ZASTOSOWANIA MATEMATYKI
APPLICATIONES MATHEMATICAE
XVI, 1 (1977)

M. RUTKOWSKA (Wrocław)

MINIMAX ESTIMATION OF THE PARAMETERS OF THE MULTIVARIATE HYPERGEOMETRIC AND MULTINOMIAL DISTRIBUTIONS

In this paper the problem of minimax estimation of the parameters of the multivariate hypergeometric and multinomial distributions is considered. As a loss function the sum of the square errors of estimation and the cost of sampling are used.

1. Introduction. We define the minimax estimation problem as follows. Let X be a random variable, with values in the space $\mathscr{X}$, whose distribution depends on a parameter $\theta \in \Theta$. On the basis of the observed value X we want to estimate the value of the parameter θ. In the sequel we assume that X and θ are vectors. Let $L[f(x), \theta_0]$ be the loss to the statistician if he applies the estimator $f(x)$ when x is the observed value of X, and θ_0 is the value of the parameter θ. If we establish the function $f(x)$ and θ, we can find the risk

$$R(f, \theta) = \mathrm{E}\{L[f(x), \theta] \mid \theta\} = \int_{\mathscr{X}} L[f(x), \theta] dF(x \mid \theta).$$

It is our aim to determine a function f^0 such that

$$\sup_{\theta \in \Theta} R(f^0, \theta) = \inf_f \sup_{\theta \in \Theta} R(f, \theta).$$

Let the prior distribution of the parameter θ be given by the distribution function $G(\theta)$. The expected risk $r(f, G)$ is

$$r(f, G) = \mathrm{E}_\theta[R(f, \theta)] = \int_{\Theta} R(f, \theta) dG(\theta).$$

The estimator $f_G(x)$ which minimizes the function $r(f, G)$ for a given G is called a *Bayesian estimator* for G. The distribution G^0, for which

$$\inf_f r(f, G^0) = \sup_G \inf_f r(f, G)$$

holds, is defined to be the *least favourable distribution*.

In this paper we make use of the theorem which has been proved by Hodges and Lehmann [1].

THEOREM 1. *If there are a set Θ_1 of values θ and an estimator f^0 such that $R(f^0, \theta) = C$ for $\theta \in \Theta_1$, and $R(f^0, \theta) \leqslant C$ for $\theta \in \Theta - \Theta_1$, and if there is a distribution G^0 of the parameter θ on Θ_1 such that the estimator f^0 is Bayesian for G^0, then f^0 is a minimax estimator and G^0 is the least favourable distribution.*

2. Minimax estimation of the parameters of the multivariate hypergeometric distribution. In practice we often meet the following situation. A lot consisting of N units of product has been produced. The units are classified into k various categories. Let us assume that the category i contains U_i units ($i = 1, \ldots, k$). To estimate $U_1, \ldots, U_k$ a sample of size n is taken from the lot in which $m_1, \ldots, m_k$ units of categories $1, \ldots, k$ are observed. Let us suppose that the examination of a unit of the i-th category causes the cost d_i ($i = 1, \ldots, k$). We have such losses when, for example, a correct classification of the examined unit destroys it entirely, and d_i is the value of the unit of the i-th category. We are looking for minimax estimators of the parameters $U_1, \ldots, U_k$ on the basis of values of the sample $m_1, \ldots, m_k$. This leads to the estimation of the parameter $U = (U_1, \ldots, U_k)$ of the multivariate hypergeometric distribution. Thus

$$\text{(2.1)} \qquad \mathrm{P}(X_1 = m_1, \ldots, X_k = m_k) = \frac{\binom{U_1}{m_1} \cdots \binom{U_k}{m_k}}{\binom{N}{n}}.$$

Define the loss function by

$$\text{(2.2)} \qquad L(f, U) = \sum_{i=1}^{k} \{c_i[f_i(m_1, \ldots, m_k) - U_i]^2 + d_i m_i\},$$

where $f = (f_1, \ldots, f_k)$ is the estimator of the parameter $U = (U_1, \ldots, U_k)$. Let us suppose that $c_i > 0$ and $n < N$. In the case $N = n$ we know the contents of the population and $m_i = U_i$. We can determine the risk for $L(f, U)$ defined by formula (2.2):

$$R(f, U) = \mathrm{E}\{L[f(X), U] \mid U\}$$
$$= \sum_{m_1, \ldots, m_k} \sum_{i=1}^{k} \{c_i[f_i(m_1, \ldots, m_k) - U_i]^2 + d_i m_i\} \frac{\binom{U_1}{m_1} \cdots \binom{U_k}{m_k}}{\binom{N}{n}},$$

where $m_1 + \ldots + m_k = n$ and $m_1 \geqslant 0, \ldots, m_k \geqslant 0$.

THEOREM 2. *Let us assume that the random variable X has a probability density function of form* (2.1) *and that the loss function is defined by formula* (2.2). *Let*

$$g_i = \frac{n}{N^2}\left(\sqrt{n\frac{N-1}{N-n}}+1\right)^2 d_i. \tag{2.3}$$

If the constants c_i and g_i $(i = 1, \ldots, k)$ are ordered according to the formula

$$c_1+g_1 \geqslant c_2+g_2 \geqslant \ldots \geqslant c_k+g_k \tag{2.4}$$

and satisfy the conditions

$$c_1+c_2 > g_1-g_2, \quad c_i > 0 \ (i = 1, \ldots, k), \tag{2.5}$$

then the minimax estimator f^0 of the parameter U is of the form

$$f_i^0(X) = N\frac{X_i+2^{-1}(1-s_i)\sqrt{n(N-n)/(N-1)}}{n+\sqrt{n(N-n)/(N-1)}} \quad (i = 1, \ldots, k), \tag{2.6}$$

where

$$s_i = \begin{cases} \dfrac{L-2+\sum\limits_{j=1}^{L} g_j/c_j}{c_i\sum\limits_{j=1}^{L} 1/c_j} - \dfrac{g_i}{c_i} & (i = 1, \ldots, L), \\ 1 & (i = L+1, \ldots, k), \end{cases} \tag{2.7}$$

and $L \leqslant k$ is the greatest positive integer such that

$$(c_L+g_L)\sum_{j=1}^{L}\frac{1}{c_j} > L-2+\sum_{j=1}^{L}\frac{g_i}{c_i}. \tag{2.8}$$

The prior distribution defined by the formulae

$$P(U_{L+1} = U_{L+2} = \ldots = U_k = 0) = 1,$$

$$P(U_1 = u_1, \ldots, U_L = u_L) = K\frac{\Gamma(a_1+u_1)\ldots\Gamma(a_L+u_L)}{u_1!\ldots u_L!}, \tag{2.9}$$

where

$$a_i = \frac{1}{2}(1-s_i)\frac{N\sqrt{n(N-n)/(N-1)}}{N-n-\sqrt{n(N-n)/(N-1)}} \quad (i = 1, \ldots, L), \tag{2.10}$$

is the least favourable distribution.

Proof. The risk for the estimator $f^0 = (f_1^0, \ldots, f_k^0)$, where f_i^0 is expressed by formula (2.6), is of the form

$$(2.11)\qquad R(f^0, U) = \frac{N}{\left(1+\sqrt{n(N-1)/(N-n)}\right)^2}\left\{\frac{N}{4}\sum_{i=1}^{k} c_i(1-s_i)^2 + \right.$$
$$\left. + \sum_{i=1}^{k}\left[c_i s_i + \left(\sqrt{n\frac{N-1}{N-n}}+1\right)^2 \frac{n}{N^2} d_i\right]\right\}.$$

It is convenient to write

$$g_i = \left(\sqrt{n\frac{N-1}{N-n}}+1\right)^2 \frac{n}{N^2} d_i.$$

We can change the numeration of constants c_i and g_i in such a way that

$$c_1+g_1 \geqslant c_2+g_2 \geqslant \ldots \geqslant c_k+g_k.$$

If $L \leqslant k$ is defined by (2.8) (assumptions (2.4) and (2.5) guarantee that such an L exists and equals at least 2), then

$$(2.12)\qquad c_i+g_i > \frac{L-2+\sum_{j=1}^{L} g_j/c_j}{\sum_{j=1}^{L} 1/c_j} \qquad (i = 1, \ldots, L).$$

We show that

$$(2.13)\qquad c_i+g_i \leqslant \frac{L-2+\sum_{j=1}^{L} g_j/c_j}{\sum_{j=1}^{L} 1/c_j} \qquad (i = L+1, \ldots, k).$$

Because of (2.4) it is sufficient to prove inequality (2.13) for $i = L+1$. It follows from definition (2.8) of the number L that

$$(c_{L+1}+g_{L+1})\sum_{j=1}^{L+1}\frac{1}{c_j} \leqslant L-1+\sum_{j=1}^{L+1}\frac{g_j}{c_j}$$

or, equivalently,

$$(c_{L+1}+g_{L+1})\sum_{j=1}^{L}\frac{1}{c_j} \leqslant L-2+\sum_{j=1}^{L}\frac{g_j}{c_j},$$

which gives (2.13).

Now, we substitute (2.7) into formula (2.11). It follows from (2.12) that $s_i \leqslant 1$. We can observe that s_i are dependent only on $\{c_i\}$ and $\{g_i\}$. The risk $R(f^0, U)$ takes then the form

$$R(f^0, U) = \frac{N}{\left(\sqrt{n(N-1)/(N-n)}+1\right)^2}\left\{N\frac{L-2+\sum_{j=1}^{L} g_j/c_j}{\sum_{j=1}^{L} 1/c_j}+\right.$$

$$+\frac{N}{4}\sum_{i=1}^{L}\frac{1}{c_i}\left(c_i+g_i-\frac{L-2+\sum_{j=1}^{L} g_j/c_j}{\sum_{j=1}^{L} 1/c_j}\right)^2+$$

$$\left.+\sum_{i=L+1}^{k}\left(c_i+g_i-\frac{L-2+\sum_{j=1}^{L} g_j/c_j}{\sum_{j=1}^{L} 1/c_j}\right)U_i\right\}.$$

Let us notice that, for $U_{L+1} = U_{L+2} = \ldots = U_k = 0$, the risk

$$R(f^0, U) = \frac{N^2}{\left(\sqrt{n(N-1)/(N-n)}+1\right)^2}\left\{\frac{L-2+\sum_{j=1}^{L} g_j/c_j}{\sum_{j=1}^{L} 1/c_j}+\right.$$

$$\left.+\frac{1}{4}\sum_{i=1}^{L}\frac{1}{c_i}\left(c_i+g_i-\frac{L-2+\sum_{j=1}^{L} g_j/c_j}{\sum_{j=1}^{L} 1/c_j}\right)^2\right\} = C$$

is a constant and, because of (2.13), $R(f^0, U) \leqslant C$ for any set of non-negative integers $U_1, \ldots, U_k$ such that

$$\sum_{i=1}^{k} U_i = N.$$

We prove now that f^0 is a Bayesian estimator for the distribution defined by (2.9) and (2.10). The expected risk takes the form

$$r(f, p) = K \sum_{u_1,\ldots,u_L}\sum_{m_1,\ldots,m_L}\left\{\sum_{i=1}^{L} c_i[f_i(m_1, \ldots, m_L, 0, \ldots, 0)-u_i]^2+\right.$$

$$+\sum_{i=L+1}^{k} c_i f_i^2(m_1, \ldots, m_L, 0, \ldots, 0)+$$

$$\left.+\sum_{i=1}^{L} g_i m_i\right\}\frac{\binom{u_1}{m_1}\cdots\binom{u_L}{m_L}\Gamma(a_1+u_1)\ldots\Gamma(a_L+u_L)}{\binom{N}{n}u_1!\ldots u_L!},$$

where $u_1+\ldots+u_L = N$, $u_1 \geqslant m_1, \ldots, u_L \geqslant m_L$, and $m_1+\ldots+m_L = n$, $m_1 \geqslant 0, \ldots, m_L \geqslant 0$.

The expected risk is a positively determined quadratic form of the variables $f_i(m_1, \ldots, m_L, 0, \ldots, 0)$. In order to find its minimum it is sufficient to solve the system of equations

$$\frac{\partial r(f, p)}{\partial f_i(m_1, \ldots, m_L, 0, \ldots, 0)} = 0,$$

$$m_1 \geqslant 0, \ldots, m_L \geqslant 0, \sum_{j=1}^{L} m_j = n \quad (i = 1, \ldots, k).$$

The Bayesian estimator $\bar{f}$ for distribution (2.9) is then of the form

$$(2.14)\qquad \bar{f}_i(m_1, \ldots, m_L, 0, \ldots, 0) = \frac{\displaystyle\sum_{u_1,\ldots,u_L} u_i \frac{\Gamma(a_1+u_1)\ldots\Gamma(a_L+u_L)}{(u_1-m_1)!\ldots(u_L-m_L)!}}{\displaystyle\sum_{u_1,\ldots,u_L} \frac{\Gamma(a_1+u_1)\ldots\Gamma(a_1+u_L)}{(u_1-m_1)!\ldots(u_L-m_L)!}} \quad (i = 1, \ldots, L),$$

where $u_1+\ldots+u_L = N$, $u_1 \geqslant m_1, \ldots, u_L \geqslant m_L$, and

$$(2.15)\qquad \bar{f}_i(m_1, \ldots, m_L, 0, \ldots, 0) = 0 \quad (i = L+1, \ldots, k).$$

Let us notice that

$$\sum_{v_1,\ldots,v_L} \frac{(N-n)!}{v_1!\ldots v_L!} \frac{\Gamma(b_1+v_1)\ldots\Gamma(b_L+v_L)}{\Gamma\big(N-n+\sum_{j=1} b_j\big)}$$

$$= \int_{p_1,\ldots,p_L}\!\!\!\ldots\int p_1^{b_1-1}\ldots p_L^{b_1-1}\Bigg(\sum_{\substack{v_1+\ldots+v_L=N-n\\ v_1\geqslant 0,\ldots,v_L\geqslant 0}} \frac{(N-n)!}{v_1!\ldots v_L!} p_1^{v_1}\ldots p_L^{v_L}\Bigg) dp_1\ldots dp_L$$

$$= \int_{p_1,\ldots,p_L}\!\!\!\ldots\int p_1^{b_1-1}\ldots p_L^{b_L-1} dp_1\ldots dp_L = \frac{\Gamma(b_1)\ldots\Gamma(b_L)}{\Gamma\big(\sum_{j=1}^{L} b_j\big)},$$

where $v_1+\ldots+v_L = N-n$, $v_1 \geqslant 0, \ldots, v_L \geqslant 0$, and $p_1+\ldots+p_L = 1$, $p_1 \geqslant 0, \ldots, p_L \geqslant 0$.

Substituting $v_i = u_i - m_i$ into formula (2.14) and using the above-mentioned identity we obtain (see [3])

$$\bar{f}_i(m_1, \ldots, m_L, 0, \ldots, 0)$$

$$= \frac{\displaystyle\sum_{u_1,\ldots,u_L} \frac{\Gamma(a_1+u_1)\ldots\Gamma(a_i+u_i+1)\ldots\Gamma(a_L+u_L)}{(u_1-m_1)!\ldots(u_L-m_L)!}}{\displaystyle\sum_{u_1,\ldots,u_L} \frac{\Gamma(a_1+u_1)\ldots\Gamma(a_L+m_L)}{(u_1-m_1)!\ldots(u_L-m_L)!}} - a_i$$

$$= \frac{\displaystyle\sum_{v_1,\ldots,v_L} \frac{(N-n)!}{v_1!\ldots v_L!} \frac{\Gamma(v_1+a_1+m_1)\ldots\Gamma(v_i+a_i+m_i+1)\ldots\Gamma(v_L+a_L+m_L)}{\Gamma\big(N+\sum_{j=1}^{L} a_j+1\big)}}{\displaystyle\sum_{v_1,\ldots,v_L} \frac{(N-n)!}{v_1!\ldots v_L!} \frac{\Gamma(v_1+a_1+m_1)\ldots\Gamma(v_L+a_L+m_L)}{\Gamma\big(N+\sum_{j=1}^{L} a_j\big)}} \times$$

$$\times\Big(N+\sum_{j=1}^{L} a_j\Big) - a_i$$

$$= \frac{\dfrac{\Gamma(a_1+m_1)\ldots\Gamma(a_i+m_i+1)\ldots\Gamma(a_L+m_L)}{\Gamma\big(\sum_{j=1}^{L} a_j+n+1\big)}}{\dfrac{\Gamma(a_1+m_1)\ldots\Gamma(a_L+m_L)}{\Gamma\big(\sum_{j=1}^{L} a_j+n\big)}} \Big(N+\sum_{j=1}^{L} a_j\Big) - a_j.$$

$$= \frac{a_i+m_i}{n+\sum_{j=1}^{L} a_j} \Big(N+\sum_{j=1}^{L} a_j\Big) - a_i,$$

where $u_1+\ldots+u_L = N$, $u_1 \geqslant m_1, \ldots, u_L \geqslant m_L$, and $v_1+\ldots+v_L = N-n$, $v_1 \geqslant 0, \ldots, v_L \geqslant 0$.

Thus

$$\bar{f}_i(m_1, \ldots, m_L, 0, \ldots, 0) = \frac{\big(N+\sum_{j=1}^{L} a_j\big)m_i + (N-n)a_i}{n+\sum_{j=1}^{L} a_j} \quad (i = 1, \ldots, L).$$

For a_i defined by (2.10) we have

$$\bar{f}_i(m_1, \ldots, m_L, 0, \ldots, 0) = f_i^0(m_1, \ldots, m_L, 0, \ldots, 0) \quad (i = 1, \ldots, L). \tag{2.16}$$

Besides, from equations (2.15) and the definition of the estimator f^0 it follows that formula (2.16) holds also for $i = L+1, \ldots, k$. The estimator $f^0 = (f_1^0, \ldots, f_k^0)$, defined by formulae (2.6), (2.7) and (2.8), is a minimax estimator for $N > n+1$, which ensures that $a_i > 0$ and, therefore, the prior distribution defined by (2.9) and (2.10) exists. For this distribution, f^0 is a Bayesian estimator. If $N = n+1$, then f^0 is a Bayesian estimator for the prior distribution defined by

$$\mathrm{P}(U_{L+1} = \ldots = U_k = 0) = 1,$$

$$\mathrm{P}(U_1 = u_1, \ldots, U_L = u_L) = \frac{N}{u_1! \ldots u_L!} p_1^{u_1} \ldots p_L^{u_L},$$

where $p_i = \frac{1}{2}(1-s_i)$ for $i = 1, \ldots, L$. In this case the expected risk takes the form

$$r(f, p) = \sum_{u_1,\ldots,u_L} \sum_{m_1,\ldots,m_L} \Big\{ \sum_{i=1}^{L} c_i [f_i(m_1, \ldots, m_L, 0, \ldots, 0) - u_i]^2 + \\ + \sum_{i=L+1}^{k} c_i f_i^2(m_1, \ldots, m_L, 0, \ldots, 0) + \\ + \frac{1}{n} \sum_{i=1}^{L} g_i m_i \Big\} \frac{\binom{u_1}{m_1} \ldots \binom{u_L}{m_L}}{\binom{N}{n}} \frac{N!}{u_1! \ldots u_L!} p_1^{u_1} \ldots p_L^{u_L},$$

where $u_1 + \ldots + u_L = N$, $u_1 \geqslant m_1, \ldots, u_L \geqslant m_L$, and $m_1 + \ldots + m_L = n$, $m_1 \geqslant 0, \ldots, m_L \geqslant 0$. The Bayesian estimator for $i = 1, \ldots, L$ is then defined by the formula

$$\bar{f}_i(m_1, \ldots, m_L, 0, \ldots, 0) \\ = \frac{\sum\limits_{u_1,\ldots,u_L} u_i \dfrac{(N-n)!}{(u_1-m_1)! \ldots (u_L-m_L)!} p_1^{u_1} \ldots p_L^{u_L}}{\sum\limits_{u_1,\ldots,u_L} \dfrac{(N-n)!}{(u_1-m_1)! \ldots (u_L-m_L)!} p_1^{u_1} \ldots p_L^{u_L}} = m_i + p_i,$$

where $u_1 + \ldots + u_L = N$, $u_1 \geqslant m_1, \ldots, u_L \geqslant m_L$, and for $i = L+1, \ldots, k$ by

$$\bar{f}_i(m_1, \ldots, m_L, 0, \ldots, 0) = 0.$$

Thus for $N = n+1$ we have

$$\bar{f}_i(m_1, \ldots, m_L, 0, \ldots, 0) = f_i^0(m_1, \ldots, m_L, 0, \ldots, 0) \quad (i = 1, \ldots, k).$$

Since

$$p_i > 0 \ (i = 1, \ldots, L), \quad \sum_{i=1}^{L} p_i = 1,$$

such a prior distribution exists. We can easily verify that the estimator f^0 satisfies the condition

$$\sum_{i=1}^{k} f_i^0 = N.$$

3. Minimax estimation of the parameters of the multinomial distribution. As $N \to \infty$, the distribution of the random vector X is convergent to the multinomial distribution with the probability function

$$\mathrm{P}(X_1 = m_1, \ldots, X_k = m_k) = \frac{n!}{m_1! \ldots m_k!} p_1^{m_1} \ldots p_k^{m_k}, \qquad \sum_{i=1}^{k} m_i = n.$$

Let us assume as before that the cost of the observation of the i-th category unit is d_i $(i = 1, \ldots, k)$.

We consider the problem of minimax estimation of the parameter $p = (p_1, \ldots, p_k)$ of the multinomial distribution for the loss function

$$L(f, p) = \sum_{i=1}^{k} \{c_i [f_i(m_1, \ldots, m_k) - p_i]^2 + d_i m_i\}, \tag{3.1}$$

where $f = (f_1, \ldots, f_k)$ is the estimator of the parameter p. Let us suppose that $c_i > 0$ for $i = 1, \ldots, k$. We can determine the risk as

$$R(f, p) = \mathrm{E}\{L(f, p) | p\}$$
$$= \sum_{m_1, \ldots, m_k} \sum_{i=1}^{k} \{c_i [f_i(m_1, \ldots, m_k) - p_i]^2 + d_i m_i\} \frac{n!}{m_1! \ldots m_k!} p_1^{m_1} \ldots p_k^{m_k},$$

where $m_1 + \ldots + m_k = n$, and $m_1 \geqslant 0, \ldots, m_k \geqslant 0$.

Let us write

$$g_i = n(\sqrt{n}+1)^2 d_i \qquad (i = 1, \ldots, k).$$

Without loss of generality we may assume that the sequence $\{c_i + g_i\}$ is non-increasing, i.e.,

$$c_1 + g_1 \geqslant c_2 + g_2 \geqslant \ldots \geqslant c_k + g_k. \tag{3.2}$$

Suppose also that

$$c_1 + c_2 > g_1 - g_2. \tag{3.3}$$

Let $L \leqslant k$ be the greatest positive integer such that

$$(c_L + g_L) \sum_{j=1}^{L} \frac{1}{c_j} > L - 2 + \sum_{j=1}^{L} \frac{g_j}{c_j}$$

(L equals at least 2). Then (cf. (2.12) and (2.13))

$$c_i + g_i > \frac{L - 2 + \sum_{j=1}^{L} g_j / c_j}{\sum_{j=1}^{L} 1/c_j} \qquad (i = 1, \ldots, L) \tag{3.4}$$

and

$$c_i + g_i \leqslant \frac{L-2+\sum_{j=1}^{L} g_j/c_j}{\sum_{j=1}^{L} 1/c_j} \qquad (i = L+1, \ldots, k). \tag{3.5}$$

THEOREM 3. *If the constants c_i and g_i ($i = 1, \ldots, k$) are ordered according to formula* (3.2) *and satisfy condition* (3.3), *then the estimator $f^0 = (f_1^0, \ldots, f_k^0)$ of the form*

$$f_i^0(X) = \frac{X_i + 2^{-1}(1-s_i)\sqrt{n}}{n+\sqrt{n}} \qquad (i = 1, \ldots, k), \tag{3.6}$$

where

$$s_i = \begin{cases} \dfrac{L-2+\sum_{j=1}^{L} g_j/c_j}{c_i \sum_{j=1}^{L} 1/c_j} - \dfrac{g_i}{c_i} & (i = 1, \ldots, L), \\ 1 & (i = L+1, \ldots, k), \end{cases} \tag{3.7}$$

is a minimax estimator of the parameter $p = (p_1, \ldots, p_k)$ of the multinomial distribution for the loss function determined by formula (3.1). *The prior distribution $G(p)$ of the parameter $p = (p_1, \ldots, p_k)$ is determined by the equations*

$$\begin{gathered} \mathrm{P}(p_{L+1} = p_{L+2} = \ldots = p_k = 0) = 1, \\ dG(p) = Kp_1^{r_1} \ldots p_L^{r_L} dp_1 \ldots dp_L, \end{gathered} \tag{3.8}$$

where

$$r_i = \frac{1}{2}(1-s_i)\sqrt{n} - 1 \qquad (i = 1, \ldots, L),$$

is the least favourable distribution.

Proof. Let us evaluate the risk $R(f^0, p)$ for f^0 determined by formulae (3.6) and (3.7):

$$R(f^0, p) = \frac{1}{(\sqrt{n}+1)^2}\left\{\frac{L-2+\sum_{j=1}^{L} g_j/c_j}{\sum_{j=1}^{L} 1/c_j} + \frac{1}{4}\sum_{i=1}^{L}\frac{1}{c_i}\left(c_i + g_i - \frac{L-2+\sum_{j=1}^{L} g_j/c_j}{\sum_{j=1}^{L} 1/c_j}\right)^2 + \right.$$
$$\left. + \sum_{i=L+1}^{k}\left(c_i + g_i - \frac{L-2+\sum_{j=1}^{L} g_j/c_j}{\sum_{j=1}^{L} 1/c_j}\right)p_i\right\}.$$

We can see that, for $p_{L+1} = p_{L+2} = \ldots = p_k = 0$,

$$R(f^0, p) =$$

$$= \frac{1}{(\sqrt{n}+1)^2}\left\{\frac{L-2+\sum_{j=1}^{L} g_j/c_j}{\sum_{j=1}^{L} 1/c_j} + \frac{1}{4}\sum_{i=1}^{L}\frac{1}{c_i}\left(c_i+g_i-\frac{L-2+\sum_{j=1}^{L} g_j/c_j}{\sum_{j=1}^{L} 1/c_j}\right)^2\right\} = C$$

and, on the basis of (3.5), $R(f^0, p) \leqslant C$ for each system of numbers $p_1, p_2, \ldots$ $\ldots, p_k$ such that

$$\sum_{i=1}^{k} p_i = 1, \quad p_i \geqslant 0 \quad (i = 1, \ldots, k).$$

In order to prove that f^0 is a minimax estimator it is sufficient to check that f^0 is a Bayesian estimator for the prior distribution determined by formula (3.8). That distribution exists, since condition (3.4) ensures that $s_i \leqslant 1$ $(i = 1, \ldots, L)$. Let us determine the expected risk for this distribution. We have

$$(3.9) \quad r(f, G) = K \int_{p_1,\ldots,p_L}\!\!\cdots\!\int \sum_{m_1,\ldots,m_L} \left\{\sum_{i=1}^{L} c_i[f_i(m_1, \ldots, m_L, 0, \ldots, 0)-p_i]^2 + \right.$$

$$+ \sum_{i=L+1}^{k} c_i f_i^2(m_1, \ldots, m_L, 0, \ldots, 0) +$$

$$\left. + \frac{1}{(\sqrt{n}+1)^2}\sum_{i=1}^{L} g_i m_i\right\}\frac{n!}{m_1! \ldots m_L!} p_1^{m_1+r_1} \ldots p_L^{m_L+r_L} dp_1 \ldots dp_L,$$

where $p_1 + \ldots + p_L = 1$, $p_1 \geqslant 0, \ldots, p_L \geqslant 0$, and $m_1 + \ldots + m_L = n$, $m_1 \geqslant 0, \ldots, m_L \geqslant 0$.

Since $r(f, G)$ is the positively determined quadratic form of the variables $f_i(m_1, \ldots, m_L, 0, \ldots, 0)$, the estimator $\bar{f}$ which minimizes $r(f, G)$ can be found from the system of equations

$$\frac{\partial r(f, G)}{\partial f_i(m_1, \ldots, m_L, 0, \ldots, 0)} = 0,$$

$$m_1 \geqslant 0, \ldots, m_L \geqslant 0, \quad \sum_{j=1}^{L} m_j = n \quad (i = 1, \ldots, k).$$

Differentiating the expression (3.9) we see that $r(f, G)$ attains its minimum for

(3.10)

$$\bar{f}_i(m_1, \ldots, m_L, 0, \ldots, 0) = \frac{\int\limits_{p_1,\ldots,p_L} \ldots \int p_1^{m_1+r_1} \ldots p_i^{m_i+r_i+1} \ldots p_L^{m_L+r_L} dp_1 \ldots dp_L}{\int\limits_{p_1,\ldots,p_L} \ldots \int p_1^{m_1+r_1} \ldots p_L^{m_L+r_L} dp_1 \ldots dp_L}$$

$$= \frac{\dfrac{\Gamma(m_1+r_1+1) \ldots \Gamma(m_i+r_i+2) \ldots \Gamma(m_L+r_L+1)}{\Gamma\left(n+\sum_{j=1}^{L} r_j+L+1\right)}}{\dfrac{\Gamma(m_1+r_1+1) \ldots \Gamma(m_i+r_i+1) \ldots \Gamma(m_L+r_L+1)}{\Gamma\left(n+\sum_{j=1}^{L} r_j+L\right)}}$$

$$= \frac{m_i+r_i+1}{n+\sum_{j=1}^{L} r_j+L} \quad (i = 1, \ldots, L),$$

where $p_1 + \ldots + p_L = 1$, $p_1 \geqslant 0, \ldots, p_L \geqslant 0$, and

$$\bar{f}_i(m_1, \ldots, m_L, 0, \ldots, 0) = 0 \quad (i = L+1, \ldots, k).$$

Substituting

$$r_i = \frac{1}{2}(1-s_i)\sqrt{n}-1 \quad (i = 1, \ldots, L),$$

into formula (3.10) we obtain

$$\bar{f}(m_1, \ldots, m_L, 0, \ldots, 0) = f^0(m_1, \ldots, m_L, 0, \ldots, 0).$$

We have demonstrated that there exists the prior distribution for which f^0 is a Bayesian estimator. This completes the proof.

It is easy to verify that f^0 satisfies the condition

$$\sum_{i=1}^{k} f_i^0 = 1.$$

4. Remarks. The problem of minimax estimation of parameters of the multivariate hypergeometric and multinomial distributions was previously studied.

Hodges and Lehmann [1] obtained the minimax estimators of the parameters of the binomial and hypergeometric distributions when the loss function is the squared error of estimation.

Steinhaus [2] solved the problem of minimax estimation of the parameter p of the multinomial distribution for the loss

$$L(f, p) = \sum_{i=1}^{k} (f_i - p_i)^2.$$

Generalizations of this result are comprised in the papers by Trybuła ([3] and [4]) who has examined the problem of minimax estimation of the parameters of the multinomial and multivariate hypergeometric distributions without taking into account the cost of observation.

References

[1] J. L. Hodges and E. L. Lehmann, *Some problems in minimax point estimation*, Ann. Math. Statist. 21 (1950), p. 182-197.

[2] H. Steinhaus, *The problem of estimation*, ibidem 28 (1957), p. 633-648.

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

[4] — *O minimaksowej estymacji parametrów w rozkładzie wielomianowym*, Zastosow. Matem. 3 (1958), p. 307-322.

INSTITUTE OF MATHEMATICS
TECHNICAL UNIVERSITY
50-370 WROCŁAW

Received on 2. 2. 1976

M. RUTKOWSKA (Wrocław)

MINIMAKSOWA ESTYMACJA PARAMETRÓW ROZKŁADU HIPERGEOMETRYCZNEGO WIELOWYMIAROWEGO I WIELOMIANOWEGO

STRESZCZENIE

W pracy rozpatrzono problem minimaksowej estymacji parametrów $U_1, \ldots, U_k$ rozkładu hipergeometrycznego wielowymiarowego dla funkcji straty, będącej sumą błędów kwadratowych estymacji i kosztów pobierania próby. Udowodniono następujące twierdzenie:

Załóżmy, że zmienna losowa X ma rozkład określony wzorem (2.1) *i że funkcja straty jest postaci* (2.2), *gdzie $f = (f_1, \ldots, f_k)$ jest estymatorem parametru $U = (U_1, \ldots, U_k)$, a d_i oznacza koszt obserwacji X_i. Niech g_i będzie określone wzorem* (2.3). *Jeżeli stałe c_i oraz g_i $(i = 1, 2, \ldots, k)$ są uporządkowane zgodnie z wzorem* (2.4) *i spełniają warunki* (2.5), *to minimaksowym estymatorem parametru U jest estymator $f^0 = (f_1^0, \ldots, f_k^0)$ postaci* (2.6), *gdzie s_i określone jest wzorem* (2.7), *a $L \leqslant k$ jest największą liczbą naturalną spełniającą* (2.8).

Analogiczne twierdzenie udowodniono dla rozkładu wielomianowego.
