KYBERNETIKA ČÍSLO 2, ROČNÍK 3/1967

On the Statistical Decision Problems with Discrete Parameter Space

IGOR VAJDA

In this paper definitions of the Bayes risk and information in a sample concerning a parameter in the framework of a classical model of statistical decision with an abstract sample space and discrete parameter space are given and a relation between them is investigated. We try also to estimate these quantities by means of simpler expressions in order to obtain a platform for a study of an asymptotic behaviour of them.

1. INTRODUCTION

Let us consider the classical model of statistical decision with a parameter probability space $(X, \mathscr{X}, \mu)$, measurable sample space $(Y, \mathscr{Y})$, set $\{v_x\}$, $x \in X$, of conditional distributions of the variable $y \in Y$ defined on the σ-algebra $\mathscr{Y}$, decision measurable space $(X, \mathscr{X})$, and with a non-negative and $\mathscr{X} \otimes \mathscr{X}$-measurable loss function w defined on $X \otimes X$. In this paper we shall assume that the set X of the possible values, in general abstract, of the parameter x is countable, i.e. that the parameter space is discrete. Besides it we shall assume without loss of generality that $\mathscr{X}$ is the σ-algebra of all subsets of X and that the prior probability $\mu(x)$ is positive for every parameter value $x \in X$.

For every $x \in X$ and for every decision function ϱ (i.e. for every $\mathscr{Y}$-measurable mapping of the measurable space $(Y, \mathscr{Y})$ into $(X, \mathscr{X})$ we define the average value of the loss corresponding to them by

$$(1.1) \qquad R(x, \varrho) = \int_Y w(x, \varrho(y)) \, \mathrm{d}v_x(y) ,$$

and for every decision function ϱ we define the average risk $R(\varrho)$ by

$$(1.2) \qquad R(\varrho) = \sum_{x \in X} \mu(x) \, R(x, \varrho) .$$

In the paper we use the well-known Bayes principle of ordering of the decision functions based on the average risk. By optimal decision function (if it exits) we understand a decision function ϱ_0 which minimizes the average risk, i.e. which satisfies the following equality

$$(1.3) \qquad R(\varrho_0) = \inf_{\mathscr{R}} R(\varrho) = r\,,$$

where $\mathscr{R}$ is the set of all possible decision functions and where the non-negative number r is the so-called *Bayes risk*. Hence, the Bayes risk seems be a fundamental characteristics of the model we have considered.

Another characteristics of great importance is the *average amount of information* I contained in the sample y concerning the parameter x. This quantity can be defined as it is described below.

Let us denote by ω the probability distribution defined on the Cartesian product σ-algebra $\mathscr{X} \otimes \mathscr{Y}$ by

$$(1.4) \qquad \omega(E) = \sum_{x \in X} \mu(x)\, \nu_x((E)_x) \quad \text{for every} \quad E \in \mathscr{X} \otimes \mathscr{Y}\,,$$

where $(E)_x$ is the x-section of the set E, i.e.

$$(E)_x = \{y \in Y : (x, y) \in E\}\,,$$

and let us denote by $\tilde{\omega}$ the marginal distribution induced by ω on the σ-algebra $\mathscr{Y}$, i.e. let

$$(1.5) \qquad \tilde{\omega}(F) = \sum_{x \in X} \mu(x)\, \nu_x(F) \quad \text{for every} \quad F \in \mathscr{Y}\,.$$

(It is easy to see that, in view of (1.4) and (1.5), $\omega \ll \mu \otimes \tilde{\omega}$, where $\mu \otimes \tilde{\omega}$ is the Cartesian product distribution on $\mathscr{X} \otimes \mathscr{Y}$.) Then the corresponding average amount of information I is defined by

$$(1.6) \qquad I = \int_{X \otimes Y} \log f(x, y)\, \mathrm{d}\omega(x, y)\,,$$

where $f(x, y)$ is the Radon-Nikodym density of the joint probability measure ω with respect to the product measure $\mu \otimes \tilde{\omega}$. (Let us note that all logarithms in this paper are taken to the base e.)

From the intuitive point of view it is clear that, though the information I does not depend on the loss function w, for a sufficiently wide class of loss functions there exists a relation between r and I. This fact is a platform for the study of statistical decision problems from the point of view of information theory. The data reduction theory recently developed by A. Perez [5] shows that the indicated relation plays a growing role in solutions of certain class of decision problems.

The purpose of this paper is to estimate the Bayes risk and information in the framework of general model of statistical decision with discrete parameter space and to investigate the indicated relation between them. This general questions are studied in the following section. The results of this section are then used in Sec. 3 devoted to the study of of the rate of convergence of information and Bayes risk in some classes of decision models as it is more precisely described below.

Let $n = 1, 2, \ldots, \infty$ be the size of mutually independent samples $y = (y_1, \ldots, y_n)$, i.e. suppose that the measurable sample space of the model is of the form

$$(1.7) \qquad (Y^n, \mathscr{Y}^n) = \bigotimes_{i=1}^{n} (Y_i, \mathscr{Y}_i), \quad n = 1, 2, \ldots, \infty,$$

where Y_i is the set of all y_i's with a given σ-algebra $\mathscr{Y}_i$ and suppose that the joint probability distribution of the n-vector $(y_1, \ldots, y_n)$ under the condition that $x \in X$ is the realized value of the parametr is a Cartesian product measure v_x^n. It is clear that, for every $n = 1, 2, \ldots, v_x^n$ is the restriction of v_x^∞ on the σ-algebra $\mathscr{Y}^n \subset \mathscr{Y}^\infty$, where the latter inclusion (as well as the inclusions $\mathscr{Y}_i \subset \mathscr{Y}^n$, $i = 1, 2, \ldots, n$, $n = 1, 2, \ldots, \infty$, that will be used below) is written in accordance with a well-known identification convention for product σ-algebras, and that

$$v_x^\infty = \bigotimes_{i=1}^{\infty} v_{xi} \quad \text{for every} \quad x \in X,$$

where v_{xi} denotes over all the paper the restriction of v_x^∞ on the sub σ-algebra $\mathscr{Y}_i \subset \mathscr{Y}^\infty$.

We shall denote by I_n or r_n the information or Bayes risk respectively corresponding to the measurable sample space (1.7). Since $\mathscr{Y}^n$, $n = 1, 2, \ldots$, is an increasing sequence of σ-algebras, I_n, $n = 1, 2, \ldots$, is a non-decreasing sequence (cf. Th. 12 in [4]) and

$$(1.8) \qquad \lim_{n \to \infty} I_n = I_\infty,$$

whereas r_n, $n = 1, 2, \ldots$, is a non-increasing sequence (cf. Th. 6.1 in [5]) and

$$(1.9) \qquad \lim_{n \to \infty} r_n = r_\infty.$$

From the point of view of application it is important to ask which is the rate of convergence above. This question was studied in [6] under the assumption that X is finite. According to [6], both r_n and I_n converge to their limit values r_∞ and I_∞ exponentially in a sufficiently wide class of decision models with finite parameter space, for example when v_x^∞, $x \in X$, are stationary and mutually different Cartesian product measures. It seems that an analogical assertion need not be true when the parameter space is infinite. Sec. 3 of this paper is devoted to the rate of convergence mentioned above under the assumption of a discrete parameter space.

The first our result it obvious. Let γ be a real valued function defined on the parameter space X by

$$\gamma(x) = \inf w(x, x')\,, \tag{2.1}$$

where the infimum is extended over all $x' \in X$ different from x, and let us denote

$$\Delta(v_x, v_{x'}) = \sup_{E \in \mathscr{Y}} |v_x(E) - v_{x'}(E)|\,.$$

Theorem 1. *If the loss function w is bounded from above by w_0, then*

$$\frac{\mu(x)\,\gamma(x)\,\mu(x')\,\gamma(x')}{\mu(x)\,\gamma(x) + \mu(x')\,\gamma(x')}\,(1 - \Delta(v_x, v_{x'})) \leqq r \leqq w_0 \sum_{x \in X} \mu(x)\,(1 - v_x(E_x)) \tag{2.2}$$

for every $x \neq x'$ such that the left side has a meaning and for every measurable disjoint decomposition $\{E_x\}$, $x \in X$, of the sample space Y.

Proof. The right inequality is clear. The left inequality is non-trivial only if both $\gamma(x)$ and $\gamma(x')$ are positive. If this condition is satisfied, then, according to (1.2) and (1.3), there exists a decision function $\varrho_\varepsilon \in \mathscr{R}$ such that

$$\sum_{x \in X} \mu(x)\,R(x, \varrho_\varepsilon) < r + \varepsilon \quad \text{for every} \quad \varepsilon > 0\,,$$

and consequently

$$R(x, \varrho_\varepsilon) < \frac{r}{\mu(x)} + \frac{\varepsilon}{\mu(x)} \quad \text{for every} \quad x \in X\,.$$

Hence, if we denote $E_{x''} = \{y \in Y : \varrho_\varepsilon(y) = x''\}$ for every $x'' \in X$, we can write

$$\gamma(x) \sum_{x'' \neq x} v_x(E_{x''}) \leqq \sum_{x'' \in X} w(x, x'')\,v_x(E_{x''}) = R(x, \varrho_\varepsilon) < \frac{r}{\mu(x)} + \frac{\varepsilon}{\mu(x)}\,,$$

and similarly

$$\gamma(x') \sum_{x'' \neq x'} v_{x'}(E_{x''}) < \frac{r}{\mu(x')} + \frac{\varepsilon}{\mu(x')}\,.$$

Since

$$\sum_{x'' \neq x} v_x(E_{x''}) = 1 - v_x(E_x)\,,$$

and since, in view of $x \neq x'$,

$$v_x(E_x) > 1 - \frac{r}{\mu(x)\,\gamma(x)} - \frac{\varepsilon}{\mu(x)\,\gamma(x)}\,,$$

$$v_{x'}(E_x) < \frac{r}{\mu(x')\,\gamma(x')} + \frac{\varepsilon}{\mu(x')\,\gamma(x')}\,,$$

it follows from the definition of $\Delta(\nu_x, \nu_{x'})$ that

$$\Delta(\nu_x, \nu_{x'}) \geqq 1 - \left(\frac{1}{\mu(x)\,\gamma(x)} + \frac{1}{\mu(x')\,\gamma(x')}\right)(r + \varepsilon)\,.$$

Since this inequality remains true for arbitrarily small $\varepsilon > 0$, we can write

$$\Delta(\nu_x, \nu_{x'}) \geqq 1 - r\left(\frac{1}{\mu(x)\,\gamma(x)} + \frac{1}{\mu(x')\,\gamma(x')}\right);$$

the desired inequality is proved.

If the parameter space is finite, then it is well-known and frequently used that $I \leqq H(\mu)$, where $H(\mu)$ is the entropy of the parameter space. The non-negative difference $H(\mu) - I$ has been called equivocation by Shannon. The following extension of validity of the latter inequality will be often used in the sequel.

Lemma 1. *If $H(\mu)$ is the entropy of the parameter probability space, i.e. if*

$$H(\mu) = -\sum_{x \in X} \mu(x) \log \mu(x)\,, \tag{2.3}$$

then

$$I \leqq H(\mu)\,. \tag{2.4}$$

If $H(\mu) < \infty$ then the sign of equality in (2.4) holds if and only if $\Delta(\nu_x, \nu_{x'}) = 1$ for every $x \neq x'$.

Proof. If $H(\mu) = \infty$, then the assertion of the Lemma is clear. Let, consequently, $H(\mu) < \infty$.

Suppose that the sample space is discrete (i.e. countable or finite) and that the σ-algebra $\mathcal{Y}$ contains all its subsets. Under this assumptions it is easily verified that the Radon-Nikodym density f of the distribution ω with respect to the product distribution $\mu \otimes \tilde{\omega}$ is of the form

$$f(x, y) = \frac{\nu_x(y)}{\sum_{x \in X} \mu(x)\,\nu_x(y)}\,.$$

Since we can write

$$-\mu(x) \log \mu(x) = -\sum_{y \in Y} \mu(x)\,\nu_x(y) \log \mu(x)$$

and since, in view of (1.6),

$$I = \sum_X \sum_Y \mu(x)\,\nu_x(y) \log \left(\frac{\nu_x(y)}{\sum_X \mu(x)\,\nu_x(y)}\right),$$

we can write

$$H(\mu) - I = \sum_X \sum_Y \psi(x, y)\,, \tag{2.5}$$

where

$$\psi(x, y) = \begin{cases} 0 \quad \text{if} \quad \nu_x(y) = 0\,, \\ \mu(x)\,\nu_x(y) \log\left(\dfrac{\sum_X \mu(x)\,\nu_x(y)}{\mu(x)\,\nu_x(y)}\right) \quad \text{if} \quad \nu_x(y) \neq 0\,. \end{cases} \tag{2.6}$$

It is clear that $\psi(x, y) \geqq 0$ and consequently the inequality (2.4) under the conditions we have considered holds. In order to extent the validity of (2.4) to the case of general measurable sample space $(Y, \mathscr{Y})$ we can use an obvious procedure based on Th. 13 in [4].

If, for every $x \neq x'$, $\Delta(\nu_x, \nu_{x'}) = 1$, then there exists a countable disjoint measurable decomposition $\{E_x\}$, $x \in X$, such that $\nu_x(E_x) = 1$. (cf. (3) in [6]). Let us consider a measurable sample space $(Y^*, \mathscr{Y}^*)$ defined by $Y^* = \{E_x\}$, $x \in X$, $\mathscr{Y}^* = \mathscr{S}(\{E_x\}, x \in X) \subset \mathscr{Y}$ (here and in the sequel $\mathscr{S}(\cdot)$ denotes the least σ-algebra generated by the indicated class of sets) and denote by I^* the information corresponding to this space. Then, by (2.5), (2.6), and $H(\mu) < \infty$, the equality $H(\mu) = I^*$ holds. Since, according to Th. 12 in [4], $I^* \leqq I$, the desired equality $H(\mu) = I$ is proved. The proof of the converse assertion will be based on the following

Theorem 2. *If the loss function w is bounded from above by w_0 and if $I < \infty$, then*

$$r \leqq \frac{w_0}{2 \log 2} (H(\mu) - I)\,. \tag{2.7}$$

Proof. If $H(\mu) = \infty$, then the assertion of the Theorem is clear. Hence let us assume $H(\mu) < \infty$.

(a) Let both X and Y be finite sets, say $X = \{1, 2, \ldots, m\}$, $Y = \{1, 2, \ldots, n\}$, and let $\mu(i)$ and $\nu_i(j)$ be defined in accordance with Sec. 1. In information theory it is usually defined the so-called minimum probability of error P by

$$P = \min \sum_{i=1}^{m} \mu(i)\, \nu_i(Y - E_i)\,,$$

where the minimum is taken over the set of all disjoint decompositions $\{E_i\}$ of Y. It was proved earlier (cf. [1], [2]) that in this case the following inequality holds

$$P \leqq \frac{1}{2 \log 2} (H(\mu) - I)\,.$$

(A similar result is proved in [2] also in case Y is the real line and ν_i, $i = 1, 2, \ldots, m$, are absolutely continuous probability distribution on it.) As, for every loss function

$w \leqq w_0$, a routine verification gives $r \leqq w_0 P$, the inequality (2.7) holds under the condition that both X and Y are finite.

(b) Let us denote the elements of X subsequently by $x_1, x_2, \ldots$, and let μ_n be a priori distribution derived from μ by

$$\mu_n(x_i) = \mu(x_i)\,, \quad i = 1, 2, \ldots, n - 1\,,$$

$$\mu_n(x_n) = \sum_{i=n}^{\infty} \mu(x_i)\,.$$

It is easily proved that in this case

$$\lim_{n \to \infty} H(\mu_n) = H(\mu)\,. \tag{2.8}$$

(c) Let us denote by r_n the Bayes risk and by I_n the information obtained by the replacing of μ by μ_n (cf. Sec. 7 in [4] and Sec. 5 in [5]). According to Th. 12 in [4], $I_n \leqq I$ and, according to Th. 6.1 in [5], $r_1 \geqq r_2 \geqq \ldots \geqq r$ and, moreover,

$$0 \leqq r_n - r \leqq \sqrt{(2 w_0 r_n (I - I_n))}\,.$$

As it follows from $w \leqq w_0$ that $r_1 \leqq w_0$, we obtain

$$0 \leqq r_n - r \leqq w_0 \sqrt{(2(I - I_n))}\,. \tag{2.9}$$

According to Th. 13 in [4], there exists a sequence $\mathscr{D}_1 \subset \mathscr{D}_2 \subset \ldots$ of finite measurable disjoint decompositions of Y such that $I_n^m \leqq I_n$,

$$\lim_{n \to \infty} I_n^m = I_n\,, \tag{2.10}$$

where I_n^m is the information defined with respect to $\mathscr{Y} = \mathscr{S}(\mathscr{D}_m)$, $\mu = \mu_n$, $m, n = 1, 2, \ldots$. One more application of Th. 6.1 in [5] together with (2.9) yields that

$$0 \leqq r_n^m - r \leqq r_n^m - r_n + r_n - r \leqq w_0 \sqrt{2}(\sqrt{(I_n - I_n^m)} + \sqrt{(I - I_n)})\,,$$

where the meaning of r_n^m is clear. Since

$$\lim_{n \to \infty} I_n = I \tag{2.11}$$

$$\lim_{n, m \to \infty} I_n^m = I \tag{2.12}$$

(cf. Th. 12 in [4]), we conclude that

$$\lim_{n, m \to \infty} r_n^m = r\,. \tag{2.13}$$

(d) The tools are now at hand to prove (2.7). If we apply this inequality to the finite parameter space $\{x_1, x_2, \ldots, x_n\}$ and finite sample space $Y = \mathscr{D}_m$, $n, m =$

$= 1, 2, \ldots$ (cf. (a)), then we get

$$r_n^m \leqq \frac{w_0}{2\log 2}(H(\mu_n) - I_n^m)$$

and using (2.12) and (2.13) we complete the desired proof.

Now we can conclude the proof of Lemma 1. It remains to prove that the equality $H(\mu) = I$ together with $H(\mu) < \infty$ implies that $\Delta(v_x, v_{x'}) = 1$ for every $x \neq x'$. According to (2.7), $H(\mu) = I$ implies $r = 0$ for every bounded loss function and hence also for $w(x, x') = 0$ or 1 depending on whether $x = x'$ or $x \neq x'$. In this special case $\gamma(x) = 1$ for all $x \in X$ and the desired assertion follows from Theorem 1.

Lemma 2. *If* $H(\mu) < \infty$, *then*

$$H(\mu) - I \leqq \sum_{x,x' \in X} \sqrt{[\mu(x)\, v_x(E_x)\, (\sum_{x'' \neq x} \mu(x'')\, v_{x''}(E_{x'}))]} + \tag{2.14}$$

$$+ \sum_{x \in X} \sqrt{[\mu(x)\, v_x(E_0)\, (\sum_{x'' \neq x} \mu(x'')\, v_{x''}(E_0))]}$$

for every class $\{E_x\}$, $E_x \in \mathscr{Y}$, $x \in X$, *where*

$$E_0 = \bigcap_{x \in X} (Y - E_x).$$

Proof. (a) Let Y be a discrete space all subsets of which are contined in $\mathscr{Y}$. By (2.5) it holds

$$H(\mu) - I \leqq \sum_{x,x' \in X} \sum_{y \in E_{x'}} \psi(x, y) + \sum_{x \in X} \sum_{y \in E_0} \psi(x, y). \tag{2.15}$$

Since for every $z > 0$

$$\log(1 + z) \leqq \sqrt{z},$$

we can write

$$\psi(x, y) \leqq \sqrt{[\mu(x)\, v_x(y)\, (\sum_{x'' \neq x} \mu(x'')\, v_{x''}(y))]}$$

(cf. (2.6)). If we apply the Schwarz's inequality to the series

$$\sum_{y \in E_{x'}} \sqrt{[\mu(x)\, v_x(y)\, (\sum_{x'' \neq x} \mu(x'')\, v_{x''}(y))]},$$

then we obtain

$$\sum_{y \in E_{x'}} \psi(x, y) \leqq \sqrt{[\mu(x)\, v_x(E_{x'})\, (\sum_{x'' \neq x} \mu(x'')\, v_{x''}(E_{x'}))]}$$

and similarly

$$\sum_{y \in E_0} \psi(x, y) \leqq \sqrt{[\mu(x)\, v_x(E_0)\, (\sum_{x'' \neq x} \mu(x'')\, v_{x''}(E_0))]}.$$

The latter two inequalities together with (2.15) imply the desired inequality (2.14).

(b) Let (Y, $\mathscr{Y}$) be an arbitrary measurable space and define a disjoint measurable decomposition $\mathscr{D}$ of Y in a following way: $E \in \mathscr{D}$ if and only if there exist sets $E_1, E_2, \ldots, E_n$, $E_i \in \{E_x\}$, $x \in X$, or $E_i = E_0$ such that

$$E = \bigcap_{i=1}^{n} E_i \, .$$

If we put in (a) $Y = \mathscr{D}$, $\mathscr{Y} = \mathscr{S}(\mathscr{D})$, and if we define

$$I^* = \int_{X \otimes Y} \log f^*(x, y) \, \mathrm{d}\omega(x, y) \, ,$$

where f^* is $\mathscr{S}(\mathscr{D})$-measurable version of the Radon-Nikodym density f then, according to (a),

$$H(\mu) - I^* \leqq \sum_{x, x' \in X} \sqrt{[\mu(x) \, v_x(E_{x'}) \, (\sum_{x'' \neq x} \mu(x'') \, v_{x''}(E_{x'}))]} + \\ + \sum_{x \in X} \sqrt{[\mu(x) \, v_x(E_0) \, (\sum_{x'' \neq x} \mu(x'') \, v_{x''}(E_0))]} \, .$$

Since it follows from Th. 12 in [4] that $I^* \leqq I$, we have $H(\mu) - I \leqq H(\mu) - I^*$ and the proof of (2.14) is complete.

Theorem 3. *If* $H(\mu) < \infty$, *then*

$$(2.16) \qquad 2 \log 2 \frac{\mu(x) \, \mu(x')}{\mu(x) + \mu(x')} \, (1 - \Delta(v_x, v_{x'})) \leqq H(\mu) - I \leqq \\ \leqq \sum_{x \in X} \sqrt{[\mu(x) \, (1 - v_x(E_x))]} \, (1 + \sum_{x \in X} \sqrt{\mu(x)})$$

for every $x \neq x'$ *and for every disjoint measurable decomposition* $\{E_x\}$, $x \in X$, *of sample space Y.*

Remark. The upper estimate has a meaning only in case $\sum \sqrt{\mu(x)} < \infty$; it is easily proved that if this condition is satisfied, then $H(\mu) < \infty$.

Proof. (a) As $\{E_x\}$, $x \in X$, is a decomposition of Y, E_0 in the preceding Lemma is empty and hence

$$(2.17) \qquad H(\mu) - I \leqq \sum_{x \in X} \sqrt{[\mu(x) \, v_x(E_x) \, (\sum_{x'' \neq x} \mu(x'') \, v_{x''}(E_x))]} + \\ + \sum_{x \in X} \sum_{x' \neq x} \sqrt{[\mu(x) \, v_x(E_{x'}) \, (\sum_{x'' \neq x} \mu(x'') \, v_{x''}(E_{x'}))]} \, .$$

The first sum can be easily estimated from above by

$$\sum_{x \in X} \sqrt{[\mu(x) \sum_{x'' \neq x} \mu(x'') \, (1 - v_{x''}(E_{x''}))]}$$

and hence also by

$$\sum_{x \in X} \sqrt{[\mu(x) \sum_{x'' \in X} \mu(x'') (1 - \nu_{x''}(E_{x''}))]}$$

or by

$$\sum_{x \in X} \sqrt{\mu(x)} (\sum_{x'' \in X} \sqrt{[\mu(x) (1 - \nu_{x''}(E_{x''}))]}) .$$

If we apply to the second sum in (2.17) the Schwarz's inequality again, then we obtain the following upper estimate of it:

$$\sum_{x \in X} \sqrt{[\sum_{x' \neq x} \mu(x) \nu_x(E_{x'}) (\sum_{x'' \neq x} \mu(x'') \sum_{x' \neq x} \nu_{x''}(E_{x'}))]} =$$

$$= \sum_{x \in X} \sqrt{[\mu(x) \nu_x(\bigcup_{x' \neq x} E_{x'}) \sum_{x'' \neq x} \mu(x'') \nu_{x''}(\bigcup_{x' \neq x} E_{x'})]} \leqq \sum_{x \in X} \sqrt{[\mu(x) (1 - \nu_x(E_x))]} .$$

In view of this estimates and in view of (2.17), the right inequality in (2.16) holds.

(b) Let us define in Theorem 2 the zero-one loss function w similarly as above. In this special case $\gamma(x)$ is identically 1 and the left inequality in (2.16) is a consequence of (2.7) and (2.1).

3. DECISION MODEL WITH INDEPENDENT SAMPLES

In this section we shall deal with the classical model of statistical decision with descrete parameter space under the assumption that for every realized value of the parameter $x \in X$ the sequence of samples $y_1, y_2, \ldots$ is an independent (not necessarily stationary) random sequence. Over all the section we shall follow the notation and terminology employed above.

Lemma 3. *If, for every $x \in X$, the sequence of samples is independent, then, for every $n = 1, 2, \ldots$ and for every $F_i \in \mathcal{Y}_i$, $i = 1, 2, \ldots, n$, there exists a disjoint measurable decomposition $\{E_x\}$, $x \in X$, of Y^n (i.e. $E_x \in \mathcal{Y}^n$) such that*

$$\nu_x^n(E_x) > 1 - 2e^{-n\eta_n(x)} , \quad n = 1, 2, \ldots , \tag{3.1}$$

where

$$\eta_n(x) = \inf_{x' \neq x} \frac{1}{5n} \left| \sum_{i=1}^{n} (\nu_{xi}(F_i) - \nu_{x'i}(F_i)) \right| , \quad n = 1, 2, \ldots \tag{3.2}$$

Proof. Let n be an arbitrary and define on $(Y^n, \mathcal{Y}^n)$ a sequence $f_1, f_2, \ldots, f_n$ of measurable functions by

$$f_i(y_1, y_2, \ldots, y_n) = \chi_{F_i}(y_i) , \quad i = 1, 2, \ldots, n ,$$

where χ is the characteristic function. It is to see that for every probability distribution ν_x^n on $(Y^n, \mathcal{Y}^n)$, f_i are independent random variables taking values between 0 and 1 with expectations $v_{xi}(F_i)$ and with variances uniformly bounded from above by $\frac{1}{4}$. Under this conditions a routine verification (using inequality § 18.1.A in [3], Chapter V) gives for every $0 < \tau < \frac{1}{4}$,

$$v_x^n(Y^n - E_x(\tau)) < 2e^{-\tau n},$$

where

$$E_x(\tau) = \left\{(y_1, \ldots, y_n) : \frac{1}{n} \left| \sum_{i=1}^{n} (f_i(y_1, \ldots, y_n) - v_{xi}(F_i)) \right| \leqq \tau \right\}.$$

Let us define

$$\tilde{E}_x = \begin{cases} E_x(\eta_n(x)) & \text{if} \quad \eta_n(x) > 0, \\ \emptyset & \text{if} \quad \eta_n(x) = 0. \end{cases}$$

Since $0 \leqq \eta_n(x) \leqq \frac{1}{5}$, the preceding inequality yields

$$v_x^n(\tilde{E}_x) > 1 - 2e^{-n\eta_n(x)} \quad \text{for every} \quad x \in X.$$

Since (3.2) implies that, for every $x \in X$ such that $\eta_n(x) > 0$,

$$\frac{1}{n} \left| \sum_{i=1}^{n} (v_{xi}(F_i) - v_{x'i}(F_i)) \right| < 2\eta_n(x) \quad \text{for every} \quad x' \neq x,$$

we conclude that $\tilde{E}_x \cap \tilde{E}_{x'} = \emptyset$ for every $x \neq x'$. If we put $E_x = \tilde{E}_x$ for all $x \in X$ except one, say x_0, and if

$$E_{x_0} = \tilde{E}_{x_0} \cup (Y^n - \bigcap_{x \in X} E_x),$$

then $\{E_x\}$, $x \in X$, is a disjoint system of sets of the desired properties.

If we use Lemma 3 together with Theorem 1 and 3, we get the following

Theorem 4. *If, for every $x \in X$, the sequence of samples is independent, then, for every sequence $F_i \in \mathcal{Y}_i$, $i = 1, 2, \ldots$,*

$$\text{(3.3)} \qquad H(\mu) - I_n \leqq (1 + \sum_{x \in X} \sqrt{\mu(x)}) \sum_{x \in X} \sqrt{(2\mu(x)\, e^{-n\eta_n(x)})}, \quad n = 1, 2, \ldots,$$

and if the loss function is bounded by w_0, then

$$\text{(3.4)} \qquad r_n \leqq 2w_0 \sum_{x \in X} \mu(x)\, e^{-n\eta_n(x)}, \quad n = 1, 2, \ldots,$$

where $\eta_n(x)$ is defined by (3.2).

We shall say that an independent random sequence $y_1, y_2, \ldots$ is stationary, if $(Y_i, \mathcal{Y}_i) = (Y_j, \mathcal{Y}_j)$ and $v_{xi} = v_{xj}$ for every $i, j = 1, 2, \ldots$

If the sequence of samples is, for every realized value of the parameter, independent and stationary, then the model of statistical decision is completely described by a parameter probability space, loss function, one-dimensional measurable sample space $(Y_1, \mathcal{Y}_1)$ and by a set of one-dimensional conditional probability distributions ν_x^1, $x \in X$, on $\mathcal{Y}_1$. This will be respected in the remainder of this paper.

The following Theorem is just a restatement of Theorem 4 to the stationary case.

Theorem 4s. *If, for every $x \in X$, the sequence of samples is independent and stationary, then, for every set $F \in \mathcal{Y}_1$,*

$$(3.5) \qquad H(\mu) - I_n \leqq (1 + \sum_{x \in X} \sqrt{\mu(x)}) \sum_{x \in X} \sqrt{(2\mu(x)\, \mathrm{e}^{-n\eta(x)})}\,, \quad n = 1, 2, \ldots$$

and if $w \leqq w_0$, then

$$(3.6) \qquad r_n \leqq 2w_0 \sum_{x \in X} \mu(x)\, \mathrm{e}^{-n\eta(x)} \quad \textit{for every} \quad n = 1, 2, \ldots,$$

where

$$(3.7) \qquad \eta(x) = \inf_{x' \neq x} \tfrac{1}{5} |\nu_x^1(F) - \nu_{x'}^1(F)| \quad \textit{for every} \quad x \in X\,.$$

Corollary 1. Let for $X = \{1, 2, \ldots\}$ the assumptions of Theorem 4s be satisfied and let there exists such $\alpha > 0$ that

$$(3.8) \qquad \eta(i) = \frac{a_i}{i^{1/\alpha}} \quad \text{for every} \quad i \in X\,,$$

where a_i's are bounded from below by $a > 0$. Let $s(t)$ and $\tilde{s}(t)$ be non-negative functions defined for $t \geqq 1$ by

$$(3.9) \qquad s(t) = \sum_{i=[t]}^{\infty} \mu(i)$$

$$\tilde{s}(t) = \sum_{i=[t]}^{\infty} \sqrt{\mu(i)}\,,$$

where $[t]$ denotes the least integer greater than or equal to t. Then $r_\infty = 0$ and

$$(3.10) \qquad r_n \leqq 2w_0[n^{\alpha(1-\varepsilon)}\, \mathrm{e}^{-n^\varepsilon a} + s(n^{\alpha(1-\varepsilon)})]\,, \quad n = 1, 2, \ldots$$

for every $0 < \varepsilon < 1$. If $\tilde{s}(1) < \infty$, then $I_\infty = H(\mu)$ and

$$(3.11) \qquad H(\mu) - I_n < \sqrt{2}(1 + \tilde{s}(1))\, [n^{\alpha(1-\varepsilon)}\, \mathrm{e}^{-(n^\varepsilon a)/2} + \tilde{s}(n^{\alpha(1-\varepsilon)})]$$

for every $0 < \varepsilon < 1$ and $n = 1, 2, \ldots$.

Proof. According to (3.6) we have, for every $n = 1, 2, \ldots$,

$$r_n \leqq 2w_0 \sum_{i=1}^{\infty} \mu(i)\, \mathrm{e}^{-n\eta(i)} \leqq 2w_0(\sum_{i=1}^{m} \mathrm{e}^{-n\eta(i)} + \sum_{i=m+1}^{\infty} \mu(i)\, \mathrm{e}^{-n\eta(i)}) \leqq$$

$$\leqq 2w_0(m\mathrm{e}^{-n \min\limits_{1 \leqq i \leqq m} \eta(i)} + s(m + 1)) \leqq 2w_0(m\mathrm{e}^{-n(a/m^{1/\alpha})} + s(m + 1))$$

for every $m = 1, 2, \ldots$ Let m be an integer satisfying the inequality

$$n^{\alpha(1-\varepsilon)} - 1 \leqq m < n^{\alpha(1-\varepsilon)},$$

where $0 < \varepsilon < 1$. Then it is easily proved that

$$m\mathrm{e}^{-n(a/m^{1/\alpha})} < n^{\alpha(1-\varepsilon)}\mathrm{e}^{-n^{\varepsilon}a},$$

$$s(m+1) \leqq s(n^{\alpha(1-\varepsilon)})$$

and hence that (3.10) holds. The equality $r_\infty = 0$ is clear.

Similarly, according to (3.5),

$$H(\mu) - I_n \leqq \sqrt{2}(1 + \tilde{s}(1))\,(m\mathrm{e}^{-(na/2m^{1/\alpha})} + \tilde{s}(m+1)),$$

for every $m = 1, 2, \ldots$ and the conclusion of the proof of (3.11) is now clear.

Since the assumption $\tilde{s}(1) < \infty$ implies that

$$\lim_{t\to\infty} \tilde{s}(t) = 0$$

the equality $I_\infty = H(\mu)$ follows from (3.11) and (1.8).

Example 1. Let us consider the case when the prior distribution μ is geometrical, i.e.

$$\mu(i) = \frac{1-\mu}{\mu}\mu^i, \quad i = 1, 2, \ldots, \quad \text{where} \quad 0 < \mu < 1.$$

In this case

$$s(t) = \mu^{[t]-1}$$

$$\tilde{s}(t) = \frac{\sqrt{(1-\mu)}}{1-\sqrt{\mu}}\mu^{([t]-1)/2}.$$

Suppose that (3.8) is satisfied for some α, say $\alpha = \frac{1}{2}$. As in this case $\tilde{s}(1) < \infty$, we get from Corollary 1 that $r_\infty = 0$, $I_\infty = H(\mu)$, where

$$H(\mu) = \left(1 + \frac{1}{\mu} + \frac{1}{\mu^2}\right)\log\frac{1}{\mu} + \log\frac{1}{1-\mu}$$

and, according to (3.10) or (3.11) for $\varepsilon = \frac{1}{2}$, we obtain that

$$r_n < 2w_0(n^{1/4}\mathrm{e}^{-n^{1/2}a} + \mu^{n^{1/4}}),$$

$$H(\mu) - I_n < \sqrt{2}\left(1 + \frac{\sqrt{(1-\mu)}}{1-\sqrt{\mu}}\right)\left(n^{1/4}\mathrm{e}^{-(n^{1/2}a/2)} + \frac{\sqrt{(1-\mu)}}{1-\sqrt{\mu}}\mu^{n^{1/4}/2}\right)$$

for every $n = 1, 2, \ldots$

Example 2. In order to find an example of decision problem satisfying the condition (3.8) let us proceed in the following manner. Let $X = \{1, 2, \ldots\}$, $Y_1 = \{0, 1\}$, $\mathcal{Y}_1 = \{\emptyset, \{0\}, \{1\}, \{0, 1\}\}$, and let, for every $i \in X$, the probability distribution ν_i^1 on $\mathcal{Y}_1$ be defined by

$$\nu_i^1(1) = \frac{1}{i}$$

$$\nu_i^1(0) = 1 - \frac{1}{i}\,.$$

If we put $F = \{1\} \in \mathcal{Y}_1$, then, according to (3.7),

$$\eta(i) = \frac{1}{5}\left(\frac{1}{i} - \frac{1}{i+1}\right) = \frac{a_i}{i^2}\,,$$

where

$$a_i = \frac{i^2}{5i(i+1)} \geqq 1/10 \quad \text{for} \quad i = 1, 2, \ldots$$

Consequently, in this case the condition (3.8) is satisfied for $a = 1/10$ and $\alpha = 1/2$.

Corollary 2. Let for $X = \{1, 2, \ldots\}$ the assumptions of Theorem 4s be satisfied and let there exists $0 < \beta < 1$ such that

$$\eta(i) = a_i \beta^i \quad \text{for every} \quad i \in X\,, \tag{3.12}$$

where the set $\{a_1, a_2, \ldots\}$ is bounded from below by $a > 0$. Then $r_\infty = 0$ and

$$r_n < 2w_0\left[\frac{\varepsilon}{\log(1/\beta)} \log n\, \mathrm{e}^{-n^\varepsilon a} + s\left(\frac{\varepsilon}{\log(1/\beta)} \log n\right)\right], \quad n = 1, 2, \ldots \tag{3.13}$$

for every $0 < \varepsilon < 1$. If $\tilde{s}(1) < \infty$, then $I_\infty = H(\mu)$ and

$$H(\mu) - I_n < \sqrt{2(1 + \tilde{s}(1))}\left[\frac{\varepsilon}{\log(1/\beta)} \log n\, \mathrm{e}^{-(n^\varepsilon a/2)} + \tilde{s}\left(\frac{\varepsilon}{\log(1/\beta)} \log n\right)\right] \tag{3.14}$$

for every $n = 1, 2, \ldots$ and $0 < \varepsilon < 1$.

Proof. It was proved above that

$$r_n < 2w_0\left(m\,\mathrm{e}^{-n \min\limits_{1 \leqq i \leqq n} \eta(i)} + s(m + 1)\right)$$

for every n, $m = 1, 2, \ldots$ so that, in view of (3.12),

$$r_n < 2w_0\left(m\,\mathrm{e}^{-na\beta^m} + s(m + 1)\right).$$

It is necessary to choose $m = m(n)$ satisfying the following two conditions

$$\lim_{n \to \infty} m(n) = 0 \,,$$

$$\lim_{n \to \infty} m(n)\, \mathrm{e}^{-na\beta^{m(n)}} = 0 \,.$$

In order to achieve this define m by

$$\frac{\varepsilon}{\log(1/\beta)} \log n - 1 \leqq m < \frac{\varepsilon}{\log(1/\mathrm{b})} \log n \,.$$

It is easily verified that

$$s(m + 1) \leqq s\left(\frac{\varepsilon}{\log(1/\beta)} \log n\right),$$

$$m\mathrm{e}^{-na\beta^{m}} \leqq \frac{\varepsilon}{\log(1/\beta)} \log n\, \mathrm{e}^{-n^{\varepsilon}a}$$

and consequently (3.13) holds. The remainder of the proof is clear.

Example 3. If the prior distribution μ on X is geometric (cf. Example 1) and if $\eta(i)$, $i \in X$, satisfy the condition (3.12) for, say, $\beta = \mathrm{e}^{-1}$, then we easily obtain by means of (3.13) and (3.14) (for $\varepsilon = \frac{1}{2}$) that in this case

$$H(\mu) - I_n < \sqrt{2}\left(1 + \frac{\sqrt{(1-\mu)}}{1-\sqrt{\mu}}\right)\left[\frac{1}{2}\log n\,\mathrm{e}^{-\sqrt{n(a/2)}} + \frac{\sqrt{(1-\mu)}}{1-\sqrt{\mu}}\left(\frac{1}{n}\right)^{1/4\log(1/\mu)}\right]$$

and (if $w \leqq w_0$)

$$r_n < 2w_0\left[\frac{1}{2}\log n\mathrm{e}^{-\sqrt{na}} + \left(\frac{1}{n}\right)^{1/4\log(1/\mu)}\right]$$

for every $n = 1, 2, \ldots$. It is clear that there exists a positive integer n_0 such that for $n > n_0$

$$H(\mu) - I_n < \text{const}\,.\left(\frac{1}{n}\right)^{1/4\log(1/\mu)},$$

$$r_n < \text{const}\,.\left(\frac{1}{n}\right)^{1/2\log(1/\mu)}.$$

Example 4. In order to give an example of decision problem satisfying the condition (3.12) we proceed in the following manner. Let $X = \{1, 2, \ldots\}$ and let $(Y_1, \mathscr{Y}_1)$ be Borel line. Let, for every $i \in X$, the probability distribution ν_i^1 on $\mathscr{Y}_1$ be Poisson distribution with parameter i, i.e. let

$$\nu_i^1(j) = \mathrm{e}^{-i}\frac{i^j}{j!} \quad j = 0, 1, 2, \ldots$$

If we put $F = \{0\} \in Y_1$, then it is clear that

$$\eta(i) = a \,.\, \mathrm{e}^{-i}$$

(cf. (3.7)), where

$$a = 1 - \mathrm{e}^{-1} > 0$$

so that the condition (3.12) is satisfied for $\beta = \mathrm{e}^{-1}$ and a given above.

(Received June 21st, 1966.)

REFERENCES

[1] L. Baladová: Minimum of average conditional entropy for given minimum probability of error. Kybernetika *2* (1966), 5, 416–422.

[2] V. A. Kovalevskij: Pattern recognition problem from the viewpoint of mathematical statistics. In: Reading automata, Kiev 1965.

[3] M. Loéve: Probability Theory. D. Van Nostrand, New York 1960.

[4] A. Perez: Notions géneralisées d'incertitude, d'entropie et d'information du point de vue de la theorie de martingales. In Trans. of First Prague Conf. on Inf. Theory, Stat. Dec. Functions and Random Processes. Praha 1957.

[5] A. Perez: Information, ε-Sufficiency and Data Reduction Problems. Kybernetika *1* (1965), 4, 297–323.

[6] I. Vajda: Rate of convergence of the information in a sample concerning a parameter, Czechosl. Mathem. Journal (in print).

VÝTAH

O statistických rozhodovacích problémech s diskrétním parametrovým prostorem

Igor Vajda

V práci je stručně definován klasický model statistického rozhodování s abstraktním výběrovým prostorem a s nejvýše spočetným (diskretním) parametrovým prostorem a základní charakteristiky tohoto modelu: informace I kterou nám poskytne výběrová hodnota o parametru a Bayesovské riziko r.

Jedním z nejdůležitějších problémů, které v souvislosti s uvažovaným modelem přicházejí v úvahu je asymptotické chování rizika r a informace I při opakovaném pozorování s rozsahem výběru konvergujícím do nekonečna. Proto prvním cílem práce je poskytnout odhady veličin r a I pomocí jednodušších výrazů, kterých by pak bylo možno použít k vyšetření zmíněných asymptotických vlastností. Výsledky jsou obsaženy v § 2, Theorem 1 a 3. Tyto výsledky jsou pak v § 3 aplikovány na studium asymptotického chování r a I za předpokladu, že posloupnost výběrů je nezávislá (Theorem 4) a nezávislá stacionární (Theorem 4s) náhodná posloupnost. V § 3 jsou též ukázány třídy rozhodovacích problémů, pro které riziko r_n resp. informace I_n, příslušné rozsahu výběru n, jsou dány vztahy $r_n = o(\lambda^n)$, $I_n = H(\mu) - o(\lambda^n)$, kde $0 < \lambda < 1$, resp. $r_n = o(n^\alpha)$, $I_n = H(\mu) - o(n^\alpha)$, kde $\alpha < 0$ a kde $H(\mu)$ je entropie parametrového prostoru.

Dále, přestože informace nezávisí na ztrátové funkci, z intuitivního hlediska je jasné, že mezi r a I existuje pro dostatečně širokou třídu ztrátových funkcí jakýsi vztah v tom smyslu, že čím je informace větší, tím je riziko menší. Vyjasnění tohoto vztahu je velmi důležité, protože existuje celá řada statistických problémů, kdy potřebujeme znát r a nepotřebujeme znát optimální rozhodovací funkci, která jedině nám umožňuje stanovit r přímo a která se obvykle konstruuje velmi obtížně. Proto druhým cílem práce je přispět k vyjasnění tohoto vztahu (Theorem 2).

Ing. Igor Vajda, Ústav teorie informace a automatizace ČSAV, Praha 2, Vyšehradská 49.