KYBERNETIKA ČÍSLO 1, ROČNÍK 3/1967

Information-Theoretic Risk Estimates in Statistical Decision*

ALBERT PEREZ

In this paper we give some information-theoretical estimates of average and Bayes risk change in statistical decision produced by a modification of the probability law in action and, in particular, by reducing or enlarging the sample space as well as the parameter space σ-algebras. These estimates, expressed in terms of information growth or *generalized f-enrotpy* not necessarily of Shannon's type, are improved versions of the estimates we obtained in previous papers.

CONTENTS

0. INTRODUCTION

Following the line of investigation of our previous papers [1, 2, 3], we try in the present paper to further extend and improve some information-theoretical estimates of the average and Bayes risk change in a statistical decision problem implied by the change of the probability law in action and, in particular, by reducing or enlarging the sample and/or parameter space σ-algebra.

As underlined namely in paper [1], the need for such estimates is growing with the complexity of the decision problems encountered in many fields of application. Due to the boundedness of the available capabilities, theoretical and/or material, for decision making it is desirable and sometimes even unavoidable to reduce complex decision problems to more simple ones as compared to the state of these capabilities. In such situations the need arises of course for indirect

* This paper partly coincides with the author's communication *Decision Quality Improvement Versus Information Growth in a Statistical Decision Problem* presented on International Congress of Mathematicians, Moscow, 16th—26th August, 1966.

methods of attack allowing to perform rationally the process of simplification or reduction in order to pick out, if it exists, a good approximation or admissible version of the initial decision problem. By "admissible" we understand such a version that its solution and the realization of the corresponding decision procedure are possible in the frame of the existing capabilities and, at the same time, the change or loss of decision quality connected with this reduced version as compared to the unreduced one is still acceptable.

In face of such a task it is only natural for a worker in the field of information theory to get the idea that the methods of information theory suitably developed could be useful in solving this task for a growing variety of decision problems. Indeed, all the theorems of Shannon's type deal essentially with criteria of transmissibility of an information source through a communication channel, i.e. criteria of discernibility or decidability in a certain asymptotic sense of the transmitted messages on the base of the received signals. These criteria, formulated in information-theoretical terms, may of course inform in certain cases on the existence or non-existence of suitable encoding and decoding procedures allowing the transmission, but they say nothing about the structure of these procedures. What we need in our case is something analogue to these criteria permitting us to judge if a given reduced version of the initial decision problem is admissible or not. Our information-theoretic risk estimates are conceived just as a contribution in the above sense.

By introducing in [1] the information-theoretical concept of *ε-sufficiency* as a natural generalization of the concept of sufficiency used in mathematical statistics (to which it reduces for $\varepsilon = 0$) we observed that it would be possible to part from other definitions of information generalizing that of Shannon, namely, the *f-informations* or *f-divergences* of Rényi and Csiszár [4], which also have the property to be conserved only with respect to sufficient transformations (in the sense of mathematical statistics) and otherwise to decrease as well as some other fundamental properties similar to those of the Shannon's information (but not the well-known additivity properties of its density which are so important in proving limit theorems of the Shannon-McMillan's type [5]). And we added that in any case it would be interesting to try to improve our estimates by using the other definitions of information. To a certain extent this plan is realized in the present paper in combination with a suitable application of the Lagrange method of multipliers for obtaining certain constrained extremum (minimum) of the corresponding generalized f-entropy. Thus, it was possible, in particular, to improve some previous estimates of the average and Bayes risk change we obtained in [1, 2, 3] in terms of the corresponding Shannon's information change or in terms of the generalized entropy [6] of the initial probability law with respect to the modified one.

1. GENERALIZED f-ENTROPY

The information-theoretic risk estimates obtained in [1, 2, 3] are based to a great extent on certain properties of the *generalized Shannon's entropy* or *Kullback-Leibler's information* of a probability measure P with respect to a probability measure Q defined on the same measurable space $(Z, \mathfrak{Z})$, i.e. of the quantity*

$$(1.1) \qquad H_Q(P) = \int u \log u \, \mathrm{d}Q \quad \text{if} \quad P \ll Q\,, \qquad = +\infty \quad \text{otherwise}\,,$$

* In the sequel by "log" we understand the *natural logarithm*.

where $u = dP/dQ$ in the first equality is the Radon-Nikodym density of P with respect to Q. These properties are essentially due to the *convexity* of the function $f(u) = u \log u$ which assures for every probability measure μ on the Borel sets of the interval $[0, +\infty]$ with $\int_0^\infty u \, d\mu(u) < \infty$ the validity of the inequality

$$\int_0^\infty f(u) \, d\mu(u) \geqq f\left(\int_0^\infty u \, d\mu(u)\right) \tag{1.2}$$

of Jensen. If, moreover, the function $f(u)$ is strictly convex, as in the case above, then the sign of equality in (1.2) holds if, and only if, the measure μ is concentrated on the unique point

$$u_0 = \int_0^\infty u \, d\mu(u) .$$

To the convexity of $f(u) = u \log u$ are essentially also due the well-known semi-martingale properties of the stochastic process $\{f(u_n), \mathfrak{Z}_n; n \geqq 1\}$ where $\{u_n, \mathfrak{Z}_n; n \geqq 1\}$ is the martingale process of the Radon-Nikodym densities $u_n = dP/dQ(\mathfrak{Z}_n)$ corresponding to an increasing sequence $\mathfrak{Z}_1 \subset \mathfrak{Z}_2 \subset \ldots \subset \mathfrak{Z}$ of sub-σ-algebras of the σ-algebra $\mathfrak{Z}$ under the assumption that the restrictions P_n and Q_n of P and Q on $\mathfrak{Z}_n \subset \mathfrak{Z}$, $n \geqq 1$, satisfy the condition $P_n \ll Q_n$ of absolute continuity (cf. [6]).

Neglecting, thus, the additivity property and retaining only the convexity property of $f(u)$ one may expect to obtain useful generalizations of the Shannon's entropy. The realization of certain aspects of this program (the martingale aspect, for instance, is neglected) was systematically undertaken by Csiszár [4] which considerably generalized, namely, the work beginned in this direction by Rényi [7]. In the remainder of this section we shall often refer to [4] without making a special mention.

Definition of the generalized f-entropy. Let $f(u)$ be a continuous and convex function of u defined for $u \in (-\infty, \infty)$ resp. $u \in (0, \infty)$ with, by definition, $f(0) = \lim_{u \to 0} f(u)$ and $0 \, . \, f(0) = 0$.

Let $(Z, \mathfrak{Z})$ be a measurable space* and P and Q two probability measures on it. Let, further, C be the set of absolute continuity of P with respect to Q, the corresponding density being denoted by u.

By *generalized f-entropy* (*f-divergence*, according to [4]) we understand the quantity

$$H_f(P, Q) = \int_C f(u) \, dQ + P(Z - C) \, . \lim_{v \to \infty} \frac{f(v)}{v} \tag{1.3}$$

which in the case $P \ll Q$ becomes

$$H_f(P, Q) = \int_Z f(u) \, dQ . \tag{1.4}$$

* That is Z is a non-empty set and $\mathfrak{Z}$ is a σ-algebra of subsets of Z.

For $f(u) = u \log u$ the generalized f-entropy coincides with the generalized Shannon's entropy (1.1), i.e.

$$H_{u \log u}(P, Q) = H_Q(P) . \tag{1.5}$$

For

$$f(u) = \begin{cases} -u^a & \text{for} \quad 0 < a < 1 , \\ u^a & \text{for} \quad a > 1 \end{cases} \tag{1.6}$$

we obtain the *generalized entropies of order* a related to the well-known Rényi's relative informations of order a (cf. [7]) by the relation (we write H_a instead of H_{u^a} resp. H_{-u^a})

$$I_a(P \parallel Q) = \frac{1}{a-1} \log |H_a(P, Q)| . \tag{1.7}$$

It is possible to see that for $a \to 1$ the latter quantity converges to the generalized Shannon's entropy (1.1). Similarly for a converging to 1 from above we obtain

$$\lim_{a \searrow 1} \frac{H_a(P, Q) - 1}{a - 1} = H_Q(P) \tag{1.8}$$

and for a converging to 1 from below we obtain

$$\lim_{a \nearrow 1} \frac{H_a(P, Q) + 1}{1 - a} = H_Q(P) . \tag{1.9}$$

On the base of these relations it is possible by extension to conceive the generalized entropy $H_Q(P)$ as a relative information or *generalized entropy of order* 1 and denote it by $H_1(P, Q)$.

The total variation of P and Q may be also considered as a special case of the generalized f-entropy (1.3) obtained for $f(u) = |u - 1|$.

Let us now recall some important properties of the generalized f-entropy coinciding with those we obtained in [6] for the special case of the generalized Shannon's entropy.

It always holds

$$H_f(P, Q) \geqq f(1) = H_f(P, P) , \tag{1.10}$$

whereas for $f(u)$ *strictly convex* (i.e. with a graph not containing a linear segment) the sign of equality in the first inequality holds if, and only if, $P = Q$.

If, for P and Q given on the measurable space $(Z, \mathfrak{Z})$, $\mathscr{A} = (A_1, A_2, \ldots, A_n)$ is a finite measurable partition of $(Z, \mathfrak{Z})$ and $P_{\mathscr{A}} = (P(A_1), P(A_2), \ldots, P(A_n))$, $Q_{\mathscr{A}} = (Q(A_1), Q(A_2), \ldots, Q(A_n))$, then it holds

$$H_f(P, Q) = \sup_{\mathscr{A}} H_f(P_{\mathscr{A}}, Q_{\mathscr{A}}) . \tag{1.11}$$

Let now, besides $(Z, \mathfrak{Z})$ be given another measurable space $(Z', \mathfrak{Z}')$ and let $(\mathfrak{Z}, \nu_z, \mathfrak{Z}')$ be a *transmission channel* with input space $(Z, \mathfrak{Z})$, output space $(Z', \mathfrak{Z}')$ and transition probabilities given by ν_z. As known, for every $z \in Z$ it is, thus, ν_z a probability measure on $(Z', \mathfrak{Z}')$ and for every $E \in \mathfrak{Z}'$ is the function $\nu_z(E)$ a $\mathfrak{Z}$-measurable function of z.

Let, further, $\bar{P}$ and $\bar{Q}$ be the marginal distributions induced on $(Z', \mathfrak{Z}')$ by the channel $(\mathfrak{Z}, \nu_z, \mathfrak{Z}')$ and the distributions P and Q, respectively, given on $(Z, \mathfrak{Z})$, i.e.

$$\bar{P}(E) = \int_Z \nu_z(E)\, \mathrm{d}P(z)\,, \quad \bar{Q}(E) = \int_Z \nu_z(E)\, \mathrm{d}Q(z)\,. \tag{1.12}$$

Then, for every convex function $f(u)$ and arbitrary probability distributions P and Q on $(Z, \mathfrak{Z})$, it holds

$$H_f(P, Q) \geqq H_f(\bar{P}, \bar{Q})\,. \tag{1.13}$$

If, moreover, the function $f(u)$ is strictly convex, then the equality sign in the above inequality holds if, and only if, either $H_f(\bar{P}, \bar{Q}) = \infty$ or

$$\nu_z(\{z' : \bar{u}(z') = u(z)\}) = 1 \quad [Q] \tag{1.14}$$

where by $u(z)$ we denote the density of P with respect to Q on the set of absolute continuity of the first with respect to the second, and, otherwise, we put $u(z) = \infty$ and by $\bar{u}(z')$ we denote the corresponding function for the pair of $\bar{P}$ and $\bar{Q}$. (As usually, the notation $[Q]$ in (1.14) means that the respective relation holds everywhere on Z with the possible exception of a set belonging to $\mathfrak{Z}$ of measure Q equal to zero.)

From the above result (cf. Theorem 1 and Complement of [4]) it follows in particular:

1. If the channel $(\mathfrak{Z}, \nu_z, \mathfrak{Z}')$ is such that ν_{z_1} and ν_{z_2} are never mutually orthogonal (singular) whatever be $z_1 \in Z$ and $z_2 \in Z$, then the sign of equality in (1.13) holds only in the case $P = Q$, provided that the function $f(u)$ is strictly convex and that $H_f(\bar{P}, \bar{Q}) < \infty$.
2. If the channel $(\mathfrak{Z}, \nu_z, \mathfrak{Z}')$ is such that there exists a partition $\{E_z, z \in Z\}$ of the output measurable space $(Z', \mathfrak{Z}')$ with the property that $\nu_z(E_z) = 1$, $z \in Z$, (i.e. if the channel is noiseless) then $H_f(P, Q) = H_f(\bar{P}, \bar{Q})$ for every pair of probability measures P and Q.
3. If the channel $(\mathfrak{Z}, \nu_z, \mathfrak{Z}')$ is such that there exists a measurable transformation T from $(Z', \mathfrak{Z})$ to $(Z, \mathfrak{Z}')$ with the property that $\nu_z(\{Tz\}) = 1$, $z \in Z$, then $\bar{P}$ and $\bar{Q}$ in (1.12) coincide with the probability measures PT^{-1} and QT^{-1} induced on $(Z', \mathfrak{Z}')$ by P and Q, respectively through the transformation T and, thus, the inequality (1.13) may be written

$$H_f(P, Q) \geqq H_f(PT^{-1}, QT^{-1})\,. \tag{1.15}$$

If $f(u)$ is strictly convex and $H_f(PT^{-1}, QT^{-1}) < \infty$, the sign of equality in (1.15) takes place if, and only if, the transformation T is *sufficient* (in the sense of mathematical statistics) with respect to the system of measures $\{P, Q\}$.

4. If $\mathfrak{Z}' \subset \mathfrak{Z}$ is a sub-σ-algebra of $\mathfrak{Z}$ and P' and Q' the measures induced by P and Q on $\mathfrak{Z}'$ (i.e. the restrictions of P and Q on $\mathfrak{Z}' \subset \mathfrak{Z}$), then

$$H_f(P, Q) \geqq H_f(P', Q'), \tag{1.16}$$

the sign of equality taking place (for $f(u)$ strictly convex and $H_f(P', Q') < \infty$) if, and only if, $\mathfrak{Z}'$ is a sufficient σ-algebra (in the sense of mathematical statistics) with respect to the system of measures $\{P, Q\}$.

As to the martingale properties mentioned at the beginning of the present section, we shall not insist here.

2. EXTREMAL METHOD OF RISK ESTIMATION IN TERMS OF THE GENERALIZED f-ENTROPY

Let us consider, as in [1, 2, 3], a classical statistical decision problem Π with *parameter* (input) *measurable space* $(X, \mathfrak{X})$, *sample* (output) *measurable space* $(Y, \mathfrak{Y})$, *decision measurable space* $(D, \mathfrak{D})$, *probability law* P on the Cartesian product measurable space $(X \times Y, \mathfrak{X} \times \mathfrak{Y})$ of the input and output, and *weight* or *loss function* $w(x, d)$, $x \in X$, $d \in D$, supposed non-negative and $\mathfrak{X} \times \mathfrak{D}$-measurable. (This function serves as a measure of the "loss" implied by taking a decision d while x is the realized value of the parameter at the input). Finally, let $\mathscr{B}$ be the set of possible *decision functions* or decision procedures applicable in connection with this problem. They may be of the *pure* or *mixed* (*randomized*) type, i.e. either measurable transformations of the measurable sample space $(Y, \mathfrak{Y})$ to the decision space $(D, \mathfrak{D})$ or random transforms-channels of the type $(\mathfrak{Y}, P_{D/y}, \mathfrak{D})$ with input measurable space $(Y, \mathfrak{Y})$ and output measurable space $(D, \mathfrak{D})$, (cf. Section 1). In the latter case, to every sample value $y \in Y$ there corresponds in general not a single decision $d \in D$ but a probability distribution on $(D, \mathfrak{D})$, so that the final choice of the decision d is made randomly according to $P_{D/y}$.

As a consequence of the application of a decision function $b \in \mathscr{B}$ there is induced by P on the Cartesian product $(X \times D, \mathfrak{X} \times \mathfrak{D})$ a probability measure which will be denoted by Pb^{-1} or $\bar{P}$ (cf. (1.12)).

The *average risk* $r(\Pi, b)$ corresponding to the decision problem Π and to the decision function $b \in \mathscr{B}$ is given by

$$r(\Pi, b) = r(P, b) = \int_{X \times D} w(x, d)\, \mathrm{d}Pb^{-1} = \int_{X \times Y} w(x, b(y))\, \mathrm{d}P = r(b)\,. \tag{2.1}$$

Let, now, $\tilde{\Pi}$ be a new decision problem differing from the above decision problem Π only in what concerns the probability law in action: in the place of P we have now the probability distribution $\tilde{P}$ on $(X \times Y, \mathfrak{X} \times \mathfrak{Y})$. By applying to $\tilde{\Pi}$ the decision function* $\tilde{b} \in \tilde{\mathscr{B}}$ there is induced by $\tilde{P}$ on $(X \times D, \mathfrak{X} \times \mathfrak{D})$ a probability measure $\tilde{P}\tilde{b}^{-1} = \tilde{\bar{P}}$ and the corresponding average risk is given by

$$(2.2) \qquad r(\tilde{\Pi}, \tilde{b}) = r(\tilde{P}, \tilde{b}) = \int_{X \times D} w(x, d) \, \mathrm{d}\tilde{P}\tilde{b}^{-1} = \int_{X \times Y} w(x, \tilde{b}(y)) \, \mathrm{d}\tilde{P} = \tilde{r}(\tilde{b}) \, .$$

In Lemma 3.1 of [1] or [3] we proved, in particular, the following inequality for the average risks (2.1) and (2.2):

$$(2.3) \qquad \tilde{r}(\tilde{b}) - r(b) \leqq \sqrt{[2\tilde{r}(w^2, \tilde{b}) \, H_{\tilde{P}\tilde{b}^{-1}}(Pb^{-1})]}$$

where by $H_{\tilde{P}\tilde{b}^{-1}}(Pb^{-1})$ we denote, as in (1.1), the generalized Shannon's entropy of Pb^{-1} with respect to $\tilde{P}\tilde{b}^{-1}$ and by

$$(2.4) \qquad \tilde{r}(w^2, \tilde{b}) = \int w^2(x, d) \, \mathrm{d}\tilde{P}\tilde{b}^{-1}$$

the average risk corresponding to the decision problem $\tilde{\Pi}(w^2)$ resulting from $\tilde{\Pi}$ by only changing the weight function w to w^2, and to the decision function $\tilde{b}$.

This inequality served as a basis in papers [1, 2, 3] for obtaining upper estimates of the average risk change on passing from the decision problem Π to the decision problem $\tilde{\Pi}$ or conversely, namely, under different conditions concerning the choice of the decision functions b and $\tilde{b}$ applied in the two cases. We can, for instance, take $\tilde{b} = b$ or, more generally, $\tilde{b} = bT$, where T is a measurable one-to-one transformation of $(Y, \mathfrak{Y})$ onto itself conserving, thus, the information. We can also consider the Bayes risk change on passing from Π to $\tilde{\Pi}$. As known, by *Bayes risk* corresponding to the decision problem Π and to the related set of available decision functions $\mathscr{B}$ we understand the following quantity.

$$(2.5) \qquad r_0(\Pi, \mathscr{B}) = \inf_{b \in \mathscr{B}} r(\Pi, b) \, ,$$

where the average risk $r(\Pi, b)$ is defined by (2.1).

In face of such a task the direct method would be to solve in each case the corresponding decision problem. However, this method, if realizable at all, is not always economic to apply, so that every indirect method of estimation of the decision possibilities (i.e. of the decision quality attainable) before beginning to solve a decision problem is always desirable.

The extremal method of risk estimation in terms of the generalized f-entropy given in the present section serves as a better basis than inequality (2.3) for obtaining estimates of the above kind as we shall prove in the following sections.

* Note that, in general, the set $\tilde{\mathscr{B}}$ of decision functions related to the decision problem $\tilde{\Pi}$ may differ from the set $\mathscr{B}$ of decision functions related to the decision problem Π.

The general method consists to determine the minimal value of the generalized f-entropy (cf. (1.3) and (1.4)) of $Pb^{-1} = \bar{P}$ with respect to $\tilde{P}\tilde{b}^{-1} = \bar{\tilde{P}}$ introduced above given $\bar{\tilde{P}}$, the weight function $w(x, d)$ (and, thus, also the value $\tilde{r}$ of the average risk $r(\tilde{\Pi}, \tilde{b})$ through (2.2)) and the value r of the average risk $r(\Pi, b)$ defined by (2.1). In the sequel we shall suppose that the second derivative $f''(u)$ of the convex function $f(u)$, involved in the definition of the generalized f-entropy, exists and, thus, due to the *strict convexity* of f we suppose in the sequel,

$$f''(u) > 0 . \tag{2.6}$$

Further, we shall suppose here for the sake of simplicity that always the probability measure $\bar{P} = Pb^{-1}$ is absolutely continuous with respect to the given probability measure $\bar{\tilde{P}} = \tilde{P}\tilde{b}^{-1} : \bar{P} \ll \bar{\tilde{P}}$, the corresponding Random-Nikodym density being denoted by $\bar{u}$.

By an heuristic application of the *Lagrange multipliers method* for obtaining the extremum value of the generalized f-entropy

$$H_f(\bar{P}, \bar{\tilde{P}}) = \int f(\bar{u})\, \mathrm{d}\bar{\tilde{P}} \tag{2.7}$$

under the two constraints

$$\int \bar{u}\, \mathrm{d}\bar{\tilde{P}} = 1 \tag{2.8}$$

and

$$\int w(x, d)\, \bar{u}(x, d)\, \mathrm{d}\bar{\tilde{P}} = r \tag{2.9}$$

we find that the "minimizing density" $\bar{u}_0$ must satisfy the equation

$$f'(\bar{u}_0) = \lambda w + \mu \tag{2.10}$$

where the constants λ and μ are such that the conditions (2.8) and (2.9) are satisfied if one replaces $\bar{u}$ by the solution $\bar{u}_0$ of (2.10).

Note. It may happen that the "minimizing density" $\bar{u}_0$ obtains with positive probability $\bar{\tilde{P}}$ negative values, so that the signed measure $\bar{P}_0$ defined by $\mathrm{d}\bar{P}_0 = \bar{u}_0\, \mathrm{d}\bar{\tilde{P}}$ is not necessarily a probability measure and, as a consequence, the expression

$$\bar{H}_f^0 = \int f(\bar{u}_0)\, \mathrm{d}\bar{\tilde{P}} \tag{2.11}$$

may in some cases (for some combinations of f and w) not to be a "true" generalized f-entropy. However, even in such cases an interesting result may be expected since anyway the strong inequality

$$H_f(\bar{P}, \bar{\bar{P}}) > \bar{H}_f^0 \tag{2.12}$$

holds for every probability measure $\bar{P}$ giving an average risk $\int w \, \mathrm{d}\bar{P}$ equal to r, provided of course that the quantity $\bar{H}_f^0$ given by (2.11) is a minimum indeed, what we are going to prove.

Lemma 2.1. *Provided that there exists a function $\bar{u}_0$ satisfying simultaneously* (2.8), (2.9) *and* (2.10) *for suitable values of the constants λ and μ, the inequality*

$$H_f(\bar{P}, \bar{\bar{P}}) \geqq \bar{H}_f^0 = \int f(\bar{u}_0) \, \mathrm{d}\bar{\bar{P}} \tag{2.13}$$

holds for every probability measure $\bar{P} \ll \bar{\bar{P}}$ on $\mathfrak{X} \times \mathfrak{D}$ giving an average risk $\int w \, \mathrm{d}\bar{P}$ equal to r.

Proof. Let us denote by $\bar{u}$ the density of $\bar{P}$ with respect to $\bar{\bar{P}}$. Then it holds

$$\int \mathrm{d}\bar{P} = \int \bar{u} \, \mathrm{d}\bar{\bar{P}} = 1 \; ; \quad \int w \, \mathrm{d}\bar{P} = \int w\bar{u} \, \mathrm{d}\bar{\bar{P}} = r \tag{a}$$

and

$$\int \bar{u}_0 \, \mathrm{d}\bar{\bar{P}} = 1 \; ; \quad \int w\bar{u}_0 \, \mathrm{d}\bar{\bar{P}} = r \, . \tag{b}$$

Further, on account of the fact that for some $v \in \langle \bar{u}, \bar{u}_0 \rangle$

$$f(\bar{u}) - f(\bar{u}_0) = f'(\bar{u}_0)(\bar{u} - \bar{u}_0) + \tfrac{1}{2} f''(v)(\bar{u} - \bar{u}_0)^2 \tag{c}$$

based on the assumption of the existence of the second derivative of f, we can successively write

$$\begin{aligned} H_f(\bar{P}, \bar{\bar{P}}) - \bar{H}_f^0 &= \int [f(\bar{u}) - f(\bar{u}_0)] \, \mathrm{d}\bar{\bar{P}} = \\ &= \int f'(\bar{u}_0)(\bar{u} - \bar{u}_0) \, \mathrm{d}\bar{\bar{P}} + \tfrac{1}{2} \int f''(v)(\bar{u} - \bar{u}_0)^2 \, \mathrm{d}\bar{\bar{P}} = \\ &= \int (\lambda w + \mu)(\bar{u} - \bar{u}_0) \, \mathrm{d}\bar{\bar{P}} + \tfrac{1}{2} \int f''(v)(\bar{u} - \bar{u}_0)^2 \, \mathrm{d}\bar{\bar{P}} = \\ &= \lambda r + \mu - (\lambda r + \mu) + \tfrac{1}{2} \int f''(v)(\bar{u} - \bar{u}_0)^2 \, \mathrm{d}\bar{\bar{P}} = \\ &= \tfrac{1}{2} \int f''(v)(\bar{u} - \bar{u}_0)^2 \, \mathrm{d}\bar{\bar{P}} \geqq 0 \, . \end{aligned} \tag{2.14}$$

Here, the second equality follows from (c), the third equality follows from (2.10). the fourth equality follows from (a) and (b) and the last inequality from (2.6) expressing the (strict) convexity of f. Thus, the lemma is proved.

Lemma 2.2. *If the decision functions b and $\tilde{b}$ applied in the decision problems Π and $\tilde{\Pi}$ introduced above are related by the equality*

$$\tilde{b} = bT \tag{2.15}$$

where T is a measurable one-to-one transformation of the sample space $(Y, \mathfrak{Y})$ onto itself, then it holds

$$H_f(P, \tilde{P}T^{-1}) \geqq \bar{H}_f^0 \tag{2.16}$$

where $\bar{H}_f^0$ is given by (2.11) *and $H_f(P, \tilde{P}T^{-1})$ is the generalized f-entropy of P with respect to $\tilde{P}T^{-1}$, provided of course that the conditions of Lemma 2.1 are fulfilled.*

In particular, it holds

$$H_f(P, \tilde{P}) \geqq \bar{H}_f^0 \tag{2.17}$$

for $b = \tilde{b}$, where of course the decision functions b and $\tilde{b}$ may be of the pure or mixed (*randomized*) *type* (cf. Section 1).

Proof. The inequality (2.16) results from the inequality (2.13) of Lemma 2.1 on the base of the inequality (1.13) which here takes the form

$$H_f(P, \tilde{P}T^{-1}) \geqq H_f(\bar{P}, \bar{\tilde{P}}) \tag{2.18}$$

since, according to (2.15), we have: $\bar{P} = Pb^{-1}$ and $\bar{\tilde{P}} = \tilde{P}\tilde{b}^{-1} = \tilde{P}T^{-1}b^{-1}$.

As to the inequality (2.17), it is a special case of (2.16) obtained for $b = \tilde{b}$. Thus, the lemma is proved.

Lemma 2.3. *If in Lemma 2.2 we interchange the roles of b and $\tilde{b}$ by taking in the place of* (2.15) *the equality*

$$b = \tilde{b}T \tag{2.19}$$

then it holds

$$H_f(PT^{-1}, \tilde{P}) \geqq \bar{H}_f^0 \,. \tag{2.20}$$

Proof. It is completely similar to the proof of Lemma 2.2.

Theorem 2.1. *Let the decision functions b and $\tilde{b}$ applied in the decision problems Π and $\tilde{\Pi}$ introduced above be of the pure type* (*i.e. measurable transformations of the measurable sample space $(Y, \mathfrak{Y})$ to the measurable decision space $(D, \mathfrak{D})$ and satisfy the equality*

$$b = \tilde{b} \,. \tag{2.21}$$

Provided that there exists a function $u_0(x, y)$ defined on the Cartesian product $X \times Y$ of the parameter and sample space which is $\mathfrak{X} \times \mathfrak{Y}$-measurable and satisfies simultaneously the following three relations

$$\int u_0 \, \mathrm{d}\tilde{P} = 1 \,, \tag{2.22}$$

$$\int w(x, b(y)) \, u_0(x, y) \, \mathrm{d}\tilde{P} = r \,, \tag{2.23}$$

$$f'(u_0) = \lambda w + \mu \,, \quad (w = w(x, b(y))) \tag{2.24}$$

or suitable values of the constants λ and μ, then the inequality

$$H_f(P, \tilde{P}) \geqq H_f^0 = \int f(u_0) \, \mathrm{d}\tilde{P} \tag{2.25}$$

holds for every probability measure $P \ll \tilde{P}$ on $\mathfrak{X} \times \mathfrak{Y}$ giving an average risk $\int w(x, b(y)) \, \mathrm{d}P$ equal to r.

Moreover, it holds

$$H_f^0 = \bar{H}_f^0 = \int f(\bar{u}_0) \, \mathrm{d}\bar{\tilde{P}} = \int f(\bar{u}_0) \, \mathrm{d}\tilde{P}b^{-1} \tag{2.26}$$

where $\bar{u}_0$ is defined as in Lemma 2.1 (cf. (2.8)–(2.11)) *for $b = \tilde{b}$ of the above type.* (*As to the convex function f involved in the definition of the generalized f-entropy we suppose that it possesses a second derivative* (cf. (2.6))).

Proof. Inequality (2.25) may be obviously derived from the assumptions above in a completely analogue manner as inequality (2.13) in Lemma 2.1. It is sufficient to replace the entities represented by a letter with a bar by those without a bar and, moreover, $w(x, d)$ by $w(x, b(y))$ and $\bar{u}(x, d) = \mathrm{d}\bar{P}/\mathrm{d}\bar{\tilde{P}} = \mathrm{d}Pb^{-1}/\mathrm{d}\tilde{P}b^{-1}$ by $u(x, y) = \mathrm{d}P/\mathrm{d}\tilde{P}$ since here, by hypothesis, b and $\tilde{b}$ are of the pure type and identical according to (2.21).

As to the equality (2.26) of the constrained minimal "entropies" (which correspond to true generalized f-entropies if the respective "minimizing densities" $u_0(x, y)$ and $\bar{u}_0(x, d)$ may be chosen non-negative; cf. Note preceding Lemma 2.1), it results from the fact that here $u_0(x, y)$ may be chosen equal to $\bar{u}_0(x, b(y))$,

$$u_0(x, y) = \bar{u}_0(x, b(y)) \,, \tag{2.27}$$

since if the first satisfies simultaneously the three conditions (2.22), (2.23) and (2.24), then (under (2.21)) the second satisfies simultaneously the three conditions (2.8), (2.9) and (2.10) for the same values of the constants λ and μ, and conversely.

Thus, the theorem is proved.

An alternative extremal method of risk estimation in terms of the generalized f-entropy is the following.

The problem is formulated in a somewhat different way: *For* $\tilde{P}$ *and* $H_f(P, \tilde{P}) = h$ *given, to find* $P \ll \tilde{P}$ *maximizing the absolute difference* $|r(\tilde{P}, b) - r(P, b)| = |\tilde{r} - r|$ *of average risks* (for given w and b).

Applying heuristically as before the Lagrange multipliers method for obtaining the extremum value of the average risk

$$(2.28)\qquad r(P, b) = r = \int w(x, b(y))\,\mathrm{d}P = \int wu\,\mathrm{d}\tilde{P} \quad \left(u = \frac{\mathrm{d}P}{\mathrm{d}\tilde{P}}\right)$$

under the conditions (constraints)

$$(2.29)\qquad \int u\,\mathrm{d}\tilde{P} = 1$$

and

$$(2.30)\qquad \int f(u)\,\mathrm{d}\tilde{P} = h$$

we find (under the hypothesis that the second derivative of the (strictly) convex function f exists) that "the minimizing density" must satisfy the equation

$$(2.31)\qquad f'(u_0) = \lambda w + \mu \quad (w = w(x, b(y)))$$

where the constants λ and μ are such that the conditions (2.29) and (2.30) are satisfied if one replaces u by the solution u_0 of (2.31).

A similar note as that preceding Lemma 2.1 applies also here. The analogue of Theorem 2.1 is the following

Theorem 2.2 *Let the decision functions* b *and* $\tilde{b}$ *applied in the decision problems* Π *and* $\tilde{\Pi}$ *introduced above be of the pure type and satisfy the equality* $\tilde{b} = b$.

Provided that there exists a function $u_0(x, y)$ *defined on the Cartesian product* $X \times Y$ *of the parameter and sample space which is* $\mathfrak{X} \times \mathfrak{Y}$*-measurable and satisfies simultaneously* (2.29), (2.30) *and* (2.31) *for suitable values of the constants* λ *and* μ, *then the inequality*

$$(2.32)\qquad |r(\tilde{P}, b) - r(P, b)| = |\tilde{r} - r| \leqq |\tilde{r} - r_0| \quad \text{with} \quad r_0 = \int wu_0\,\mathrm{d}\tilde{P}$$

holds for every probability measure $P \ll \tilde{P}$ *on* $\mathfrak{X} \times \mathfrak{Y}$ *giving a generalized f-entropy with respect to* $\tilde{P}$ *equal to* h.

Proof. Let us denote by u the density of P with respect to $\tilde{P}$. Then it holds

$$(\mathrm{a})\qquad \int u\,\mathrm{d}\tilde{P} = 1\,; \quad \int f(u)\,\mathrm{d}\tilde{P} = h\,,$$

and

(b) $$\int u_0 \, \mathrm{d}\tilde{P} = 1 \; ; \quad \int f(u_0) \, \mathrm{d}\tilde{P} = h \, .$$

Further, on account of the fact that for some $v \in \langle u, u_0 \rangle$

(c) $$f(u) - f(u_0) = f'(u_0)(u - u_0) + \tfrac{1}{2} f''(v)(u - u_0)^2$$

or, for some $v' \in \langle 1, u_0 \rangle$,

(d) $$f(1) = f(u_0) + f'(u_0)(1 - u_0) + \tfrac{1}{2} f''(v')(1 - u_0)^2 \, ,$$

we can successively write

(e) $$0 = \int [f(u) - f(u_0)] \, \mathrm{d}\tilde{P} = \int f'(u_0)(u - u_0) \, \mathrm{d}\tilde{P} + \tfrac{1}{2} \int f''(v)(u - u_0)^2 \, \mathrm{d}\tilde{P} =$$
$$= \int (\lambda w + \mu)(u - u_0) \, \mathrm{d}\tilde{P} + \tfrac{1}{2} \int f''(v)(u - u_0)^2 \, \mathrm{d}\tilde{P} =$$
$$= \lambda \, . \, (r - r_0) + \tfrac{1}{2} \int f''(v)(u - u_0)^2 \, \mathrm{d}\tilde{P}$$

(cf. (a), (b), and (2.31)), or

(f) $$f(1) = \int f(u_0) \, \mathrm{d}\tilde{P} + \int f'(u_0)(1 - u_0) \, \mathrm{d}\tilde{P} + \tfrac{1}{2} \int f''(v')(1 - u_0)^2 \, \mathrm{d}\tilde{P} =$$
$$= h + \int (\lambda w + \mu)(1 - u_0) \, \mathrm{d}\tilde{P} + \tfrac{1}{2} \int f''(v')(1 - u_0)^2 \, \mathrm{d}\tilde{P} =$$
$$= h + \lambda(r - r_0) + \tfrac{1}{2} \int f''(v')(1 - u_0)^2 \, \mathrm{d}\tilde{P} \, .$$

Since by assumption $f''(v) > 0$, $\int f''(v)(u - u_0)^2 \, \mathrm{d}\tilde{P} = 0$ if, and only if, $u = u_0[\tilde{P}]$, which immediately implies that

$$r = \int wu \, \mathrm{d}\tilde{P} = \int wu_0 \, \mathrm{d}\tilde{P} = r_0 \, ,$$

or that (2.32) holds. Similarly, $\int f''(v')(1 - u_0)^2 \, \mathrm{d}\tilde{P} = 0$ if, and only if, $u_0 = 1[\tilde{P}]$, which implies $u = 1[\tilde{P}]$ (cf. (a) and (1.10)), or $\tilde{r} = r$.

Hence we can restrict ourselves to the case when simultaneously $\int f''(v)(u - u_0)^2 \, . \, \mathrm{d}\tilde{P} > 0$, $\int f''(v')(1 - u_0)^2 \, \mathrm{d}\tilde{P} > 0$; in this case however,

$$\lambda \, . \, (r - r_0) < 0 \quad \text{(cf. (e))}$$
$$\lambda \, . \, (\tilde{r} - r_0) < 0 \quad \text{(cf. (f), (1.10))} \, .$$

If, now, $\tilde{r} > r_0$ then $\lambda < 0$ and, thus, $r > r_0$ i.e. r_0 = minimum and $0 < \tilde{r} - r_0 =$ = maximum.

If, on the other hand, $\tilde{r} < r_0$, then $\lambda > 0$, and, thus, $r < r_0$ i.e. r_0 = maximum and, as a consequence, $0 < r_0 - \tilde{r}$ = maximum.

In other words, it always holds $|\tilde{r} - r| \leqq |\tilde{r} - r_0|$ and, thus, the theorem is proved.

In the following sections we shall apply the results of the present section and, namely, Theorems 2.1 and 2.2.

3. RISK ESTIMATES IN TERMS OF $H_1(P, \tilde{P})$

In this section we shall apply the extremal method of risk estimation developed in Section 2 to the special case of the *generalized Shannon's entropy* or *generalized entropy of order* 1 (cf. (1.1) and (1.6)–(1.9)), corresponding to $f(u) = u \log u$.

Placing us in the frame of Theorem 2.1, we find that the "minimizing density" u_0 must satisfy the following three conditions (corresponding to (2.22)–(2.24))

$$u_0 = e^{\lambda w + \mu} \quad (w = w(x, b(y))) , \tag{3.1}$$

$$\int e^{\lambda w + \mu} \, \mathrm{d}\tilde{P} = 1 , \tag{3.2}$$

$$\int w e^{\lambda w + \mu} \, \mathrm{d}\tilde{P} = r \tag{3.3}$$

for suitable values of the constants λ and μ, in order to assure the validity of the inequality

$$H_1(P, \tilde{P}) \geqq H_1^0 = \int u_0 \log u_0 \, \mathrm{d}\tilde{P} = \int (\lambda w + \mu) \, \mathrm{d}P_0 = \lambda r + \mu \tag{3.4}$$

corresponding to the statement of Theorem 2.1 given by (2.25).

In relation with the Note preceding Lemma 2.1, let us remark that here the "minimizing density" is non-negative, so that the signed measure P_0 defined by $\mathrm{d}P_0 = u_0 \, \mathrm{d}\tilde{P}$ is a probability measure and H_1^0 is, thus, a true generalized entropy of order 1,

$$H_1^0 = H_1(P_0, \tilde{P}) . \tag{3.5}$$

As a consequence, the sign of equality in (3.4) may really occur.

By assuming that λ and μ as functions of r are derivable, we obtain from (3.1) – (3.3) that

$$\begin{cases} \lambda'(r) = \dfrac{1}{r(w^2) - r^2} , \\[2ex] \mu'(r) = - \dfrac{r}{r(w^2) - r^2} \end{cases} \tag{3.6}$$

with

$$r(w^2) = \int w^2(x, b(y))\, \mathrm{d}P_0 \tag{3.7}$$

and, further,*

$$\frac{\mathrm{d}}{\mathrm{d}r} H_1^0(r) = \lambda(r) \quad \text{with} \quad \lambda(\tilde{r}) = \mu(\tilde{r}) = H_1^0(\tilde{r}) = 0\,. \tag{3.8}$$

By developing $H_1^0(r)$ according to the powers of $r - \tilde{r}$ we, thus, obtain for some $h \in \langle r, \tilde{r} \rangle$

$$H_1^0(r) = \tfrac{1}{2}\lambda'(\tilde{r})(r - \tilde{r})^2 + \tfrac{1}{6}\lambda''(h)(r - \tilde{r})^3\,. \tag{3.9}$$

On the base of (3.4), (3.6) and (3.9) we derive immediately the following result.

Theorem 3.1. *If*

$$(r - \tilde{r})\,\lambda''(h) \geqq 0 \quad \text{for} \quad h \in \langle r, \tilde{r} \rangle \tag{3.10}$$

then it holds

$$|\tilde{r} - r| \leqq \sqrt{\{2[\tilde{r}(w^2) - \tilde{r}^2]\, H_1(P, \tilde{P})\}} \tag{3.11}$$

where $\tilde{r}(w^2) = \int w^2(x, b(y))\, \mathrm{d}\tilde{P}$ *and* $\tilde{r} = \int w(x, b(y))\, \mathrm{d}\tilde{P}$, $r = \int w(x, b(y))\, \mathrm{d}P$.

Corollary 3.1. *Let the weight function* w *take the values* 0 *or* 1.
Then, for $r \leqq \tilde{r} \leqq \frac{1}{2}$ *it holds*

$$\tilde{r} - r \leqq \sqrt{[2\tilde{r}(1 - \tilde{r})\, H_1(P, \tilde{P})]}\,. \tag{3.12}$$

Proof. According to (3.6) and to the fact that here

$$r(w^2) - r^2 = r - r^2 = r(1 - r) \tag{a}$$

we obtain that

$$\lambda'(r) = \frac{1}{r(w^2) - r^2} = \frac{1}{r(1 - r)} \tag{b}$$

and, thus,

$$(r - \tilde{r})\,\lambda''(h) = \frac{(\tilde{r} - r)(1 - 2h)}{h^2(1 - h)^2} \geqq 0 \quad \text{for} \quad h \in \langle r, \tilde{r} \rangle \tag{c}$$

according to our assumption that $r \leqq \tilde{r} \leqq \frac{1}{2}$. Thus, the corollary is proved.

Note that this result is better than that obtained in [1] on the base of Lemma 3.1 (cf. (2.3)). However, a better result is contained in the following theorem.

* The first equality is valid for every $f(u)$ and not only for $f(u) = u \log u$.

Theorem 3.2. *Let the weight function w be uniformly bounded from above by w_0, i.e. $w(x, d) \leqq w_0$, $x \in X$, $d \in D$.*
Then it holds

$$H_1(P, \tilde{P}) \geqq \frac{1}{w_0}\left[r \log\frac{r}{\tilde{r}} + (w_0 - r)\log\frac{w_0 - r}{w_0 - \tilde{r}}\right]. \tag{3.13}$$

For the special case of a weight function w taking the values 0 or 1 (the corresponding average risk is the so-called probability of error) we obtain the inequality

$$H_1(P, \tilde{P}) \geqq r \log\frac{r}{\tilde{r}} + (1 - r)\log\frac{1 - r}{1 - \tilde{r}}, \tag{3.14}$$

where the sign of equality may also take place.

Proof. According to (3.6) and to the fact that here

$$r(w^2) - r^2 \leqq w_0 r - r^2 = r(w_0 - r) \tag{a}$$

we obtain that

$$\lambda'(r) = \frac{1}{r(w^2) - r^2} \geqq \frac{1}{r(w_0 - r)}. \tag{b}$$

On the base of (3.8) we obtain by a two-fold integration of (b) the inequality

$$H_1^0(r, \tilde{r}) = H_1(P_0, \tilde{P}) \geqq \frac{1}{w_0}\left[r \log\frac{r}{\tilde{r}} + (w_0 - r)\log\frac{w_0 - r}{w_0 - \tilde{r}}\right] \tag{3.15}$$

for the corresponding minimal generalized entropy of order 1.

The inequality (3.13) follows immediately from (3.15).

On the other hand, if w is of the special type "0 or 1", the sign of equality holds in (a) by taking $w_0 = 1$ and, thus, the sign of equality holds also in (3.15) for $w_0 = 1$. As a consequence, (3.14) holds where the sign of equality is also possible.

Thus, the theorem is proved.

Let, under the latter assumptions concerning w, $\tilde{r}$ be the risk corresponding to the distribution $\tilde{P}$ on $\mathfrak{X} \times \mathfrak{Y}$ and let us assume that X is a discrete (i.e. finite or countable) set and that $\mathfrak{X}$ contains all subsets of X. If we define a new decision problem with sample space $(Y \times X, \mathfrak{Y} \times \mathfrak{X})$ and with a probability distribution P generated on $\mathfrak{X} \times (\mathfrak{Y} \times \mathfrak{X})$ by $P_X = \tilde{P}_X$ (marginal cf $\tilde{P}$ on $\mathfrak{X}$) and by

$$P_{Y \times X|x}(E) = \tilde{P}_{Y|x} \times \tilde{P}_{X|x}(E), \quad E \in \mathfrak{Y} \times \mathfrak{X}, \tag{3.16}$$

where $\tilde{P}_{X|x}$ on $\mathfrak{X}$ is defined by $\tilde{P}_{X|x}(F) = \chi_F(x)$, $F \in \mathfrak{X}$, where χ_F is the characteristic function of the set F and where $\tilde{P}_{Y|x}$ or $P_{Y \times X|x}$ is the conditional distribution on $\mathfrak{Y}$ or $\mathfrak{Y} \times \mathfrak{X}$ given x corresponding to $\tilde{P}$ or P respectively. Under this assumptions it is to see that the rick r corresponding to P is equal to 0 and, moreover, that $\tilde{P}$ is a restriction of P on $\mathfrak{Y} \subset \mathfrak{Y} \times \mathfrak{X}$ and that the marginal measures P_X, $\tilde{P}_X$ of P, $\tilde{P}$ on $\mathfrak{X}$ coincide. Hence, according to Corollary 4.1 of [1], there exists a probability distribution $\tilde{\tilde{P}}$ on $\mathfrak{X} \times (\mathfrak{Y} \times \mathfrak{X})$ satisfying the following relations

$$I(\tilde{\tilde{P}}) = I(\tilde{P}), \tag{3.17}$$

$$H_1(P, \tilde{\tilde{P}}) = I(P) - I(\tilde{P}),$$

where $I(\cdot)$ denotes the Shannon's information (cf. (2.5) of [1]). The first equality implies $\tilde{\tilde{r}} = r(\tilde{\tilde{P}}) = \tilde{r} = r(\tilde{P})$ (cf. Theorem 4.3 of [1]) so that, in view of (3.14), we can write for w of the type "0 or 1" that

$$\tilde{r} \leqq 1 - e^{-H_1(P, \tilde{\tilde{P}})} = 1 - e^{-(I(P) - I(\tilde{P}))}, \tag{3.18}$$

whereas $r(P) = 0$ implies that $I(P) = H(P_X) = H(\tilde{P}_X)$, where $H(\cdot)$ denotes the ordinary Shannon's entropy of the parameter space X. Hence we can conclude that

$$\tilde{r} \leqq 1 - e^{-\tilde{h}} \tag{3.19}$$

where $\tilde{h} = H(\tilde{P}_X) - I(\tilde{P}) = \tilde{H}(X \mid Y)$ is the so-called *equivocation* (i.e. the average conditional entropy of the parameter given the sample value) corresponding to the probability law $\tilde{P}$.

The value $\tilde{h}_0 = \log(1 - \tilde{r})$ of the equivocation $\tilde{h}$ for a given value of the average risk (probability of error) equal to $\tilde{r}$, coincides on the points $\tilde{r} = (m - 1)/m$, $m = 1, 2, \ldots$, with that found by V. A. Kovalenskij [8] who has established the following expression for the minimal equivocation given the probability of error $\tilde{r}$:

$$\tilde{h}_{\min}(\tilde{r}) = \log m + m(m + 1) \log \left(\frac{m + 1}{m}\right)\left(\tilde{r} - \frac{m - 1}{m}\right) \tag{3.20}$$

for

$$\frac{m - 1}{m} \leqq \tilde{r} < \frac{m}{m + 1}, \quad m = 1, 2, \ldots$$

Theorem 3.3. *If the dominating measure $\tilde{P}$ is Gaussian, if the weight function $w(x, d) = (x - d)^2$ (the parameter and decision measurable spaces coinciding with the real Borel line) and if the decision function b is linear, then it holds*

$$H_1(P, \tilde{P}) \geqq \frac{1}{2}\left[\frac{r}{\tilde{r}} - 1 - \log \frac{r}{\tilde{r}}\right] \geqq 0 \tag{3.21}$$

where $r = \int w(x, b(y))\, dP$ and $\tilde{r} = \int w(x, b(y))\, d\tilde{P}$.

Proof. According to (3.6) and to the fact that here (cf. [3])

$$r(w^2) - r^2 \leqq 2r^2 \tag{a}$$

for $r(w^2) = \int w^2(x, b(y))\, dP_0$ and $r = \int w(x, b(y))\, dP_0$ since our assumptions imply that the minimizing probability measure P_0 is also Gaussian, we obtain that

$$\text{(b)} \qquad \lambda'(r) = \frac{1}{r(w^2) - r^2} \geqq \frac{1}{2r^2}\,.$$

On the base of (3.8) we obtain by a two-fold integration of (b) the inequality

$$\text{(3.22)} \qquad H_1^0(r, \tilde{r}) \geqq \frac{1}{2}\left[\frac{r}{\tilde{r}} - 1 - \log\frac{r}{\tilde{r}}\right] \geqq 0$$

for the corresponding minimal generalized entropy of order 1.

The inequality (3.21) follows immediately from (3.22).

Thus, the theorem is proved.

By applying a completely similar procedure as that used in papers [1, 2, 3] it is possible to deduce from the above inequalities, concerning the average risks, corresponding inequalities relative to the Bayes risks and, namely, to the Bayes risk increase resulting by a reduction of the sample space σ-algebra as well as of the parameter space σ-algebra. The obtained estimates are, in general better. However, we shall not insist on these questions in the present paper.

4. RISK ESTIMATES IN TERMS OF $H_a(P, \tilde{P})$

As in Section 3, we shall in this section apply the extremal method of risk estimation developed in Section 2 to another special case and, namely, to the case of the *generalized entropy of order* $a(0 < a \neq 1)$ introduced in Section 1 (cf. (1.6)), i.e.

$$\text{(4.1)} \qquad H_a(P, \tilde{P}) = \int u^a\, d\tilde{P} \quad \text{for} \quad a > 1\,,$$

$$H_a(P, \tilde{P}) = -\int u^a\, d\tilde{P} \quad \text{for} \quad 0 < a < 1\,,$$

where we suppose $P \ll \tilde{P}$ with $u = dP/d\tilde{P}$.

As stated in Section 1, the relations (1.8) and (1.9) exist between them and the generalized Shannon's entropy or generalized entropy of order 1.

Placing us again in the frame of Theorem 2.1, we find that the "minimizing density" u_0 must satisfy here the following three conditions (corresponding to (2.22)–(2.24))

$$\text{(4.2)} \qquad u_0^{a-1} = \lambda w + \mu \quad (w = w(x, b(y)))\,,$$

$$\text{(4.3)} \qquad \int (\lambda w + \mu)^{1/(a-1)}\, d\tilde{P} = 1\,,$$

$$\text{(4.4)} \qquad \int w(\lambda w + \mu)^{1/(a-1)}\, d\tilde{P} = r$$

for suitable values of the constants λ and μ, in order to assure the validity of the inequality

$$(4.5) \qquad H_a(P, \tilde{P}) \geqq H_a^0 = \operatorname{sign}(a-1)\int u_0^a \, \mathrm{d}\tilde{P} = \operatorname{sign}(a-1)\int u_0^{a-1}\, \mathrm{d}P_0 =$$

$$= \operatorname{sign}(a-1)\int (\lambda w + \mu)\, \mathrm{d}P_0 = \operatorname{sign}(a-1)(\lambda r + \mu)$$

corresponding to the statement of Theorem 2.1 given by (2.25).

As it concerns the Note preceding Lemma 2.1, let us remark that in the present case it may happen that the signed measure P_0 defined by $\mathrm{d}P_0 = u_0\, \mathrm{d}\tilde{P}$ is not a true probability measure. However, even in such a case the strong inequality (4.4) holds.

Special case: $a = 2$. In the case of the *generalized entropy of order 2* we can obtain explicitly the "minimizing density":

$$(4.6) \qquad u_0 = 1 + \frac{(\tilde{r} - r)(\tilde{r} - w)}{\tilde{r}(w^2) - \tilde{r}^2}$$

and, thus, the following result holds.

Theorem 4.1. *The following inequality holds*

$$(4.7) \qquad |\tilde{r} - r| \leqq \sqrt{\{[\tilde{r}(w^2) - \tilde{r}^2] \cdot [H_2(P, \tilde{P}) - 1]\}}$$

where

$$\tilde{r}(w^2) = \int w^2(x, b(y))\, \mathrm{d}\tilde{P}\,, \quad \tilde{r} = \int w(x, b(y))\, \mathrm{d}\tilde{P}\,, \quad r = \int w(x, b(y))\, \mathrm{d}P\,.$$

Special case of w of the type "0 or 1". If the weight function w takes the values 0 or 1, then the following theorem holds in terms of the generalized entropy of order $0 < a \neq 1$.

Theorem 4.2. *Let w be of the type "0 or 1". Then the inequality*

$$(4.8) \qquad H_a(P, \tilde{P}) \geqq H_a^0 = \left[\tilde{r}\left(\frac{r}{\tilde{r}}\right)^a + (1 - \tilde{r})\left(\frac{1-r}{1-\tilde{r}}\right)^a\right] \operatorname{sing}(a-1)$$

holds. The "minimizing density" u_0 coincides with that of H_1^0.

Proof. From (4.2)–(4.4) we derive easily that for such a w

$$(4.9) \qquad \left|\begin{array}{l} \lambda = \left(\dfrac{r}{\tilde{r}}\right)^{a-1} - \left(\dfrac{1-r}{1-\tilde{r}}\right)^{a-1}, \\[2ex] \mu = \left(\dfrac{1-r}{1-\tilde{r}}\right)^{a-1}, \end{array}\right.$$

and, thus, from (4.5) the inequality (4.8) follows immediately.

On the other hand, from (4.2) we obtain on the base of (4.9) that

$$u_0 = \left(\frac{r}{\tilde{r}}\right)^w \left(\frac{1 - r}{1 - \tilde{r}}\right)^{1-w} \tag{4.10}$$

whatever be $a > 0$. Thus, the theorem is proved.

(Received October 10th, 1966.)

REFERENCES

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

[2] Perez A.: Information and ε-Sufficiency. Paper No. 41 presented at the 35th Session of the International Statistical Institute, Beograd, 1965.

[3] Perez A.: Information Theory Methods in Reducing Complex Decision Problems. To appear in: Transactions of the Fourth Prague Conference on Information Theory, Statistical Decision Problems and Random Processes (1965).

[4] Csiszár I.: Eine informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markoffschen Ketten. Publications of the Mathematical Institute of the Hungarian Academy of Sciences *VIII* (1963), Series A, Fasc. 1—2, 85—108.

[5] Perez A.: Extensions of the Shannon-McMillan's Limit Theorem to more general Stochastic Processes. In: Transactions of the Third Prague Conference on Information Theory, Statistical Decision Functions and Random Processes (1962), Prague 1964, 545—574.

[6] Perez A.: Notions généralisées d'incertitude, d'entropie et d'information du point de vue de la théorie des partingales. In: Transactions of the First Prague Conference on Information Theory, Statistical Decision Functions and Random Processes (1956), Prague 1957, 183—208.

[7] Rényi A.: On measures of entropy and information. In: Proceedings of the 4th Berkeley Symposium on Probability and Statistics, I, Berkeley, 1960, 547—561.

[8] Kovalevskij V. A.: The problem of pattern recognition from the point of view of mathematical statistics. In: Reading Automata, Kiev 1965, 8—37 (in Russian).

VÝTAH

Informačně-teoretické odhady rizika ve statistickém rozhodování

Albert Perez

Článek navazuje na autorovy publikace [1, 2, 3] s cílem dále rozšířit a zlepšit některé informačně-teoretické odhady změny středního a Bayesova rizika spojené se změnou působícího pravděpodobnostního zákona v daném rozhodovacím problému, a zejména spojené s redukcí nebo rozšířením σ-algeber výběrového nebo parametrového prostoru. Potřeba takových odhadů roste úměrně se složitostí rozhodovacích problémů, se kterými se setkáváme čím dále tím více v mnoha aplikačních oblas-

tech. Zatímco odhady v [1, 2, 3] byly vyjádřeny pomocí zobecněné Shannonovské entropie (viz (1.1)), zde jsou odvozené odhady založeny na širším pojmu *zobecněné f-entropie* (*f-divergence*, podle [4]; viz (1.3) a (1.44), kde f je spojitá *konvexní* funkce, ne nutně typu $f(u) = u \log u$.

V části 1 (*Zobecněná f-entropie*) jsou připomenuty některé základní vlastnosti zobecněné f-entropie, které jsou založeny na konvexnosti funkce f a které jsou využité v dalším.

V části 2 (*Extremální metoda odhadu rizika v termínech zobecněné f-entropie*) se heuristickou aplikací metody Lagrangeových multiplikátorů dosahuje podmíněného extrému (minima) zobecněné f-entropie $H_f(P, \tilde{P})$ resp. podmíněného extrému (maxima) absolutní změny rizika $|\tilde{r} - r|$ (viz zejména Lemma 2.1 a Teoremy 2.1 a 2.2).

V části 3 (*Odhady rizika v termínech* $H_1(P, \tilde{P})$) se extremální metoda odhadu rizika z části 2 aplikuje ve speciálním případě zobecněné Shannonovské entropie nebo zobecněné entropie prvního řádu (viz (1.1) a (1.6)–(1.9)), která odpovídá $f(u) = u \log u$. Jako typické příklady lze uvést nerovnosti (3.12), (3.13) a (3.21).

V části 4 (*Odhady rizika v termínech* $H_a(P, \tilde{P})$) se extremální metoda odhadu rizika aplikuje v speciálním případě zobecněné entropie a-tého řádu ($0 < a \neq 1$) (viz (4.1)). Jako typické příklady lze zde uvést nerovnosti (4.7) a (4.8).

Dr. Albert Perez, DrSc., Ústav teorie informace a automatizace ČSAV, Vyšehradská 49, Praha 2.