# Minimax Procedures in Double Sampling

*Kim Ik-chan, Shin Min-woong**

# 二重 標本抽出에서의 Minimax 節次

金益贊, 中敏雄*

## INTRODUCTION

In conducting sample survey, we have adopted one of the following two procedures: ( 1) to get an estimate of maximum precision for a given total cost of the survey, or ( 2) to get an estimate of given precision for a minimum total cost of the survey.

We may consider jointly the losses resulting from the errors in the estimates and from the cost of sampling, and to employ such sampling and estimation procedures as will, in some sénce, "minimize" the total expected loss. We shall take as loss function the sum of two components, one proportional to the squre of the error of the estimate and the other proportional to the cost of obtaining the sample.

## MINIMAX ESTIMATES

We are given a sample space X, a space of probability distribution on X, $P_\Omega = \{P_w : w \in \Omega\}$, where $\Omega$ is an index set and a function g defined on $\Omega$ whose value $g(w)$.

A non-randomized decision function for us, usually called an estimate, is a numerical function $\delta$ defined on X, specifying for each x the number $a \in A$ which will be chosen to estimate $g(w)$ when that x is observed. The loss function L defined on $\Omega$xA is non-negative and is the loss incurred when $g(w)$ is estimated by a. The risk function R is defined by

$R(w, \delta) = E_w L(w, \delta)$.

The average risk corresponding to an a priori distribution $\xi$ for nature and a decision func-

師範大學 助教授, 韓國外大 教授*

tion $\delta$ used for estimation of g($w$) is obviously

$$R(\xi,\ \delta)=\underset{X\ \Omega}{E\ E}[(\delta(x)-g((w))^2 \mid x].$$

This is minimized by choosing, for each x, the number $\delta$(x) which does it is clearly

$$\delta_\xi(x)=E(g(w)\mid x).$$

## MINIMAX ESTIMATION IN DOUBLE SAMPLING

The population of N units is divided into subpopulation of $N_1$, $N_2$, ……, $M_L$ units. The $n_h$ is the number of units in the h-th stratum. The $\overline{Y}_h=\Sigma y_{hi}/N_h$ is a true mean and $\bar{y}_h=\Sigma y_{hi}/n_h$ is a sample mean in the h-th stratum. We consider the risks in double sampling for stratification. The first sample is a simple random sample of size n'.

Let

$$w_h=n_h'/n'$$

= proportion of first sample falling in stratum h.

The second sample is a stratified random sample of size n in which the $y_{hi}$ are measured, $n_h$ units are drawn from stratum h. The second sample in stratum h is a random subsample from the $n_h'$ in the stratum. The objective of the first sample is to estimate the strata weights; that of the second sample is to estimate the strata means $\overline{Y}_h$. The population mean $\overline{Y}=\Sigma W_h\overline{Y}_h$. As an estimate we use

$$\overline{y}_{st}=\Sigma W_h\overline{y}_h$$

Let $c_h$ be the known cost of collecting information from a unit in stratum h and let $c'$ be the cost of classification per unit. We shall investigate for stratification sampling some Bayes and minimax procedures, first for an infinite and then for finite populations.

### 1. Infinite populations

Suppose that the i-th stratum consists of an infinite population with unknown mean $\mu_i$ and known upper bound $\sigma_i^2$ for all variances, i= 1,2,…, k and that we have to estimate a linear function of the $\mu_i$, say $U=\Sigma a_i\mu_i$ where the $a_i$ are some given real numbers. Without loss of generality we may take $\Sigma a_i=1$. For the sake of simplicity we shall assume that none of the $a_i$ is zero. The loss function U is given by

$$L(U,\ \delta)=(\delta-U)^2+\Sigma c_h n_h+c'n'$$

where $n_h(>0)$ is the size of the sample chosen from the h-th stratum, and $\delta$ is a function of the sample $\{X_{ij};\ i=1,2,\cdots\cdots,k;\ j=1,2,\cdots\cdots,n_i\}$, where $X_{ij}$ is the j-th observation from the i-th stratum.

We may regard the $n_i$ as fixed for the purpose of finding the estimates. Letting $\delta^*$ be a minimax estimate for given $n_i$, we shall choose the $n_h$ so as to minimize the risk,

$$(3.2)\quad R(U,\ \delta^*)=E(\delta^*-U)^2+\Sigma c_i n_i+c'n'$$

as a function of the $n_i$ and n'.

#### A. Minimax estimate for given $n_i$.

Letting $\theta\to\infty$, we see that Bayes risk $r_\theta\to r$, where $r=\Sigma a_i\,\sigma_i^2/n_i$. If we can find some estimate $\delta^*$ with risk ≤r, then $\delta^*$ will be a minimax estimate. Let us try the limiting Bayes estimate,

$$\lim_{\theta}\ \delta_\theta(x)=\Sigma a_i\overline{X}_i=\delta^*(x).$$

Since the $\overline{X}_i$ are normal and independent with means $\mu_i$ and variances $\sigma_i^2/n_i$, $\Sigma a_i\overline{X}_i$ is normal with mean $\Sigma a_i\mu_i=U$ and variance $\Sigma a_i\sigma_i^2/n_i$, where $\overline{X}_i$ is the sample mean from the i-th stratum. Hence the risk corresponding to the estimate $\delta^*(x)=\Sigma a_i\overline{X}_i$ is equal to r which proves that $\Sigma a_i \overline{X}_i$ is a minimax estimate of U for given $n_i$

B. Minimax strategy for choosing the $n_i$.

Restoring the term $\Sigma c_i n_i$ and $c' n'$ in the risk function, we can choose "optimum" $n_i$ if the variances of the populations in different strata are known by minimizing the risk as a function of the $n_i$ and $n'$

We choose optimum $n_i$ corresponding to the variances in the different strata as the $\sigma_i^2$. For given $n_i$, the risk corresponding to $\delta^*$ is given by

$$(3.3)\quad R(w,\ \delta^*) = \Sigma \frac{W_h^2 S_h^2}{n_h} + \frac{g'}{n'} \Sigma W_h (\overline{Y}_h - \overline{Y})^2 + \Sigma c_h n_h + c' n'$$

where $g' = (N-n')/(N-1)$ and $\frac{n_h}{N_h}$ is negligible in infinite population.

Theorem 1. Suppose that we have the risk of the form (3.3) in the infinite population. Then the risk is minimum when $n_i$ is the integer nearest to $(W_h^2 S_h^2 / c_h + \frac{1}{4})^{\frac{1}{2}}$ and $n'$ is the integer nearest to

$$(g' \Sigma W_h (\overline{Y}_h - \overline{Y})^2 / c' + \frac{1}{4})^{\frac{1}{2}}$$

Proof. We want to choose the $n_h$ and $n'$ so that the risk is minimum under the restriction that the $n_h$ and $n'$ are positive integers. The i-th term on the right hand side of (3.3) depends on $n_h$ and $n'$. It is sufficient to minimize $W_h^2 S_h^2 / n_h + c_h n_h$ and $g'/n' \Sigma W_h (\overline{Y}_h - \overline{Y})^2$. Denoting $W_h^2 S_h^2 / n_h + c_h n_h$ by $f(n_h)$, we see that

$$(3\cdot4)\quad f(n_h + 1) - f(n_h) = c_h - W_h^2 S_h^2 / n_h (n_h + 1)$$

To minimize $f(n_h)$, we choose the smallest positive integral value for $n_h$ for which the difference (3.4) is positive, in other words, the smallest positive integer $n_h$ for which $(n_h + \frac{1}{2})^2$ exceeds $W_h^2 S_h^2 / c_h + \frac{1}{4}$ This gives

$n_h$ = integer nearest to $(W_h^2 S_h^2 / c_h + \frac{1}{4})^{\frac{1}{2}}$

Similarly, denoting $g'/n' \Sigma W_h^2 (\overline{Y}_h - \overline{Y})^2$ by g( n' we see that

$$g(n'+1) - g(n') = c' - g' \Sigma W_h (\overline{Y}_h - \overline{Y})^2 / n'(n'+1)$$

and

$n'$ = integer nearest to

$$(g' \Sigma W_h (\overline{Y}_h - \overline{Y})^2 / c' + \frac{1}{4})^{\frac{1}{2}}$$

2. Finite population

Suppose that $x_{ij}$ ( i= 1,2,……, k; j= 1,2, ……, $N_i$) denotes the j-th unit in the i-th stratum. Suppose further that the $N_i$ are known, the means $\mu_i$ of the strata are unknown. We now choose the $n_i$ so that the minimax risk for given $n_i$ and largest allowed variances in the strata is minimum under the restriction that the $n_i$ are positive integers. This risk is given by

$$(3.5)\quad R(w,\ \delta^*) = \Sigma W_h^2 (\frac{1}{n_h} - \frac{1}{N_h}) S_h^2 + g'/n' \Sigma W_h (\overline{Y}_h - \overline{Y}) + \Sigma c_h n_h + c' n'$$

This expression differs from that in (3.3) by a quantity which is independent of $n_1, n_2, \cdots, n_k$.

Theorem 2. Suppose that we have the risk of the form (3.5) in the finite population. Then the risk is minimum when $n_i$ is the integer nearest to $((N_h^2/N^2)(S_h^2/c_h) + \frac{1}{4})^{\frac{1}{2}}$ and $n'$ is the integer nearest to

$$((g'c') \Sigma (N_h/N)(\overline{Y}_h - \overline{Y})^2 + \frac{1}{4})^{\frac{1}{2}}$$

Proof. The i-th term or the right hand side of (3.5) depends on $n_h$ and $n'$. It is sufficient to minimize $W_h^2 S_h^2 (1/[n_h(n_h + 1)] - 1/[N_i(N_i + 1)])$ and $g'n' \Sigma W_h (\overline{Y}_h - \overline{Y})^2 + c'n'$.

Denoting $W_n^2 S_h^2(1/n_h - 1/N_h) + c_h n_h$ by $f(n_h)$ we see that $f(n_i+1) - f(n_i) = c_i - [(1/(n_i+1) - 1/(N_i + 1) - (1/n_i - 1/N_i)] W_h^2 S_h^2$

Hence the smallest positive integer $n_h$ minimizing the risk is

$n_h$ = integer nearest to

$$((N_h^2/N^2)(S_h^2/c_h) + \frac{1}{4})^{\frac{1}{2}}$$

Similarly, we get

$n'$= integer nearest to

$$((g'/c')\Sigma(N_h/N)(\overline{Y}_h - \overline{Y})^2 + \frac{1}{4})^{\frac{1}{2}}$$

## Literature Cited

Aggarwal, O. P. 1959. Bayes and minimax procedures in sampling from finite and infinite population- I. J. Ann. Math. Stat., 30; 206-218.

__________ 1966. Bayes and minimax procedures for estimating the arithmetic mean of a population with two-stage sampling. J. Ann. Math. Stat., 37; 1186-1196.

Cochran, W. G. 1973. Sampling Technique, 3rd edition. John Wiley & Sons, Inc.

Rao, J. N. K 1973. On double sampling for stratification and analytical surveys. Biometrika, 60; 125-133.

## 國 文 抄 錄

本 論文에서는 二重 層別 標本抽出에서 推定誤差의 제곱과 費用의 合에 의한 risk 함수를 정의하고 minimax 절차에 의해서 risk를 최소화 하는 標本의 크기를, 無限母集團과 有限母集團에서 구하는 問題를 研究하였다.