ZASTOSOWANIA MATEMATYKI
APPLICATIONES MATHEMATICAE
XVIII, 4 (1985), p.571-576

S. TRYBUŁA and M. WILCZYŃSKI (Wrocław)

ESTIMATION UNDER SAMPLING FROM A FINITE POPULATION

In the paper the problem of minimax estimation of the parameter $M/N = (M_1/N, \ldots, M_r/N)$ of the multivariate hypergeometric distribution (1) is considered in the case when M and N are unknown and the loss function is given by (2). It is proved that there exists a minimax estimator $d = (d_1, \ldots, d_r)$ of the parameter M/N which is of the form (4).

In practice, we often meet the following situation. A lot consisting of N units of a product has been produced. The units are classified into r categories, the ith category containing M_i units $i = 1, \ldots, r$. A sample of size n is taken from the lot in which $x_1, \ldots, x_r$ units of categories $1, \ldots, r$ are observed. The numbers $M_1, \ldots, M_r, N$ are unknown and $n < N$. The problem is to estimate $M/N = (M_1/N, \ldots, M_r/N)$. This leads to the estimation of the parameter $M/N = (M_1/N, \ldots, M_r/N)$ of a multivariate hypergeometric distribution with unknown M and N. Thus let

$$P(X = x) = P(X_1 = x_1, \ldots, X_r = x_r) = \frac{\binom{M_1}{x_1} \cdot \ldots \cdot \binom{M_r}{x_r}}{\binom{N}{n}}. \tag{1}$$

Let the loss function be

$$L(M, N, d) = \sum_{i,j=1}^{r} c_{ij}\left(\frac{M_i}{N} - d_i\right)\cdot\left(\frac{M_j}{N} - d_j\right) \tag{2}$$

where $d = (d_1, \ldots, d_r)$ is an estimate of $M/N = (M_1/N, \ldots, M_r/N)$ and the matrix $C = \|c_{ij}\|$ is nonnegative definite.

Let $d(x) = (d_1(x), \ldots, d_r(x))$ be an estimator of

$$M/N = (M_1/N, \ldots, M_r/N)$$

and let

$$R(M, N, d) = \mathrm{E}\big(L(M, N, d(X))\big|M, N\big)$$

be the risk function.

DEFINITION. An estimator $d^0(x) = (d_1^0(x), \ldots, d_r^0(x))$ of M/N is called *minimax* if

$$\sup_{(M,N)} R(M, N, d^0) = \inf_{d} \sup_{(M,N)} R(M, N, d). \tag{3}$$

LEMMA 1. *Let the loss function be of the form* (2) *and let the estimator* $d = (d_1, \ldots, d_r)$ *of* M/N *fulfil the condition* $d_i(X) = a_i X_i + b_i$ *for* $i = 1, \ldots, r$, *where* a_i, b_i *are constants. If the matrix* $C = \|c_{ij}\|$ *is nonnegative definite and* $M_i/N = p_i$, $i = 1, \ldots, r$, *are fixed then the risk* $R(M, N, d)$ *is a nondecreasing function of* N.

Proof. It is well known that

$$m_i = \mathrm{E}(X_i|M, N) = n\frac{M_i}{N},$$

$$\mathrm{E}\big((X_i - m_i)^2|M, N\big) = n\frac{N-n}{N-1}\cdot\frac{M_i}{N}\cdot\left(1-\frac{M_i}{N}\right), \quad i = 1, \ldots, r,$$

$$\mathrm{E}\big((X_i - m_i)(X_j - m_j)|M, N\big) = -n\frac{N-n}{N-1}\cdot\frac{M_i}{N}\cdot\frac{M_j}{N}, \quad i, j = 1, \ldots, r,\ i \neq j.$$

We obtain

$$\begin{aligned} R(M, N, d) &= \mathrm{E}\left(\sum_{i,j=1}^{r} c_{ij}\left(\frac{M_i}{N} - a_i X_i - b_i\right)\left(\frac{M_j}{N} - a_j X_j - b_j\right)\middle| M, N\right) \\ &= -n\frac{N-n}{N-1}\sum_{i,j=1}^{r} c_{ij} a_i a_j \frac{M_i}{N}\frac{M_j}{N} + n\frac{N-n}{N-1}\sum_{i,j=1}^{r} c_{ii} a_i^2 \frac{M_i}{N} + \\ &\qquad + \sum_{i,j=1}^{r} c_{ij}\left(b_i - (1 - a_i n)\frac{M_i}{N}\right)\left(b_j - (1 - a_j n)\frac{M_j}{N}\right) \\ &= \frac{n}{2}\frac{N-n}{N-1}\sum_{i,j=1}^{r} (c_{ii} a_i^2 + c_{jj} a_j^2 - 2c_{ij} a_i a_j)\frac{M_i}{N}\cdot\frac{M_j}{N} + \\ &\qquad + \sum_{i,j=1}^{r} c_{ij}\left(b_i - (1 - a_i n)\frac{M_i}{N}\right)\left(b_j - (1 - a_j n)\frac{M_j}{N}\right). \end{aligned}$$

Then the lemma results from the fact that for a nonnegative definite matrix C

$$c_{ii}a_i^2 + c_{jj}a_j^2 - 2c_{ij}a_i a_j \geqslant 0 \quad \text{for each } i, j = 1, \ldots, r.$$

Let

$$d_i(X) = \frac{X_i + \beta_i \cdot \sqrt{n}}{n + \sqrt{n}} \tag{4}$$

where $\beta = (\beta_1, \ldots, \beta_r) \in P = \{p = (p_1, \ldots, p_r)\colon p_i \geqslant 0,\ i = 1, \ldots, r,\ p_1 + \ldots + p_r = 1\}$. Then

$$R(M, N, d) = \frac{1}{(\sqrt{n}+1)^2}\left\{\sum_{i,j=1}^{r} c_{ij}\left[\frac{n-1}{N-1}\cdot\frac{M_i}{N}\cdot\frac{M_j}{N} + \left(\beta_i - 2\frac{M_i}{N}\right)\beta_j\right] + \sum_{i,j=1}^{r} c_{ii}\frac{N-n}{N-1}\cdot\frac{M_i}{N}\right\}.$$

When $N \to \infty$ and $M_i/N = p_i$ for $i = 1, \ldots, r$, we obtain

$$R(M, N, d) \xrightarrow[N\to\infty]{} \frac{1}{(\sqrt{n}+1)^2}\left\{\sum_{i,j=1}^{r} c_{ij}\beta_i\beta_j + \sum_{i,j=1}^{r}(c_{ii} - 2c_{ij})\beta_j p_i\right\}. \tag{5}$$

THEOREM. *There exists a point* $p_0 = (p_1^0, \ldots, p_r^0) \in P$ *such that the estimator* d *defined by* (4) *with* $\beta_i = p_i^0$, $i = 1, \ldots, r$, *is minimax.*

Proof. Let $p_0 = (p_1^0, \ldots, p_r^0)$ be a solution of the equation (see [5]):

$$\sum_{i=1}^{r} c_{ii}p_i^0 - \sum_{i,j=1}^{r} c_{ij}p_i^0 p_j^0 = \max_{p\in P}\left\{\sum_{i=1}^{r} c_{ii}p_i - \sum_{i,j=1}^{r} c_{ij}p_i p_j\right\}.$$

It is easy to deduce that there exist a set $A = \{i_1, \ldots, i_k\} \subset \{1, \ldots, r\}$ and a constant y_0 such that

$$\begin{aligned} &2\cdot\sum_{j\in A} c_{ij}p_j^0 - c_{ii} = y_0 \quad \text{for each } i \in A, \\ &2\cdot\sum_{j\in A} c_{ij}p_j^0 - c_{ii} \geqslant y_0 \quad \text{for each } i \notin A \end{aligned} \tag{6}$$

and moreover, $k = 1$ iff $c_{ij} = c_0$ for each i, j. In the following we assume that there exists an A such that $k \geqslant 2$. Otherwise, each estimator $d = (d_1, \ldots, d_r)$ with $d_1 + \ldots + d_r = 1$ is minimax.

In view of a result obtained in [3] it follows that for each $N > n+1$ the estimator $d^N = (d_1^N, \ldots, d_r^N)$ where

$$d_i^N = \frac{X_i + p_i^0 \cdot \sqrt{n\cdot\dfrac{N-n}{N-1}}}{n + \sqrt{n\cdot\dfrac{N-n}{N-1}}}$$

is Bayes with respect to the prior distribution P_N of (M, N) in which N is given and

$$P_N(M_{i_1} = m_{i_1}, \ldots, M_{i_k} = m_{i_k}) = K \cdot \frac{\Gamma(m_{i_1} + e_{i_1}) \cdot \ldots \cdot \Gamma(m_{i_k} + e_{i_k})}{m_{i_1}! \cdot \ldots \cdot m_{i_k}!},$$

(7)

$$P_N(M_j = 0, j \notin A) = 1,$$

where

$$e_{i_j} = p_{i_j}^0 \cdot \frac{N \cdot \sqrt{n\frac{N-n}{N-1}}}{N - n - \sqrt{n\frac{N-n}{N-1}}}.$$

The Bayes risk is

$$(8) \quad r(P_N, d^N) = \frac{\frac{N-n}{N-1}}{\left(\sqrt{n} + \sqrt{\frac{N-n}{N-1}}\right)^2} \cdot \left\{ \sum_{i,j=1}^{r} c_{ij} p_i^0 p_j^0 - y_0 \right\}$$

$$\xrightarrow[N \to \infty]{} \frac{1}{(\sqrt{n}+1)^2} \left\{ \sum_{i,j=1}^{r} c_{ij} p_i^0 p_j^0 - y_0 \right\} = c.$$

Let $d = (d_1, \ldots, d_r)$ be an estimator mentioned in the theorem. From Lemma 1, and (5), (6) it follows that

$$R(M, N, d) \leqslant c.$$

Since $\sum_{i,j=1}^{r} c_{ij} p_i^0 p_j^0$ is a continuous function of p_i^0, $i = 1, \ldots, r$, the theorem results from the following lemma well known in the theory of decision functions [1]:

LEMMA 2. *Let $\{\hat{d}^N\}$ be a sequence of Bayes estimators of M/N for the prior distributions $\{\hat{P}_N\}$, let $\{r(\hat{P}_N, \hat{d}^N)\}$ be the corresponding sequence of Bayes risks defined for a given loss function, and let $\hat{c}$ be a constant. If $r(\hat{P}_N, \hat{d}^N) \to \hat{c}$ as $N \to \infty$ and d is an estimator such that $R(M, N, d) \leqslant \hat{c}$ for each (M, N) then d is minimax.*

Let $X = (X_1, \ldots, X_r)$ be a sample from the multinomial distribution

$$P(X = x) = P(X_1 = x_1, \ldots, X_r = x_r) = \frac{n!}{x_1! \cdot \ldots \cdot x_r!} \cdot p_1^{x_1} \cdot \ldots \cdot p_r^{x_r}$$

and let $d = (d_1, \ldots, d_r)$ be given by (4). From [5] it follows that the same theorem, as proved above, is true for the estimation of the parameter $p = (p_1, \ldots, p_r)$ of the multinomial distribution. Then, in the situation considered here, i.e. when N is unknown and $N \geqslant n$, one can suppose that N

is infinite in order to determine a minimax estimator of the parameter $p = M/N$. Then we can apply here some results proved for the multinomial distribution.

Examples.

(a) Let $r = 3$. Let the loss function be given by (2) where the matrix $C = \|c_{ij}\|$ is positive definite. Then a minimax estimator of the parameter M/N, of the form (4), can be easily determined (see [4]).

(b) Let $X = (X_{11}, \ldots, X_{1s_1}, \ldots, X_{r1}, \ldots, X_{rs_r})$ be a random variable distributed according to the multivariate hypergeometric distribution with parameters $M = (M_{11}, \ldots, M_{1s_1}, \ldots, M_{r1}, \ldots, M_{rs_r})$ and N. Let the loss function be

$$L(M, N, d) = \sum_{i=1}^{r} c_i\left(\frac{M_i}{N} - d_i\right) + \sum_{i=1}^{r}\sum_{j=1}^{r} c_{ij}\left(\frac{M_{ij}}{N} - d_{ij}\right),$$

where $M_i = \sum_{i=1}^{s_i} M_{ij}$ $(i = 1, \ldots, r)$, d_i, d_{ij} are estimates of M_i/N, M_{ij}/N respectively, $c_i \geqslant 0$, $c_{ij} > 0$ for $i = 1, \ldots, r$, $j = 1, \ldots, s_i$. Then there always exists a minimax estimator $d = (d_{11}, \ldots, d_{1s_1}, \ldots, d_{r1}, \ldots, d_{rs_r})$ of M/N of the form

$$d_{ij}(X) = \frac{X_{ij} + \beta_{ij} \cdot \sqrt{n}}{n + \sqrt{n}}$$

with $\beta_{ij} \geqslant 0$, $\sum_{i=1}^{r}\sum_{j=1}^{s_i} \beta_{ij} = 1$ (see [4]). In [4] is given a method how to determine the constants β_{ij}.

(c) In the case when $X = (X_1, \ldots, X_r)$, $M = (M_1, \ldots, M_r)$ and the loss function is given by

$$L(M, N, d) = \sum_{i=1}^{r} c_i\left(\frac{M_i}{N} - d_i\right)^2,$$

$c_i \geqslant 0$, the constants β_i in the minimax estimator (4) can be determined as follows:

Without loss of generality we may assume $c_1 \geqslant c_2 \geqslant \ldots \geqslant c_r \geqslant 0$. Let r_0 be the greatest index i for which $c_i \neq 0$ and let

$$L = \max_s \left[s \leqslant r_0 \colon \sum_{j=1}^{s} \frac{1}{c_j} > \frac{s-2}{c_s} \right], \qquad \delta = \frac{L-2}{\sum_{i=1}^{L} 1/c_i}.$$

Then

$$\beta_i = \begin{cases} \frac{1}{2}\left(1-\frac{\delta}{c_i}\right), & \text{when } i \leqslant L, \\ 0, & \text{when } i > L, \end{cases}$$

if $c_2 > 0$ and $\beta_1 = 1/2$ if only $c_1 > 0$ (see [3]).

In the special case $r = 2$ the estimator

$$d(X) = \frac{X + (\sqrt{n}/2)}{n + \sqrt{n}}$$

is a minimax estimator of the parameter M/N of the hypergeometric distribution

$$P(X = x) = \binom{M}{x}\binom{N-M}{n-x}\Big/\binom{N}{n},$$

when M and $N \geqslant n$ are unknown and the loss function is quadratic

$$L(M, N, d) = \left(\frac{M}{N} - d\right)^2.$$

This last result is probably known but we could not find the references.

For other results concerning the multinomial distribution, which can be applied here, see [2].

References

[1] J. L. Hodges and E. L. Lehmann, *Some problems in minimax point estimation*, Ann. Math. Statist. 21 (1950), p. 182-196.

[2] M. Rutkowska, *Minimax estimation of the parameters of the multivariate hypergeometric and multinomial distributions*, Zastos. Mat. 16 (1977), p. 9-21.

[3] S. Trybuła, *Some problems in simultaneous minimax estimation*, Ann. Math. Statist. 29 (1958), p. 245-253.

[4] –, *Some investigations in minimax estimation theory*, Dissertationes Math. 240, Warszawa 1985, 44 pp.

[5] M. Wilczyński, *Minimax estimation for the multinomial and multivariate hypergeometric distributions*, Sankhyā, to be published.

INSTITUTE OF MATHEMATICS
TECHNICAL UNIVERSITY WROCŁAW
50-370 WROCŁAW

Received on 8. 10. 1982