우에는 比例配定과 일치한다.

Ⅳ. 比例配定과 最適配定의 比較

最適配定 (Neyman 配定)에 의한 $\overline{X_{st}}$ 의 分散은

$$V(\overline{X_{opt}}) = \frac{(\sum^{k} W_h S_h)^2}{n} - \frac{\sum^{k} W_h S_h^2}{N} \qquad (W_h = \frac{N_h}{N})$$

이고, 比例配定에 의한 $\overline{X_{st}}$ 의 分散은

$$V(\overline{X_{prop}}) = \frac{N-n}{N} \sum^{k} \frac{W_h S_h^2}{n} \qquad (fpc \fallingdotseq 1)$$

$$= \sum^{k} \frac{W_h S_h^2}{n} - \sum^{k} \frac{W_h S_h^2}{N} \qquad (W_h = \frac{N_h}{N})$$

이므로 이들의 차를 구해보면

$$V(\overline{X_{prop}}) - V(\overline{X_{opt}})$$

$$= \frac{1}{n} [\sum^{k} W_h S_h^2 - (\sum^{k} W_h S_h)^2]$$

$$= \frac{1}{n} [\sum^{k} W_h (S_h^2 - 2 S_h (\sum^{k} W_h S_h) + (\sum^{k} W_h S_h)^2)]$$

$$= \frac{1}{n} [\sum^{k} W_h (S_h - \sum^{k} W_h S_h)^2]$$

$$= \frac{1}{n} \sum^{k} W_h (S_h - \overline{S_h})^2 \quad \cdots\cdots\cdots\cdots\cdots (9)$$

여기서 $\overline{S_h} = \sum^{k} W_h S_h$는 S_h의 加重平均이다. 식 (9)를 다시쓰면

$$V(\overline{X_{prop}}) = V(\overline{X_{opt}}) + \frac{1}{n} \sum^{k} W_h (S_h - \overline{S_h})^2 \quad \cdots\cdots (10)$$

으로 나타낼 수 있다. 식(10)에서 $\sum^{k} W_h (S_h - \overline{S_h})^2$은 層間 S_h의 變動이며 이것이 커질수록 $V(\overline{X_{opt}})$이 $V(\overline{X_{prop}})$ 보다 작아진다. 즉 S_h와 $\overline{S_h}$의 差가 크면 클수록 $(S_h - \overline{S_h})^2$의 값이 커져서 $V(\overline{X_{opt}})$와 $V(\overline{X_{prop}})$의 差가 더욱 커지게 된다. 그러므로 層의 크기 N_h의 差가 크고, 층별 S_h간의 差가 커질수록 最適配定이 比例配定 보다 效率이 높아진다는 것을 알 수 있다. 예를들면 기업체의 생산량 조사는 업체별 變動 S_h가 크고, 層間 N_h의 差도 크므로 最適配定을 사용하는 것이 比例配定을 사용하는 것보다 效率이 높아진다.

< 예 1 > 어느 도시에 있는 1000개의 병원 중 규모에 따라 3개의 層으로 나누었을 때 각 層에 대한 기본적인 자료는 다음과 같다.

층	N_h	S_h	$N_h S_h$	$N_h S_h^2$	C_h	$N_h S_h \sqrt{C_h}$	$N_h S_h / \sqrt{C_h}$
1	600	20	12000	240000	1	12000	12000
2	300	30	9000	270000	2	12730	6360
3	100	50	5000	250000	3	8660	2890
계	1000		26000	760000		33390	21250

이 자료를 이용하여 標本의 크기 n = 500 일때 比例配定, 最適配定과 Neyman 配定에 의하여 각각의 分散을 구하자.

1) 比例配定

$$V(\overline{X_{prop}}) = \frac{N-n}{N}\sum\frac{N_h}{N}\frac{S_h^2}{n} \qquad (fpc \fallingdotseq 1)$$

$$= \frac{1000-500}{1000}\frac{760000}{1000\cdot 500}$$

$$= 0.76$$

2) 最適配定

$$V(\overline{X_{opt}}) = \frac{1}{N^2}\frac{1}{n}\sum N_h S_h\sqrt{C_h}\sum\frac{N_h S_h}{\sqrt{C_h}} - \frac{1}{N^2}\sum N_h S_h^2$$

$$= \frac{1}{1000^2}\frac{1}{500}33390\cdot 21250 - \frac{1}{1000^2}760000$$

$$= 0.66$$

3) Neyman 配定

$$V(\overline{X_{opt}}) = \frac{1}{N^2}\frac{1}{n}(\sum N_h S_h)^2 - \frac{1}{N^2}\sum N_h S_h^2$$

$$= \frac{1}{1000^2}\frac{1}{500}26000^2 - \frac{1}{1000^2}760000$$

$$= 0.59$$

따라서 이 < 예 1 >에서는

$$V(\overline{X_{opt}}) \leq V(\overline{X_{prop}})$$

인 관계가 성립한다.

推定의 精度는 分散의 크기에 의해서 결정되므로 最適配定法이 比例配定法보다 精度가 높다는 것을 알 수 있다. 즉 最適配定法이 比例配定法보다 더 效率的이다.

참고문헌

1 . Berger, J.O. : Statistical Decision Theory, springerverlag (1980).

2 . Chatterjee, S. : " A study of optimum allocation in multivariable stratified surveys", skand, Akt., **55**, 73-80 (1972).

3 . Cochran, W. G. : Sampling Techniques, 3rd Ed. John Wiley and Sons (1977).

4 . Hogg, R. V. and Craig, A.T. : Introduction to Mathematical Statistics, 4th Ed., Collier Macmillan International, Inc. (1978).

5 . Yamane, Taro : Elementary Sampling Theory, Prentice-Hall, Inc., Englewood Cliffs, N. J. (1967).

6 . 김종호, 표본조사법, 서울, 자유아카데미, (1991).

7 . 남궁평·김연형, 표본이론, 서울, 박영사, (1986).

8 . 박홍래, 통계조사론, 서울, 영지문화사, (1989).

A Study on the proportional allocation and optimum allocation in the stratified sampling

Song, Jeong
Dept. of Clinical Pathology
Kwangju Health Junior College

> *Abstract* <

The simplest and most useful way of allocating a sample among strata is to allocate it proportionally to the size of the strata.

This method of allocating the sample n among strata so as to minimize $V(\overline{X}_{st})$ is called optimum allocation.

This paper has been made on which one of the proportional allocation and optimum allocation is more efficient.

It has been found that in order to obtain the higher precision of estimates, variance must be smaller. As a result of the comparison of the above mentioned sampling methods with each other, the following has been arrived at :

$$V(\overline{X}_{opt}) \leq V(\overline{X}_{prop})$$

According to the above inequality equations it can be concluded that the variation of the optimum allocation is smaller than the proportional allocation because of the precision of estimates is determined by the size of variance.

Therefore it can be decided that the optimum allocation is more efficient.