ZASTOSOWANIA MATEMATYKI
APPLICATIONES MATHEMATICAE
21,4 (1993), pp. 555–560

S. TRYBUŁA (Wrocław)

MINIMAX ESTIMATION OF A CUMULATIVE DISTRIBUTION FUNCTION FOR A SPECIAL LOSS FUNCTION

A minimax estimator of a cumulative distribution function is determined for the loss function (1). The problem of minimax prediction of a sample distribution function is also solved for a similar loss function.

1. Minimax estimation of a cumulative distribution function. Suppose that a random variable X is distributed according to an unknown cumulative distribution function $F(t)$. Let $\hat{X} = (X_1, \ldots, X_n)$ be a random sample from F, $X_1, \ldots, X_n$ being independent. Let $\varphi(t) = \varphi(t, \hat{X})$ be an estimator of $F(t)$. We suppose that the loss function associated with the estimator $\varphi(t)$ is

$$L(F, \varphi) = \int_{-\infty}^{\infty} \frac{(\varphi(t) - F(t))^2}{F(t)(1 - F(t)) + c^2}\, w(dt)\,, \tag{1}$$

where w is a non-zero finite measure on $(\mathbb{R}, \mathcal{B})$, $\mathcal{B}$ being the σ-field of Borel subsets of $\mathbb{R} = (-\infty, \infty)$.

The problem is to determine a minimax estimator of $F(t)$ for this loss function.

Set

$$\hat{F}(t) = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}(t)\,,$$

where, for a random variable Z,

$$\delta_Z(t) = \begin{cases} 1 & \text{if } Z \le t, \\ 0 & \text{if } Z > t. \end{cases}$$

1991 *Mathematics Subject Classification*: Primary 62F15.

Key words and phrases: minimax estimation, minimax prediction, cumulative distribution function, sample distribution function.

Let us study an estimator of the form

$$\varphi(t) = a\hat{F}(t) + b\,.$$

The risk function for this estimator is

$$\begin{aligned} R(F,\varphi) &= E[L(F,\varphi(t,\hat{X}))] \\ &= \int_{-\infty}^{\infty} \frac{E[a\hat{F}(t)+b-F(t)]^2}{F(t)(1-F(t))+c^2}\,w(dt) \\ &= \int_{-\infty}^{\infty} \frac{\frac{a^2}{n}F(t)(1-F(t)) + (b-(1-a)F(t))^2}{F(t)(1-F(t))+c^2}\,w(dt)\,. \end{aligned}$$

Let $b = (1-a)/2$. Then

$$\begin{aligned} R(F,\varphi) &= \int_{-\infty}^{\infty} \frac{\left(\frac{a^2}{n} - (1-a)^2\right)F(t)(1-F(t)) + \frac{(1-a)^2}{4}}{F(t)(1-F(t))+c^2}\,w(dt) \\ &= \frac{(1-a)^2}{4c^2}\int_{-\infty}^{\infty} w(dt) \stackrel{\mathrm{df}}{=} K \end{aligned} \tag{2}$$

if

$$a = \frac{1}{1+\frac{2c}{\sqrt{1+4c^2}}\frac{1}{\sqrt{n}}}\,, \tag{3}$$

i.e. if

$$\varphi(t) = \frac{\hat{F}(t) + \frac{c}{\sqrt{1+4c^2}}\frac{1}{\sqrt{n}}}{1+\frac{2c}{\sqrt{1+4c^2}}\frac{1}{\sqrt{n}}} \stackrel{\mathrm{df}}{=} \varphi_0(t)\,. \tag{4}$$

We shall prove that the estimator $\varphi_0(t)$ is minimax.

The considered problem of determining a minimax estimator of $F(t)$ can be viewed as the problem of finding the optimal strategy in a game against nature. The nature chooses a cumulative distribution function $F(t)$, the statistician chooses an estimator $\varphi(t)$ and the payoff function is given by the risk

$$R(F,\varphi) = \int_{-\infty}^{\infty} \frac{E(\varphi(t)-F(t))^2}{F(t)(1-F(t))+c^2}\,w(dt)\,. \tag{5}$$

Let us define a sequence τ_k of mixed strategies of nature which will be used in the proof of the optimality of the strategy of $\varphi_0(t)$.

Choose the parameter p according to the density

$$g(p) = \begin{cases} C[p(1-p)+c^2][p(1-p)]^{\alpha-1} & \text{if } 0 < p < 1, \\ 0 & \text{otherwise,} \end{cases} \tag{6}$$

where C is a normalizing constant, and then, for given p, choose the distribution $F(t)$ of the form

$$F(t) = \begin{cases} 0 & \text{if } t < -k, \\ p & \text{if } -k \le t < k, \\ 1 & \text{if } t \ge k. \end{cases} \tag{7}$$

Let $F(t)$ be given by (7), where p has distribution (6). For the strategy τ_k the expected risk is

$$r(\tau_k, \varphi) = \int_{-\infty}^{\infty} E_{\tau_k}\left[\frac{E(\varphi(t)-F(t))^2}{F(t)(1-F(t))+c^2}\right] w(dt),$$

where $E_{\tau_k}(\cdot)$ is the expectation with respect to the density $g(p)$.

In order to minimize the expected risk $r(\tau_k, \varphi)$, it is sufficient to minimize

$$E_{\tau_k}\left[\frac{E(\varphi(t)-F(t))^2}{F(t)(1-F(t))+c^2}\right]$$

for any fixed t. This leads to the Bayes estimator with respect to τ_k given by

$$\varphi_{\tau_k}(t) = \begin{cases} 0 & \text{if } t < -k, \\ \dfrac{\hat{F}(t)+\alpha/(2n)}{1+\alpha/n} & \text{if } -k \le t < k, \\ 1 & \text{if } t \ge k. \end{cases}$$

Let

$$\alpha = \frac{2c}{\sqrt{1+4c^2}}\sqrt{n}.$$

In this case $\varphi_{\tau_k}(t) = \varphi_0(t)$ if $-k \le t < k$, and, by (2),

$$r(\tau_k, \varphi_{\tau_k}) = \frac{(1-a)^2}{4c^2}\int_{-\infty}^{\infty} I_{[-k,k)}(t)\, w(dt),$$

where $I_A(t)$ is the characteristic function of the set A and the constant a is given by (3).

From the above it follows that

$$\lim_{k\to\infty} r(\tau_k, \varphi_{\tau_k}) = K, \tag{8}$$

where K is defined in (2).

From (8) and the fact that the estimator $\varphi_{\tau_k}(t)$ is Bayes with respect to τ_k and $\varphi_{\tau_k}(t) = \varphi_0(t)$ for $-k \le t < k$, it follows that the *estimator* $\varphi_0(t)$ *given by* (4) *is a minimax estimator of* $F(t)$.

If the measure w is concentrated at one point, say t_0, then the problem reduces to that of determining a minimax estimator of the parameter $p = F(t_0)$ for the loss function

$$L(p,a) = \frac{(a-p)^2}{p(1-p)+c^2}.$$

This problem was solved in [5].

2. Minimax prediction of a sample distribution function. Let $\hat{X} = (X_1, \ldots, X_n)$, $\hat{Y} = (Y_1, \ldots, Y_m)$ be two independent samples from a distribution $F(t)$ and let

$$\hat{F}(t) = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}(t), \qquad \check{F}(t) = \frac{1}{m}\sum_{i=1}^{m} \delta_{Y_i}(t)$$

be the sample distribution functions from the samples $\hat{X}$ and $\hat{Y}$, respectively. Let $\psi(t) = \psi(t, \hat{X})$ be a predictor of $F(t)$ and let

$$L(\check{F}, \psi) = \int_{-\infty}^{\infty} \frac{(\psi(t) - F(t))^2}{F(t)(1-F(t))+c^2}\, w(dt) \tag{9}$$

be the loss function connected with the predictor ψ, where, as before, $w(\cdot)$ is a non-zero finite measure on $(\mathbb{R}, \mathcal{B})$.

The problem is to determine a minimax predictor of $\check{F}(t)$.

For the loss function (9) the risk function takes the form

$$\begin{aligned} R(F,\psi) &= E(L(F, \psi(t,\hat{X}))) \\ &= \int_{-\infty}^{\infty} \frac{E(\psi(t)-F(t))^2}{F(t)(1-F(t))+c^2}\, w(dt) \\ &= \int_{-\infty}^{\infty} \frac{E(\psi(t)-F(t))^2 + \frac{F(t)(1-F(t))}{m}}{F(t)(1-F(t))+c^2}\, w(dt). \end{aligned}$$

Let

$$\psi(t) = a\hat{F}(t) + (1-a)/2. \tag{10}$$

Then

$$R(F,\psi) = \int_{-\infty}^{\infty} \frac{\left(\frac{a^2}{n} - (1-a)^2 + \frac{1}{m}\right)F(t)(1-F(t)) + \frac{(1-a)^2}{4}}{F(t)(1-F(t))+c^2}\, w(dt)$$

and it is independent of F; moreover,

$$R(F,\psi) = \frac{(1-a)^2}{4c^2} \int_{-\infty}^{\infty} w(dt) \stackrel{\text{df}}{=} R(F,\psi_0) \tag{11}$$

if

$$\frac{a^2}{n} - (1-a)^2 + \frac{1}{m} = \frac{(1-a)^2}{4c^2},$$

i.e. if

$$a = \frac{1}{\frac{4c^2+1}{4c^2} - \frac{1}{n}} \left(\frac{4c^2+1}{4c^2} - \sqrt{\frac{4c^2+1}{4c^2}\left(\frac{1}{m} + \frac{1}{n}\right) - \frac{1}{mn}} \right). \tag{12}$$

It is easy to see that for any m, n, c we have $0 < a < 1$.

Define a mixed strategy σ_k of nature in the same way as τ_k, with α given by (13) below and a given by (12). Now the risk is

$$R(F,\psi) = \int_{-\infty}^{\infty} \frac{E(\psi(t) - F(t))^2 + \frac{F(t)(1-F(t))}{m}}{F(t)(1-F(t)) + c^2} \, w(dt)$$

and for the strategy σ_k the expected risk is

$$r(\sigma_k, \psi) = \int_{-\infty}^{\infty} E_{\sigma_k}\left[\frac{E(\psi(t) - F(t))^2}{F(t)(1-F(t)) + c^2}\right] w(dt) + r_0(\sigma_k),$$

where $r_0(\sigma_k)$ does not depend on ψ and $E_{\sigma_k}(\cdot)$ is the expectation with respect to the density $g(p)$ in the strategy σ_k. In the same manner as in the case of estimation, this leads to the Bayes predictor with respect to σ_k given by

$$\psi_{\sigma_k}(t) = \begin{cases} 0 & \text{if } t < -k, \\ \dfrac{n\hat{F}(t) + \alpha/2}{n+\alpha} & \text{if } -k \le t < k, \\ 1 & \text{if } t \ge k. \end{cases}$$

For

$$\frac{n}{n+\alpha} = a, \tag{13}$$

where a is now given by (12), $\psi_{\sigma_k} = \psi_0$ if $-k \le t < k$, and the Bayes risk $r(\sigma_k, \psi_{\sigma_k})$ is

$$r(\sigma_k, \psi_{\sigma_k}) = \frac{(1-a)^2}{4c^2} \int_{-\infty}^{\infty} I_{[-k,k)}(t) \, w(dt).$$

Then as before we prove that the *predictor* $\psi_0(t) = a\hat{F}(t) + (1-a)/2$, *where* a *is given by* (12), *is a minimax predictor of* $\check{F}(t)$.

For problems of estimation of a cumulative distribution function see [1], [2], [4], [6]. Minimax estimators of a cumulative distribution function for 4 loss functions different from (1) were found by Phadia in [3].

References

[1] O. P. Aggarwal, *Some minimax invariant procedures for estimating a cumulative distribution function*, Ann. Math. Statist. 26 (1955), 450–463.

[2] A. Dvoretzky, J. Kiefer and J. Wolfowitz, *Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator*, ibid. 27 (1956), 642–669.

[3] E. G. Phadia, *Minimax estimation of a cumulative distribution function*, Ann. Statist. 1 (1973), 1149–1157.

[4] R. R. Read, *The asymptotic inadmissibility of the sample distribution function*, Ann. Math. Statist. 43 (1972), 89–95.

[5] S. Trybuła, *Two problems of minimax estimation*, Zastos. Mat. 14 (1974), 41–47.

[6] —, *Estimation of the difference of cumulative distribution functions*, Bull. Polish Acad. Sci. Math. 32 (1984), 243–246.

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

Received on 9.7.1992