ON SOME BASIC PROBLEMS OF STATISTICS FROM THE POINT OF VIEW OF INFORMATION THEORY

ALFRÉD RÉNYI
MATHEMATICAL INSTITUTE OF THE HUNGARIAN ACADEMY OF SCIENCES

1. Introduction

Many problems of mathematical statistics consist in that one has to extract from certain observations the information which is needed. In other words, one has to separate the relevant information from the irrelevant. For instance, it is a generally accepted view that if for a parameter there exists a sufficient statistic, its sufficiency means that it contains all the information which is present in the sample and is relevant for determining the parameter. Although this is generally admitted, it is not usual to go a step further and ask: *how much information is contained in a statistic (sufficient or not) concerning a parameter?* (According to the author's knowledge, the first to consider this question was D. V. Lindley [1].) In view of the success of information theory in other fields (especially in the theory of information-transmission), which success was achieved by attributing a numerical measure to amounts of information, this question is a very natural one. It seems to the author that the reason why this question is usually not asked in current statistical practice is that a meaningful answer to this question can be given only if one accepts the Bayesian point of view; that is, if one considers the unknown parameter as a random variable and attributes to it a prior distribution. As a matter of fact, the amount of information in a random variable concerning another random variable is a well-defined concept of information theory, whereas the amount of information in a random variable concerning a constant is always zero. The amount of information in a random variable ξ concerning another random variable θ is equal to the average decrease of uncertainty (entropy) concerning θ which results if ξ is observed. In order to measure this decrease of uncertainty, our prior knowledge about θ has to be taken into account.

If we have some prior knowledge about θ, this causes no difficulty. If we have no prior knowledge about θ, except that we know the set of its possible values, from the point of view of information theory it seems to be natural to attribute to θ that prior distribution on the admissible set of values which has the largest entropy, that is which corresponds to maximal uncertainty. Even if the parameter is in reality a constant, if its value is unknown to us, I do not find any logical fault in attributing to θ a prior distribution, if this is needed to compare

different statistics and to choose that which is the best for our purposes. Usually the choice depends only weakly on the prior distribution.

In this paper we do not want to go into the philosophical aspects of the question—however interesting they may be—as our aim is to deal with certain purely mathematical problems which arise when one tries to apply the concepts of information theory to the mentioned statistical problems.

The problem which will be discussed in what follows is: *how can one decide whether or not a certain sequence of observations contains all the information which is needed* (for example, to find the true value of the parameter)? In general, if a statistician is confronted with a concrete problem, his first task is to decide whether or not the required information is fully present. If the answer is positive, then the second step consists in trying to find an appropriate decision procedure, for instance one which is optimal in some respect. However, if the answer to the first question is negative, then it is futile to take the second step, as even the "best" decision procedure will not yield the required information. In this case the statistician has to look for some other source of information, that is, make some other observations.

In the present paper we shall discuss from this point of view the following problem. Let us be given an infinite sequence $\{\xi_n\}$, $(n = 1, 2, \cdots)$ of observations. We suppose that the distributions of the random variables ξ_n, $(n = 1, 2, \cdots)$ depend on a parameter θ, whose set of possible values is *finite*. We suppose further that for each fixed value of θ the random variables ξ_n, $(n = 1, 2, \cdots)$ are independent, but in general do not have the same distribution. We shall be especially interested in the case where the amount of information on θ contained in the observation ξ_n decreases when n increases. Such problems are often encountered. Let us imagine for instance that an event $\mathcal{E}$ happened at time $t = 0$, and ξ_n is the value of some quantity connected with the aftereffects of this event measured by some instrument at time $t = n$. Usually as time passes, the aftereffects of the event $\mathcal{E}$ become weaker and weaker, and thus as n increases, ξ_n gives gradually less and less information on the event $\mathcal{E}$.

In section 2 we shall consider the amount of information on θ which is still missing after having observed the values $\xi_1, \xi_2, \cdots, \xi_n$ and compare it with the error of the "standard" decision, consisting in deciding always in favor of the hypothesis which has the largest posterior probability.

In section 3 we give an upper bound for the amount of missing information (theorem 2) and give a necessary and sufficient condition for the convergence to 0 of this quantity for n tending to $+\infty$.

In section 4 we shall discuss some special cases, some of which have been discussed in previous papers of the author (see [2], [3], [4]), and also others which are presented here for the first time. In section 5 we compare some of our results with a theorem of S. Kakutani [5].

For the sake of brevity we deal in detail only with the case where θ may have only two values. The generalization to the case when the set of possible values of θ is an arbitrary *finite* set is quite straightforward and presents no difficulty. It

should be added, however, that our results cannot be generalized immediately to the case when the set of possible values of θ is infinite. This case presents some peculiar difficulties. We hope to return to this question in another paper.

2. An inequality between the amount of missing information and the error of the standard decision

Let $\{\xi_n\}$, $(n = 1, 2, \cdots)$ be a sequence of random variables. Let us suppose that the distribution of ξ_n depends on a parameter θ, which may take on two different values θ_0 and θ_1. We suppose that the random variables ξ_n are independent under the condition $\theta = \theta_0$, as well as under the condition $\theta = \theta_1$. We suppose further (this restriction is made only to simplify notations) that all the distributions in question are absolutely continuous. Let $f_n(x)$ and $g_n(x)$ denote the density functions of ξ_n under the conditions $\theta = \theta_0$ and $\theta = \theta_1$, respectively. As regards θ we suppose that it is a random variable, taking on the values θ_0 and θ_1 with the corresponding positive probabilities W_0 and W_1, $(W_0 + W_1 = 1)$.

These suppositions can be formulated as follows. Let $[\Omega, \mathfrak{A}, P] = S$ be a probability space. Let us be given a partition of Ω into two $\mathfrak{A}$-measurable sets Ω_0 and $\Omega_1 = \Omega - \Omega_0$, with $P(\Omega_0) = W_0$, $P(\Omega_1) = W_1 = 1 - W_0$. Let S_0 and S_1 denote the probability spaces $S_0 = [\Omega, \mathfrak{A}, P_0]$ and $S_1 = [\Omega, \mathfrak{A}, P_1]$ where $P_0(A) = (P(A\Omega_0)/P(\Omega_0))$ and $P_1(A) = (P(A\Omega_1)/P(\Omega_1))$ for $A \in \mathfrak{A}$. Let $\xi_n = \xi_n(\omega)$, $(\omega \in \Omega)$, $(n = 1, 2, \cdots)$ be a sequence of $\mathfrak{A}$-measurable real functions. Then we can consider the ξ_n as random variables on the probability space S as well as on the probability spaces S_0 and S_1. We suppose that the random variables ξ_n are independent on S_0 as well as on S_1. Note that on the probability space S the random variables ξ_n are usually not independent.

Let us suppose now that we observe the values of all the variables ξ_n, and we want to decide on this basis whether $\theta = \theta_0$ or $\theta = \theta_1$. In other words, we suppose that the sequence of values $\xi_n(\omega^*)$, $(n = 1, 2, \cdots)$ is given for a single unknown $\omega^* \in \Omega$, and want to decide, whether $\omega^* \in \Omega_0$ or $\omega^* \in \Omega_1$. We especially want to find out under what conditions on the density functions $f_n(x)$ and $g_n(x)$ can a correct decision be made for almost all $\omega^* \in \Omega$ (with respect to the measure P)?

Let us denote by ζ_n the random n-dimensional vector with components $(\xi_1, \xi_2, \cdots, \xi_n)$. Let I_n denote the amount of information contained in ζ_n concerning θ. Then we have

$$I_n = H(\theta) - E(H(\theta|\zeta_n)), \tag{2.1}$$

where

$$H(\theta) = W_0 \log \frac{1}{W_0} + W_1 \log \frac{1}{W_1} \tag{2.2}$$

is the entropy of the random variable θ, and $H(\theta|\zeta_n)$ is the conditional entropy of θ given ζ_n, that is

$$H(\theta|\zeta_n) = P(\theta = \theta_0|\zeta_n) \log \frac{1}{P(\theta = \theta_0|\zeta_n)} + P(\theta = \theta_1|\zeta_n) \log \frac{1}{P(\theta = \theta_1|\zeta_n)}, \tag{2.3}$$

and $E(H(\theta|\zeta_n))$ denotes the expectation of the random variable $H(\theta|\zeta_n)$. Here and in what follows log always denotes logarithm with base 2.

According to our supposition and the Bayes' theorem, one has

$$P(\theta = \theta_0|\zeta_n) = \frac{W_0 f_1(\xi_1) f_2(\xi_2) \cdots f_n(\xi_n)}{W_0 f_1(\xi_1) \cdots f_n(\xi_n) + W_1 g_1(\xi_1) \cdots g_n(\xi_n)} \tag{2.4}$$

and similarly,

$$P(\theta = \theta_1|\zeta_n) = \frac{W_1 g_1(\xi_1) g_2(\xi_2) \cdots g_n(\xi_n)}{W_0 f_1(\xi_1) \cdots f_n(\xi_n) + W_1 g_1(\xi_1) \cdots g_n(\xi_n)}. \tag{2.5}$$

For the sake of brevity, we introduce the notations $x^{(n)} = (x_1, \cdots, x_n)$, $\varphi_n(x^{(n)}) = f_1(x_1) \cdots f_n(x_n)$, and $\psi_n(x^{(n)}) = g_1(x_1) \cdots g_n(x_n)$.

With these notations, letting $\chi_n(x^{(n)}) = W_0\varphi_n(x^{(n)}) + W_1\psi_n(x^{(n)})$ we have

$$E(H(\theta|\zeta_n)) = \int_{X_n} \left[W_0\varphi_n(x^{(n)}) \log \frac{\chi_n(x^{(n)})}{W_0\varphi_n(x^{(n)})} + W_1\psi_n(x^{(n)}) \log \frac{\chi_n(x^{(n)})}{W_1\psi_n(x^{(n)})} \right] dx^{(n)} \tag{2.6}$$

where X_n is the n-dimensional Euclidean space and $dx^{(n)}$ stands for $dx_1 dx_2 \cdots dx_n$.

The quantity $E(H(\theta|\zeta_n))$ may be interpreted as *the amount of missing information on θ after observing ζ_n.*

It is easy to see that I_n is nondecreasing for $n = 1, 2, \cdots$ and $I_n \leq H(\theta)$. Thus $\lim_{n\to+\infty} I_n = I^*$ always exists. If $I^* = H(\theta)$, we shall say that *the sequence of observations $\{\xi_n\}$, $(n = 1, 2, \cdots)$ gives us full information on θ*, whereas in the case $I^* < H(\theta)$ we shall say that the observations $\{\xi_n\}$ do not give full information on θ.

Clearly, the most natural decision after having observed ζ_n is to accept θ_0 if $P(\theta_0|\zeta_n) > P(\theta_1|\zeta_n)$ and to accept θ_1 if $P(\theta_1|\zeta_n) > P(\theta_0|\zeta_n)$, and if $P(\theta_0|\zeta_n) = P(\theta_1|\zeta_n)$ to make a random choice between θ_0 and θ_1 with probabilities W_0 and W_1. We shall call this the *standard decision.* Let us define the random variable $\Delta_n = \Delta_n(\xi_n)$ as follows:

$$\Delta_n = \begin{cases} \theta_0 & \text{if the standard decision means acceptance of } \theta_0, \\ \theta_1 & \text{if the standard decision means acceptance of } \theta_1. \end{cases} \tag{2.7}$$

The *error ϵ_n of the standard decision* after taking n observations is defined as the probability of the standard decision being false. We have clearly

$$\epsilon_n = P(\Delta_n \neq \theta) = W_0 P(\Delta_n = \theta_1|\theta = \theta_0) + W_1 P(\Delta_n = \theta_0|\theta = \theta_1), \tag{2.8}$$

where $P(A|B)$ denotes the conditional probability of the event A under condition B. Obviously, $\Delta_n = \theta_0$ if $(\varphi_n(\zeta_n)/\psi_n(\zeta_n)) > (W_1/W_0)$ and $\Delta_n = \theta_1$ if $(\varphi_n(\zeta_n)/\psi_n(\zeta_n)) < (W_1/W_0)$. Thus

$$\epsilon_n = W_0 \int_{A_n} \varphi_n(x^{(n)})\, dx^{(n)} + W_1 \int_{B_n} \psi_n(x^{(n)})\, dx^{(n)} \tag{2.9}$$

where

$$A_n = \left\{\frac{\psi_n(x^{(n)})}{\varphi_n(x^{(n)})} \geq \frac{W_0}{W_1}\right\} \quad \text{and} \quad B_n = \left\{\frac{\psi_n(x^{(n)})}{\varphi_n(x^{(n)})} < \frac{W_0}{W_1}\right\}. \tag{2.10}$$

It is easy to prove the following.

THEOREM 1. *One has*

$$\frac{\epsilon_n}{2} \leq H(\theta) - I_n \leq h(\epsilon_n) \tag{2.11}$$

where

$$h(p) = p \log \frac{1}{p} + (1 - p) \log \frac{1}{1 - p}, \qquad (0 \leq p \leq 1). \tag{2.12}$$

PROOF OF THEOREM 1. One has clearly

$$E(H(\theta|\zeta_n)) = W_0 E(H(\theta|\zeta_n)|\theta = \theta_0) + W_1 E(H(\theta|\zeta_n)|\theta = \theta_1); \tag{2.13}$$

$E(\eta|B)$ denotes the conditional expectation of η under condition B. Now evidently

(2.14)

$$W_1 E(H(\theta|\zeta_n)|\theta = \theta_1) \geq \int_{A_n} \frac{\varphi_n(x^{(n)})W_0}{\chi_n(x^{(n)})} \log\left(1 + \frac{W_1\psi_n(x^{(n)})}{W_0\varphi_n(x^{(n)})}\right) W_1\psi_n(x^{(n)})\, dx^{(n)}$$

and thus, as on the set defined by $(\psi_n(x^{(n)})/\varphi_n(x^{(n)})) \geq (W_0/W_1)$ one has $(W_1\psi_n(x^{(n)})/\varphi_n(x^{(n)})) \geq \frac{1}{2}$, it follows that

$$W_1 E(H\theta|\zeta_n)|\theta = \theta_1) \geq \tfrac{1}{2} \int_{A_n} W_0\varphi_n(x^{(n)})\, dx^{(n)} = \frac{W_0}{2} P(\Delta = \theta_1|\theta = \theta_0). \tag{2.15}$$

Similarly we obtain

$$W_0 E(H(\theta|\zeta_n)|\theta = \theta_0) \geq \frac{W_1}{2} P(\Delta = \theta_0|\theta = \theta_1). \tag{2.16}$$

Thus it follows that

$$E(H(\theta|\zeta_n)) \geq \frac{\epsilon_n}{2}. \tag{2.17}$$

This proves the lower inequality in (2.11).

This proof was given earlier in [2]; it is reproduced here for the convenience of the reader. To prove the upper inequality in (2.11) we apply the following integral form of Jensen's inequality:

$$\frac{\int_A h(a(x))k(x)\, dx}{\int_A k(x)\, dx} \leq h\left(\frac{\int_A a(x)k(x)\, dx}{\int_A k(x)\, dx}\right), \tag{2.18}$$

valid for every nonnegative function $k(x)$ and every concave function $h(x)$; we apply (2.18) to the concave function $h(x)$ defined by (2.12) and the domain A in X_n defined by $(\varphi_n(x^{(n)})W_0/\psi_n(x^{(n)})W_1) \geq 1$ and for the complementary domain $\bar{A}$. Thus we obtain

$$E(H(\theta|\zeta_n)) \leq \alpha h(p) + \beta h(q) \tag{2.19}$$

where

$$\alpha = \int_A \chi_n(x^{(n)})\, dx^{(n)}, \quad \beta = \int_{\bar{A}} \chi_n(x^{(n)})\, dx^{(n)} = 1 - \alpha,$$

(2.20)

$$p = \frac{W_1 \int_A \psi_n(x^{(n)})\, dx^{(n)}}{\alpha} \quad \text{and} \quad q = \frac{W_0 \int_{\bar{A}} \varphi_n(x^{(n)})\, dx^{(n)}}{\beta}.$$

Again applying Jensen's inequality in the form

$$\alpha h(p) + \beta h(q) \leq h(\alpha p + \beta q), \tag{2.21}$$

valid for any concave function h and for $\alpha \geq 0$, $\beta \geq 0$, $\alpha + \beta = 1$, since $\alpha p + \beta q = \epsilon_n$, it follows that

$$E(H(\theta|\zeta_n)) \leq h(\epsilon_n) = \epsilon_n \log \frac{1}{\epsilon_n} + (1 - \epsilon_n) \log \frac{1}{1 - \epsilon_n}. \tag{2.22}$$

Thus theorem 1 is proved.

We obtain easily from theorem 1 the following corollary.

COROLLARY. *The error of the standard decision tends to zero for $n \to \infty$ if and only if the amount of missing information tends to zero, that is,* $\lim_{n \to +\infty} \epsilon_n = 0$ *if and only if* $\lim_{n \to \infty} I_n = H(\theta)$.

It should be added that by a slight modification of the usual proof of the Neyman-Pearson fundamental lemma one can prove (see [3]) that the error of any decision function is at least as large as that of the standard decision. Thus if $\lim_{n \to \infty} I_n < H(\theta)$, that is, if the amount of information in the first n observations does not tend to the total amount of information needed, then there cannot exist a decision procedure whose error tends to zero, whereas if $\lim_{n \to \infty} I_n = H(\theta)$, there certainly exists such a procedure, namely the standard decision.

3. An upper bound for the amount of missing information and a criterion for obtaining full information

We prove now the following theorem.

THEOREM 2. *Let us write*

$$\lambda_k = \int_{-\infty}^{+\infty} \sqrt{f_k(x) g_k(x)}\, dx, \qquad (k = 1, 2, \cdots). \tag{3.1}$$

Then the following inequality holds:

$$0 \leq H(\theta) - I_n \leq B \sqrt{W_0 W_1} \prod_{k=1}^{n} \lambda_k \tag{3.2}$$

where $B > 0$ is an absolute constant.

PROOF OF THEOREM 2. The function $(h(x)/\sqrt{x})$ where $h(x)$ is defined by (2.12) is clearly continuous in the closed interval $0 \leq x \leq 1$. Let

$$C = \max_{0 \leq x \leq 1} \frac{h(x)}{\sqrt{x}}. \tag{3.3}$$

Since $h(x) = h(1 - x)$, we also have

$$C = \max_{0 \le x \le 1} \frac{h(x)}{\sqrt{1-x}}. \tag{3.4}$$

It follows that

$$H(\theta|\zeta_n) \le C\sqrt{P(\theta = \theta_1|\zeta_n)} \tag{3.5}$$

and also that

$$H(\theta|\zeta_n) \le C\sqrt{P(\theta = \theta_0|\zeta_n)}. \tag{3.6}$$

Thus we have, in view of (2.4) and (2.5),

$$H(\theta|\zeta_n) \le C\sqrt{\frac{W_0}{W_1}} \prod_{k=1}^{n} \sqrt{\frac{f_k(\xi_k)}{g_k(\xi_k)}} \tag{3.7}$$

and

$$H(\theta|\zeta_n) \le C\sqrt{\frac{W_1}{W_0}} \prod_{k=1}^{n} \sqrt{\frac{g_k(\xi_k)}{f_k(\xi_k)}}. \tag{3.8}$$

From (3.8) we obtain

$$W_0 E(H(\theta|\zeta_n)|\theta = \theta_0) \le C\sqrt{W_0 W_1} \prod_{k=1}^{n} \lambda_k, \tag{3.9}$$

and from (3.7) we obtain

$$W_1 E(H(\theta|\zeta_n)|\theta = \theta_1) \le C\sqrt{W_0 W_1} \prod_{k=1}^{n} \lambda_k. \tag{3.10}$$

Adding (3.9) and (3.10) we get

$$E(H(\theta|\zeta_n)) \le 2C\sqrt{W_0 W_1} \prod_{k=1}^{n} \lambda_k, \tag{3.11}$$

which proves theorem 2 with $B = 2C$.

Note that according to the Cauchy-Schwarz inequality, $0 \le \lambda_k \le 1$ and $\lambda_k = 1$ if and only if $f_k(x) = g_k(x)$ almost everywhere. Further, $\lambda_k = 0$ if and only if the intersection of the sets on which $f_k(x) > 0$ and $g_k(x) > 0$ is of Lebesgue measure zero. In this case, of course, the observation of ξ_k alone is sufficient with probability 1 to decide whether $\theta = \theta_0$ or $\theta = \theta_1$.

We shall now prove the following.

THEOREM 3. *One has*

$$\prod_{k=1}^{n} \lambda_k \le \sqrt{\frac{\epsilon_n}{W_0 W_1}} \tag{3.12}$$

where λ_k is defined by (3.1), *and ϵ_n is the error of the standard decision.*

PROOF OF THEOREM 3. Clearly,

$$\prod_{k=1}^{n} \lambda_k = \int_{X_n} \sqrt{\varphi_n(x^{(n)})\psi_n(x^{(n)})}\, dx^{(n)}. \tag{3.13}$$

Let us denote again by A the subset of X_n on which $(W_0\varphi_n(x^{(n)})/W_1\psi_n(x^{(n)}) \ge 1$ and put $\bar{A} = X_n - A$. Taking into account that φ_n is a density function, the Cauchy-Schwarz inequality gives

$$\int_A \sqrt{\varphi_n(x^{(n)})\psi_n(x^{(n)})}\, dx^{(n)} \le \left(\int_A \psi_n(x^{(n)})\, dx^{(n)}\right)^{1/2}. \tag{3.14}$$

Similarly we obtain

$$\int_{\bar{A}} \sqrt{\varphi_n(x^{(n)})\psi_n(x^{(n)})}\, dx^{(n)} \leq \left(\int_{\bar{A}} \varphi_n(x^{(n)})\, dx^{(n)}\right)^{1/2}. \tag{3.15}$$

Thus using again the Cauchy inequality we obtain

$$\prod_{k=1}^{n} \lambda_k \leq \left(\frac{1}{W_0} + \frac{1}{W_1}\right)^{1/2} \sqrt{\epsilon_n} = \sqrt{\frac{\epsilon_n}{W_0 W_1}}. \tag{3.16}$$

This proves theorem 3.

Now we can prove the following theorem.

THEOREM 4. *If $\lambda_k > 0$ for $k = 1, 2, \cdots$, the sequence of observations ξ_n $(n = 1, 2, \cdots)$ contains full information on θ if and only if the series*

$$\sum_{k=1}^{\infty} (1 - \lambda_k) \tag{3.17}$$

is divergent.

As regards the connection of theorem 4 with a theorem of Kakutani, see section 4.)

PROOF OF THEOREM 4. Since $1 - x \leq e^{-x}$, if the series $\sum_{k=1}^{\infty} (1 - \lambda_k)$ is divergent, one has $\lim_{n\to\infty} \prod_{k=1}^{n} \lambda_k = 0$, and thus by theorem 2 it follows that

$$\lim_{n\to+\infty} I_n = H(\theta). \tag{3.18}$$

This proves the "if" part of the theorem. On the other hand, using the inequality $1 - x \geq e^{-(x/1-x)}$, $(0 \leq x \leq 1)$, we obtain

$$\prod_{k=1}^{n} \lambda_k \geq \exp\left\{-\sum_{k=1}^{n} \left(\frac{1 - \lambda_k}{\lambda_k}\right)\right\}. \tag{3.19}$$

Now if $\sum_{k=1}^{\infty} (1 - \lambda_k)$ is convergent, then $\lim_{k\to\infty} \lambda_k = 1$, and since by assumption $\lambda_k > 0$ for $k = 1, 2, \cdots$, it follows that the sequence λ_k has a positive lower bound: $\lambda_k \geq c > 0$ for $k = 1, 2, \cdots$. It follows that the series $\sum_{k=1}^{\infty} (1 - \lambda_k/\lambda_k)$ is also convergent, and thus $\prod_{k=1}^{n} \lambda_k$ has a positive lower bound. By theorem 3 this implies that ϵ_n has a positive lower bound. Therefore, by theorem 1 the sequence $H(\theta) - I_n$ has a positive lower bound too. This proves the "only if" part of theorem 4.

THEOREM 5. *A sequence of statistics $\alpha_n = \alpha_n(\xi_1, \cdots, \xi_n)$ converging in probability to the true value of the parameter (real-valued) can exist only if the sequence ξ_n $(n = 1, 2, \cdots)$ contains full information with respect to θ, that is if* (3.18) *holds. Conversely if* (3.18) *holds, there exists a sequence of statistics α_n which converges with probability* 1 *to θ.*

PROOF OF THEOREM 5. If $\lim_{n\to\infty} P(|\alpha_n - \theta| > \epsilon) = 0$ for every $\epsilon > 0$, then we can construct the following decision function:

$$\begin{cases} \text{if } |\alpha_n - \theta_1| < |\alpha_n - \theta_0|, & \text{accept } \theta_1; \\ \text{if } |\alpha_n - \theta_0| < |\alpha_n - \theta_1|, & \text{accept } \theta_0; \\ \text{if } |\alpha_n - \theta_0| = |\alpha_n - \theta_1|, & \text{choose at random between } \theta_0 \text{ and } \theta_1 \text{ with probabilities } W_0 \text{ and } W_1. \end{cases} \tag{3.20}$$

Let us suppose that $\theta_0 < \theta_1$. Clearly,

$$P(|\alpha_n - \theta_0| \leq |\alpha_n - \theta_1| \,|\theta = \theta_1) = P\left(\alpha_n \leq \frac{\theta_0 + \theta_1}{2}|\theta = \theta_1\right), \tag{3.21}$$

and thus by assumption,

$$\lim_{n\to\infty} P(|\alpha_n - \theta_0| \leq |\alpha_n - \theta_1| \,|\theta = \theta_1) = 0. \tag{3.22}$$

Similarly,

$$\lim_{n\to\infty} P(|\alpha_n - \theta_1| \leq |\alpha_n - \theta_0| \,|\theta = \theta_0) = 0. \tag{3.23}$$

Thus the error of the decision in question tends to zero for $n \to \infty$. A fortiori, the error of the standard decision tends to zero which implies by theorem 1 that (3.18) holds. This proves the "only if" part of theorem 5. Conversely, if (3.18) holds and θ_0 and θ_1 are different real numbers, let us choose a sequence n_r such that the series $\sum_{r=1}^{\infty} \epsilon_{n_r}$ is convergent. Then by the Borel-Cantelli lemma the standard decisions Δ_{n_r} tend with probability 1 to θ. More exactly, $\Delta_{n_r} = \theta$ for all but a finite number of values of r. Now if we put $\alpha_n = \Delta_{n_r}$ for $n_r \leq n < n_{r+1}$, then the sequence of statistics $\alpha_n = \alpha_n(\xi_1, \cdots, \xi_n)$ tends with probability 1 to θ.

If the variables ξ_n have a discrete distribution, all our results remain valid; only the quantity λ_k has to be defined accordingly. If ξ_n can take the values $a_1, \cdots, a_l, \cdots$ and if

$$P(\xi_k = a_l|\theta = \theta_0) = p_{k,l}, \tag{3.24}$$

whereas

$$P(\xi_k = a_l|\theta = \theta_1) = q_{k,l}, \tag{3.25}$$

write

$$\lambda_k = \sum_{l=1}^{\infty} \sqrt{p_{k,l}q_{k,l}}. \tag{3.26}$$

All our results remain valid for this case.

4. Some examples

EXAMPLE 1. Let us suppose that $f_k(x) = f(x)$ and $g_k(x) = g(x)$, $(k = 1, 2, \cdots)$; that is, the random variables ξ_k are identically distributed under condition $\theta = \theta_i$, $(i = 1, 2)$. Suppose further that

$$1 > \lambda = \int_{-\infty}^{+\infty} \sqrt{f(x)g(x)}\, dx > 0. \tag{4.1}$$

Then it follows from theorem 2 that

$$H(\theta) - I_n \leq B\sqrt{W_0W_1}\lambda^n, \tag{4.2}$$

and thus the missing information tends exponentially to zero for $n \to \infty$. This question was treated in [2] and [3]. In [4] we have been concerned with the

special case where the random variables ξ_n take only the values 0 and 1, and where one has

$$P(\xi_n = 1|\theta = \theta_j) = p_j, \qquad P(\xi_n = 0|\theta = \theta_j) = q_j = 1 - p_j \tag{4.3}$$

for $j = 1, 2$. We have shown that the smallest value of λ for which (4.2) holds is $\lambda = 2^{-d(p_1,p_2)}$ where

$$d(p_1, p_2) = \rho \log \frac{\rho}{p_1} + (1 - \rho) \log \frac{1 - \rho}{1 - p_1} = \rho \log \frac{\rho}{p_2} + (1 - \rho) \log \frac{1 - \rho}{1 - p_2} \tag{4.4}$$

with

$$\rho = \frac{\log (q_1/q_2)}{\log (p_2q_1/p_1q_2)}. \tag{4.5}$$

In the special case where $p_1 = q_2 = p$, $q_1 = p_2 = q = 1 - p$, one has simply $\lambda = 2\sqrt{pq}$.

EXAMPLE 2. Let ξ_n be normally distributed with variance S_n^2 and with mean $m = m_0$ or $m = m_1 \neq m_0$ according to whether $\theta = \theta_0$ or $\theta = \theta_1$. We want to find the true value of m. Clearly

$$\lambda_k = \exp\left(-\frac{(m_0 - m_1)^2}{8S_k^2}\right). \tag{4.6}$$

It follows that we have full information on m if and only if the series $\sum_{k=1}^{\infty} (1/S_k^2)$ is divergent. Note that the statistic

$$\eta_n = \frac{\sum_{k=1}^{n} (\xi_k/S_k^2)}{\sum_{k=1}^{n} (1/S_k^2)}, \tag{4.7}$$

being the unbiased linear estimate of m with the least variance, is normally distributed with mean m and variance $(\sum_{k=1}^{n} (1/S_k^2))^{-1}$. If $\sum_{k=1}^{\infty} (1/S_k^2) = +\infty$, then η_n converges in probability to m and a suitably chosen subsequence η_{n_r}, such that $\sum_{r=1}^{\infty} (\sum_{k=1}^{n_r} (1/S_k^2))^{-1} < \infty$, converges with probability 1 to m.

EXAMPLE 3. Let an urn contain $a + b$ balls ($a > 0$, $b > 0$, $a \neq b$) of which either a are red and b white, or conversely, b are red and a white. Suppose that both cases have the prior probability $\frac{1}{2}$. We draw a ball from the urn, notice its color and put it back, and add, independently of the color of the ball drawn, $c_1 \geq 1$ red balls. After mixing the balls we draw again a ball, notice its color and put it back, adding $c_2 \geq 1$ red balls.

Let us continue this process indefinitely so that after the n-th step $c_n \geq 1$ new red balls are added. Can we determine with probability 1 the original composition of the urn?

Let us put $\xi_n = 1$ if the ball drawn at the n-th step is white and $\xi_n = 0$ if it is red. Let θ denote the number of white balls contained originally in the urn; thus $\theta_0 = b$ and $\theta_1 = a$. In this case

$$p_{k,1} = P(\xi_k = 1|\theta = \theta_0) = \frac{b}{a + b + \sum_{i=1}^{k-1} c_i} \tag{4.8}$$

and

$$q_{k,1} = P(\xi_k = 1|\theta = \theta_1) = \frac{a}{a + b + \sum_{i=1}^{k-1} c_i}. \tag{4.9}$$

Putting $N_k = a + b + \sum_{i=1}^{k-1} c_i$, one has

$$\lambda_k = \frac{\sqrt{ab}}{N_k} + \sqrt{\left(1 - \frac{a}{N_k}\right)\left(1 - \frac{b}{N_k}\right)}, \tag{4.10}$$

and thus

$$\lambda_k = 1 - \frac{\frac{a+b}{2} - \sqrt{ab}}{N_k} + 0\left(\frac{1}{N_k^2}\right). \tag{4.11}$$

Therefore, $\sum_{k=1}^{\infty}(1 - \lambda_k)$ is divergent or convergent according to whether $\sum_{k=1}^{\infty}(1/N_k)$is divergent or convergent. Thus $\sum_{k=1}^{\infty}(1/N_k) = +\infty$, for example if $c_n = 1$ for $n = 1, 2, \cdots$, one can find out the original composition of the urn with probability 1, whereas if $\sum_{k=1}^{\infty}(1/N_k) < \infty$ (for instance if $(c_n = n)$ this is impossible.

EXAMPLE 4. Suppose that ξ_n has under the condition that $\theta = \theta_i$ a Poisson distribution with mean value $\alpha_i\delta_n$, $(i = 1, 2)$ where $\alpha_1 > 0$, $\alpha_2 > 0$, $\alpha_1 \neq \alpha_2$ and $\delta_n > 0$, $\lim_{n\to\infty} \delta_n = 0$. In this case

$$\lambda_k = \exp\left(-\left(\frac{\alpha_1 + \alpha_2}{2} - \sqrt{\alpha_1\alpha_2}\right)\delta_k\right). \tag{4.12}$$

Clearly, $\sum_{k=1}^{\infty}(1 - \lambda_k)$ is divergent if and only if $\sum_{k=1}^{\infty}\delta_k$ is divergent. Thus, for instance, if $\delta_k = (\gamma/k)$, $(k = 1, 2, \cdots; \gamma > 0)$, it is possible to decide with probability 1 whether $\theta = \theta_1$ or $\theta = \theta_2$, but if $\delta_k = (\gamma/k^2)$, this is not possible.

EXAMPLE 5. Let ξ_n have a binomial distribution of fixed order N and parameter $\delta_n\theta_i$ under the condition that $\theta = \theta_i$, $(i = 0, 1)$ where $\theta_0 > 0$, $\theta_1 > 0$, $\theta_0 \neq \theta_1$. Then we can easily see that the observations ξ_n, $(n = 1, 2, \cdots)$ contain full information on θ if and only if $\sum_{k=1}^{\infty}\delta_n = +\infty$.

EXAMPLE 6. Let η_n have the density function $f(x)$ where $f(x)$ is everywhere positive and has a continuous derivative such that $\int_{-\infty}^{+\infty}(f'^2(v)/f(v))\,dv < +\infty$. Suppose that $\xi_n = m_0 + c_n\eta_n$ if $\theta = \theta_0$ and $\xi_n = m_1 + c_n\eta_n$ if $\theta = \theta_1$ where $m_0 \neq m_1$ and $c_n \to \infty$. Writing $d = m_1 - m_0$, we have

$$1 - \lambda_k = \tfrac{1}{2}\int_{-\infty}^{+\infty}\left(\sqrt{f\left(u + \frac{d}{c_n}\right)} - \sqrt{f(u)}\right)^2 du. \tag{4.13}$$

Write $p(u) = \sqrt{f(u)}$; then we have

$$1 - \lambda_k = \tfrac{1}{2}\int_{-\infty}^{+\infty}\left(\int_u^{u+(d/c_n)} p'(v)\,dv\right)^2 du. \tag{4.14}$$

Thus

$$1 - \lambda_k \leq \frac{d^2}{8c_k^2}\int_{-\infty}^{+\infty}\frac{f'^2(v)}{f(v)}\,dv. \tag{4.15}$$

It follows that if the series $\sum_{k=1}^{\infty} (1/c_k^2) < +\infty$, then $\sum_{k=1}^{\infty} 1 - \lambda_k < +\infty$, and we do not get full information on θ.

On the other hand, the divergence of the series $\sum (1/c_k^2)$ ensures that the sequence of observations $\{\xi_n\}$ does contain full information on θ. As a matter of fact, since $p'(x)$ is continuous, there exists an interval (a, b) in which $p'(x)$ does not change sign and $|p'(x)| \geq \delta > 0$. Letting $h_n = d/c_n$, it follows that

$$1 - \lambda_n \geq \delta^2(b - a - h_n)h_n^2. \tag{4.16}$$

Thus if $\sum_{n=1}^{\infty} (1/c_n^2) = +\infty$, then $\sum_{n=1}^{\infty} (1 - \lambda_n) = +\infty$.

5. Further remarks

S. Kakutani [5] has proved the following.

THEOREM K. *Let $(\Omega_n, \mathfrak{A}_n)$ be measurable spaces $(n = 1, 2, \cdots)$. Let Ω be the product space $\prod_{n=1}^{\infty} \Omega_n$. Let μ_n and ν_n be two equivalent probability measures on Ω_n, and let μ and ν denote the product measures $\mu = \prod_{n=1}^{\infty} \mu_n$ and $\nu = \prod_{n=1}^{\infty} \nu_n$ on Ω. Then the measures μ and ν are either equivalent or orthogonal according to whether the infinite product $\prod_{n=1}^{\infty} \rho(\mu_n, \nu_n)$ is convergent or divergent. Here $\rho(\mu_n, \nu_n)$ denotes the Hellinger distance of the measures μ_n and ν_n; that is, if m is any measure such that μ_n and ν_n are both absolutely continuous with respect to m,*

$$\rho(\mu_n, \nu_n) = \int_{\Omega_n} \sqrt{\frac{d\mu_n}{dm} \cdot \frac{d\nu_n}{dm}}\, dm \tag{5.1}$$

where $(d\mu_n/dm)$ and $(d\nu_n/dm)$ are Radon-Nikodym derivatives.

(The value of ρ is clearly independent of the choice of m.)

Obviously our results are closely connected with the above mentioned theorem of S. Kakutani. (My thanks are due to Professor K. Jacobs for calling my attention to this fact.)

As a matter of fact, according to Kakutani's theorem, if μ_n and ν_n denote the measures with density $f_n(x)$ and $g_n(x)$ with respect to the Lebesgue measure on the real line, then according to whether $\theta = \theta_0$ or $\theta = \theta_1$, one obtains in the sequence-space $x = (\xi_1, \xi_2, \cdots, \xi_n, \cdots)$ the product measure μ or ν. Further, our λ_k is equal to the Hellinger distance ρ_k of μ_k and ν_k. It follows that $\mu \sim \nu$ or $\mu \perp \nu$ according to whether $\prod_{k=1}^{\infty} \lambda_k > 0$ or $\prod_{k=1}^{\infty} \lambda_k = 0$.

If $\prod_{k=1}^{\infty} \lambda_k = 0$, that is, if $\mu \perp \nu$, then there exists a measurable set A in X such that $\mu(A) = 1$ and $\nu(A) = 0$. In this case given the infinite sequence $\{\xi_n\}$ one can clearly decide with probability 1 whether $\theta = \theta_0$ or $\theta = \theta_1$: if the infinite sequence $(\xi_1, \cdots, \xi_n, \cdots) \in A$ we decide for $\theta = \theta_0$, and in the opposite case for $\theta = \theta_1$. Now one can obviously find a sequence of sets A_n such that A_n is a cylinder set of X, its base belonging to the finite product set $\prod_{k=1}^{n} \Omega_k$, such that $\mu(A_n) \to 1$ and $\nu(A_n) \to 0$ for $n \to +\infty$. Thus if after observing $\xi_1, \cdots, \xi_n$ we decide for θ_0 or θ_1 according to whether or not the point $(\xi_1, \cdots, \xi_n)$ belongs to the base of A_n, the error of our decision tends to 0 for $n \to \infty$.

Conversely, it is easy to see that such a sequence of sets cannot exist if $\mu \sim \nu$.

Thus, in view of the corollary of theorem 1, theorem 4 can be deduced from theorem K. Note, however, that our direct approach is not only more elementary than this one via Kakutani's theorem, it also gives somewhat more, as it shows not only that $\epsilon_n \to 0$ if and only if $\lim_{n\to\infty} \prod_{k=1}^{n} \lambda_k = 0$, but also gives estimates between these quantities for finite values of n.

REFERENCES

[1] D. V. LINDLEY, "On a measure of the information provided by an experiment," *Ann. Math. Statist.*, Vol. 27 (1956), pp. 986–1005.

[2] A. RÉNYI, "On the amount of information concerning an unknown parameter in a sequence of observations," *Publ. Math. Inst. Hungar. Acad. Sci.*, Vol. 9 (1964), pp. 617–624.

[3] ———, "On the amount of missing information and the Neyman-Pearson lemma," *Festschrift for J. Neyman*, Wiley, 1966, pp. 281–288.

[4] ———, "On the amount of information in a frequency-count," *The 35th Session of the International Statistical Institute at Beograd*, 1965, pp. 1–8.

[5] S. KAKUTANI, "On equivalence of infinite product measures," *Ann. of Math.*, Vol. 49 (1948), pp. 214–226.