ZASTOSOWANIA MATEMATYKI
APPLICATIONES MATHEMATICAE
21,2 (1991), pp. 153–159

S. TRYBUŁA (Wrocław)

SOME PROBLEMS OF PREDICTION OF THE DIFFERENCE OF RANDOM VARIABLES

The problem of minimax prediction of the difference of binomial random variables is solved for a quadratic loss function. Also, the problems of minimax prediction of $Y_i - Y_j$, $i \neq j$, for the multinomial and multivariate hypergeometric random variable $(Y_1, \ldots, Y_r)$ are solved.

1. Let the random variables X_1 and Y_1 have binomial distributions with parameters m and p_1, p_2, respectively, and let X_2 and Y_2 have binomial distributions with parameters n and p_1, p_2. Assume that $X_1, Y_1; X_2, Y_2$ are independent. The problem is to determine a minimax predictor d of $X_2 - Y_2$ based on X_1, Y_1 for the quadratic loss function

$$L(X_2 - Y_2, d) = [d(X_1, Y_1) - X_2 + Y_2]^2 .$$

Let us try a predictor of the form

$$d(X_1, Y_1) = a(X_1 - Y_1) . \tag{1}$$

For this d the risk is

$$R(p_1, p_2, d) = E[L(X_2 - Y_2, d(X_1, Y_1))] = E[a(X_1 - Y_1) - X_2 + Y_2]^2$$
$$= (a^2 m + n)[p_1(1 - p_1) + p_2(1 - p_2)] + (n - am)^2 (p_1 - p_2)^2 .$$

Assume that

$$a^2 m + n = 2(n - am)^2 .$$

This equation surely holds if

$$a = \frac{\frac{2n-1}{m}}{2 + \sqrt{\frac{2}{m} + \frac{2}{n} - \frac{1}{mn}}} \tag{2}$$

1991 *Mathematics Subject Classification*: Primary 62F15.

Key words and phrases: minimax prediction, multinomial distribution, multivariate hypergeometric distribution.

and for this a the risk is

$$R(p_1,p_2,d) = (n-am)^2[-(p_1+p_2)^2 + 2(p_1+p_2)].$$

The function $R(p_1,p_2,d)$ attains its maximum when

$$p_1 + p_2 = 1. \tag{3}$$

On the other hand, for any predictor d

$$\begin{aligned} R(p_1,p_2,d) &= E\{d(X_1,Y_1) - X_2 + Y_2\}^2 \\ &= E\{d(X_1,Y_1) - n(p_1-p_2)\}^2 + n[p_1(1-p_1) + p_2(1-p_2)]. \end{aligned} \tag{4}$$

Notice that the expression in square brackets is independent of d.

Let p_1, p_2 satisfy (3) and let p_1 have the beta density

$$g(p_1) = \begin{cases} \dfrac{1}{B(\alpha,\beta)} p_1^{\alpha-1}(1-p_1)^{\beta-1} & \text{if } 0 < p_1 < 1, \\ 0 & \text{otherwise}. \end{cases}$$

For this a priori distribution the expected risk attains its minimum if

$$\begin{aligned} d(X_1,Y_1) &= nE(p_1 - p_2 \mid X_1,Y_1) \\ &= nE(2p_1 - 1 \mid X_1,Y_1) = n\left(2\frac{X_1 - Y_1 + m + \alpha}{2m+\alpha+\beta} - 1\right) \end{aligned}$$

and is equal to the predictor defined by (1) and (2) if

$$\alpha = \beta = \frac{m}{2n-1}\left(1 + n\sqrt{\frac{2}{m} + \frac{2}{n} - \frac{1}{mn}}\right).$$

Thus *the predictor $d(X_1,Y_1)$ defined by* (1) *and* (2) *is minimax.*

2. Let the random variables $X = (X_1,\ldots,X_r)$ and $Y = (Y_1,\ldots,Y_r)$ be distributed as follows:

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

$$P(Y = y) = P(Y_1 = y_1,\ldots,Y_r = y_r) = \frac{n!}{y_1!\ldots y_r!} p_1^{y_1}\ldots p_r^{y_r}.$$

The problem is to determine a minimax predictor d of $Y_1 - Y_2$ on the basis of X for the loss function

$$L(Y_1 - Y_2, d) = [d(X) - Y_1 + Y_2]^2.$$

For the predictor

$$d(X) = a(X_1 - X_2) \tag{5}$$

the risk is

$$\begin{aligned}R(p,d) &= E[a(X_1-X_2)-Y_1+Y_2]^2\\ &= (a^2m+n)[p_1+p_2-(p_1-p_2)^2]+(n-am)^2(p_1-p_2)^2\\ &= (a^2m+n)(p_1+p_2)\end{aligned}$$

provided $a^2m+n=(n-am)^2$. This equation is surely satisfied if

$$a=\frac{\frac{n-1}{m}}{1+\sqrt{\frac{1}{m}+\frac{1}{n}-\frac{1}{mn}}}. \tag{6}$$

When $n=1$ we obtain $a=0$.

This $R(p,d)$ attains its maximum if $p_1+p_2=1$.

Let us look for the least favourable distribution among the following:

(a) $p_1+p_2=1$,
(b) p_1 has a beta distribution with parameters α, β.

It follows from (a) that $X_1+X_2=m$, and from (b) that the following is the Bayes predictor:

$$\begin{aligned}d(X) &= nE(p_1-p_2\mid X_1,X_2)=nE(2p_1-1|X_1)\\ &= n\left(2\frac{X_1+\alpha}{m+\alpha+\beta}-1\right)=a(2X_1-m)\end{aligned}$$

provided

$$\frac{n}{m+\alpha+\beta}=a,\quad n\left(\frac{2\alpha}{m+\alpha+\beta}-1\right)=-am.$$

Solving this system of equations for a given by (6) we obtain

$$\alpha=\beta=\frac{m}{2(n-1)}\left(1+n\sqrt{\frac{1}{m}+\frac{1}{n}-\frac{1}{mn}}\right)$$

when $n>1$.

When $n=1$, $a=0$. In this case $d(X)\equiv 0$ is a minimax predictor. It is Bayes with respect to the a priori distribution

$$P(p_1=p_2=\tfrac{1}{2})=1.$$

Thus in both cases *the predictor* (5) *with* a *given by* (6) *is minimax.*

3. Consider the following problem. A lot consisting of N units of a product is given. The units are classified into r categories, the ith category containing M_i units ($i=1,\ldots,r$). Two samples of size m and n respectively are taken from the lot one after another; $x_1,\ldots,x_r$, respectively $y_1,\ldots,y_r$

units of categories $1, \ldots, r$ are observed in the samples. Then we have

$$P(X = x) = P(X_1 = x_1, \ldots, X_r = x_r) = \frac{\binom{M_1}{x_1} \ldots \binom{M_r}{x_r}}{\binom{N}{m}},$$

$$P(Y = y \mid X = x) = P(Y_1 = y_1, \ldots, Y_r = y_r \mid X_1 = x_1, \ldots, X_r = x_r)$$
$$= \frac{\binom{M_1 - x_1}{y_1} \ldots \binom{M_r - x_r}{y_r}}{\binom{M-m}{n}}.$$

The problem is to determine a minimax predictor $d(X)$ of the difference $Y_1 - Y_2$ for a quadratic loss function.

We take once again a predictor of the form

$$d(X) = a(X_1 - X_2).$$

The risk is then

$$\begin{aligned}
&R(m, d) \\
&= E[a(X_1 - X_2) - Y_1 + Y_2]^2 \\
&= E\{E\{[a(X_1 - X_2) - Y_1 + Y_2]^2 \mid X\}\} \\
&= E\Big\{E\Big\{\Big[a(X_1 - X_2) - n\frac{M_1 - X_1}{N - m} + n\frac{M_2 - X_2}{N - m} \\
&\quad - \Big(Y_1 - n\frac{M_1 - X_1}{N - m}\Big) + \Big(Y_2 - n\frac{M_2 - X_2}{N - m}\Big)\Big]^2 \Big| X\Big\}\Big\} \\
&= E\Big[a(X_1 - X_2) - n\Big(\frac{M_1 - X_1}{N - m} - \frac{M_2 - X_2}{N - m}\Big)\Big]^2 \\
&\quad + E\Big\{E\Big[\Big(Y_1 - n\frac{M_1 - X_1}{N - m}\Big)^2 \Big| X\Big]\Big\} + E\Big\{E\Big[\Big(Y_2 - n\frac{M_2 - X_2}{N - m}\Big)^2 \Big| X\Big]\Big\} \\
&\quad - 2E\Big\{E\Big[\Big(Y_1 - n\frac{M_1 - X_1}{N - m}\Big)\Big(Y_2 - n\frac{M_2 - X_2}{N - m}\Big)\Big| X\Big]\Big\} \\
&= E\Big[\Big(a + \frac{n}{N - m}\Big)(X_1 - X_2) - \frac{n}{N - m}(M_1 - M_2)\Big]^2 \\
&\quad + n\frac{N - m - n}{N - m - 1}\Big[E\Big(\frac{M_1 - X_1}{N - m}\frac{N - M_1 - m + X_1}{N - m}\Big) \\
&\quad + E\Big(\frac{M_2 - X_2}{N - m}\frac{N - M_2 - m + X_2}{N - m}\Big) + 2E\Big(\frac{M_1 - X_1}{N - m}\frac{M_2 - X_2}{N - m}\Big)\Big] \\
&= \Big(a + \frac{n}{N - m}\Big)^2 \Big[- m\frac{N - m}{N - 1}\frac{1}{N^2}(M_1 - M_2)^2 + m\frac{N - m}{N - 1}\frac{1}{N}(M_1 + M_2)\Big] \\
&\quad + \Big(\frac{m}{N}a - \frac{n}{N}\Big)^2 (M_1 - M_2)^2
\end{aligned}$$

$$+\frac{n(N-m-n)}{(N-m-1)(N-m)^2}\left[(N-m)\frac{N-m-1}{N-1}(M_1+M_2)\right.$$
$$\left.+\left(-1+\frac{2m}{N}-\frac{m(m-1)}{N(N-1)}\right)(M_1-M_2)^2\right]$$

if $N > m+1$, $M = (M_1, \dots, M_r)$.

Assume that the coefficient of $(M_1 - M_2)^2$ in the above expression vanishes:

$$-m\frac{N-m}{N-1}\frac{1}{N^2}\left(a+\frac{n}{N-m}\right)^2+\left(\frac{m}{N}a-\frac{n}{N}\right)^2$$
$$+\left(-1+\frac{2m}{N}-\frac{m(m-1)}{N(N-1)}\right)\frac{n(N-m-n)}{(N-m-1)(N-m)^2}=0,$$

or after expanding the expressions in the first two brackets and reduction

$$\frac{m(m-1)a^2-2mna+n(n-1)}{N(N-1)}=0.$$

The solution a of this equation does not depend on N:

$$a=\frac{\frac{n-1}{m}}{1+\sqrt{\frac{1}{m}+\frac{1}{n}-\frac{1}{mn}}}. \tag{7}$$

Thus a is the same as for the multinomial case as one can expect from the above remark.

Notice that $a = 0$ when $n = 1$.

On the other hand, we obtain

$$\begin{aligned}R(M,d) &= E[d(X)-Y_1+Y_2]^2 = E\{E[(d(X)-Y_1+Y_2)^2 \mid X]\}\\
&= E\left\{E\left[\left(d(X)-n\frac{M_1-X_1}{N-m}+n\frac{M_2-X_2}{N-m}\right.\right.\right.\\
&\qquad \left.\left.\left.-\left(Y_1-n\frac{M_1-X_1}{N-m}\right)+\left(Y_2-n\frac{M_2-X_2}{N-m}\right)\right)^2\middle|X\right]\right\}\\
&= E\left[d(X)-\frac{n}{N-m}(M_1-X_1-M_2+X_2)\right]^2+R_0(M)\\
&= E\left[d_0(X)-\frac{n}{N-m}(M_1-M_2)\right]^2+R_0(M),\end{aligned}$$

where R_0 is independent of d and

$$d_0(X)=d(X)+\frac{n}{N-m}(X_1-X_2). \tag{8}$$

Let an a priori distribution of the parameter M be defined as follows:

$$M_1 + M_2 = N, \quad P(M_1 = u) = \binom{N}{u} \frac{B(u+\alpha, N-u+\beta)}{B(\alpha,\beta)}.$$

The expected risk attains its minimum if

$$d_0(X) = \frac{n}{N-m} E(M_1 - M_2 \mid X) = \frac{n}{N-m} E(2M_1 - N \mid X)$$
$$= \frac{n}{N-m}\left[2\frac{(N+\alpha+\beta)X_1 + (N-m)\alpha}{m+\alpha+\beta} - N\right]$$

(see [1]). By (8) and $X_1 + X_2 = m$ we obtain

$$d(X) = n\left(\frac{2X_1}{m+\alpha+\beta} - \frac{m-\alpha+\beta}{m+\alpha+\beta}\right) = a(X_1 - X_2) = a(2X_1 - m)$$

provided

$$\frac{n}{m+\alpha+\beta} = a, \quad n\frac{m-\alpha+\beta}{m+\alpha+\beta} = am,$$

where a is given by (7). This leads to

$$\alpha = \beta = \frac{m}{2(n-1)}\left(1 + n\sqrt{\frac{1}{m} + \frac{1}{n} - \frac{1}{mn}}\right)$$

if $n > 1$. Thus for $n > 1$

$$d(X) = \frac{\frac{n-1}{m}(X_1 - X_2)}{1 + \sqrt{\frac{1}{m} + \frac{1}{n} - \frac{1}{mn}}} \tag{9}$$

is a minimax predictor of $Y_1 - Y_2$.

We conjecture that for $n = 1$ a minimax predictor does not exist.

For the predictor (9),

$$R(M,d) = \left[\left(a + \frac{n}{N-m}\right)^2 m\frac{N-m}{N-1}\frac{1}{N} + n\frac{N-m-n}{N-m}\frac{1}{N-1}\right](M_1 + M_2).$$

For related problems of minimax estimation and prediction see [1]–[5].

References

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

[2] S. Trybuła, *Some problems of simultaneous minimax estimation*, ibid. 29 (1958), 245–253.

[3] —, *Some investigations in minimax estimation theory*, Dissertationes Math. 240 (1985).

[4] —, *Minimax estimation and prediction for random variables with bounded sum*, Zastos. Mat. 19 (1987), 85–91.

[5] M. Wilczyński, *Minimax estimation for multinomial and multivariate hypergeometric distribution*, Sankhyā Ser. A 47 (1985), 128–132.

STANISŁAW TRYBUŁA
INSTITUTE OF MATHEMATICS
TECHNICAL UNIVERSITY OF WROCŁAW
WYBRZEŻE WYSPIAŃSKIEGO 27
50-370 WROCŁAW, POLAND

Received on 15.9.1989