Studia Scientiarum Mathematicarum Hungarica 2 (1967) 249—256.

STATISTICS AND INFORMATION THEORY[1]

by

A. RÉNYI

§ 0. Introduction

In the present paper we deal with certain basic questions connected with the information-theoretic point of view on statistics. This paper is a continuation of the papers [1], [2], [3], [4], [5] of the author; most of the results of these previous papers are presented here in an improved (sharper or more general) form.

§ 1. On the Amount of Information in a Random Variable Concerning Another

In this section we collect the basic definitions and well known results needed in what follows.

Let $S=(\Omega, \mathscr{A}, \mathbf{P})$ be a probability space, i. e. Ω an arbitrary nonempty set, $\mathscr{A}$ a σ-algebra of subsets of Ω and $\mathbf{P}$ a probability measure on $\mathscr{A}$. In what follows θ will always denote a discrete valued random variable in S, i. e., a function $\theta=\theta(\omega)$ defined for $\omega\in\Omega$, taking on only a finite number of different values $\theta_1, \theta_2, \ldots, \theta_r$ ($r\geqq 2$) for which the set (event) $H_k=\{\omega: \theta(\omega)=\theta_k\}$ belongs to $\mathscr{A}$ for $k=1, 2, \ldots, r$. Here $\theta_1, \theta_2, \ldots, \theta_r$ may be numbers, or any distinguishable symbols: their values will be in what follows irrelevant. We shall usually interpret θ as the parameter of a probability distribution and the event H_k as the *hypothesis that the true value of the parameter θ is equal to θ_k*; we shall use the notation

$$p_k=\mathbf{P}(H_k)=\mathbf{P}(\theta=\theta_k) \qquad (k=1, 2, \ldots, r) \tag{1. 1}$$

and call the distribution $(p_1, p_2, \ldots, p_r)$ of θ (contrasting it with the conditional (or posterior) distribution of θ given certain observations, to be introduced later) the *prior distribution* of θ. The (unconditional) entropy of θ is defined by Shannon's formula [2]

$$\mathbf{H}(\theta)=\sum_{k=1}^{r} p_k \log_2 \frac{1}{p_k} \tag{1. 2}$$

where the numbers p_k ($k=1, 2, \ldots, r$) are those defined by (1. 1) $\mathbf{H}(\theta)$ will be interpreted as *the amount of missing information on θ* when nothing else is known about θ except that its prior distribution is given.

[1] This paper has been presented to the 1st European Meeting of Statisticians held in London, 5—10 September 1966.

[2] $\log_2 x$ denotes the logarithm with base 2 of the positive number x; $0 \log_2 \frac{1}{0}$ always means 0.

Let now $\xi=\xi\{\omega)=(\xi_1(\omega\}, \ldots, \xi_n(\omega))$ be an n-dimensional vector valued random variable, i. e., an $\mathscr{A}$-measurable function defined on Ω and with values in the Euclidean space E_n of dimension n. We shall interpret ξ as an *observed sample*. As ξ and θ are random variables on the same probability space, by observing ξ we usually get some information on θ (except when θ and ξ are independent). After having observed ξ we may consider the *conditional* (or posterior) *distribution*

$$p_k(\xi)=\mathbf{P}(H_k|\xi) \tag{1. 3}$$

of H_k given the value of ξ. The conditional probability of an event $A\in\mathscr{A}$ given the observed value of ξ is as usual defined as follows: Let $\mathscr{A}_\xi$ denote the least σ-algebra of subsets of Ω on which ξ is measurable (i. e., the σ-algebra generated by ξ). By supposition $\mathscr{A}_\xi$ is a subalgebra of $\mathscr{A}$. The conditional probability $\mathbf{P}(A|\xi)$ of an event A, given the value of ξ, is defined as an $\mathscr{A}_\xi$-measurable function (random variable) such that for every $B\in\mathscr{A}_\xi$, one has

$$\int_B \mathbf{P}(A|\xi)\,dP = \mathbf{P}(AB). \tag{1. 4}$$

As well known, $\mathbf{P}(A|\xi)$ is by (1. 4) uniquely defined up to a set of measure 0 and $\{\mathbf{P}(H_1|\xi), \ldots, \mathbf{P}(H_r|\xi)\}$ is with probability one a probability distribution, i. e., $\mathbf{P}\left(\sum_{k=1}^{r}\mathbf{P}(H_k|\xi)=1\right)=1$. Let us consider now the entropy of the conditional (a posteriori) distribution of θ given ξ, i. e., the quantity

$$\mathbf{H}(\theta|\xi)=\sum_{k=1}^{r} p_k(\xi)\log_2\frac{1}{p_k(\xi)}. \tag{1. 5}$$

We interpret $\mathbf{H}(\theta|\xi)$ as *the amount of information concerning θ still missing after having observed the sample ξ*. Clearly $\mathbf{H}(\theta|\xi)$ itself is a random variable (which is not only $\mathscr{A}$-measurable but also $\mathscr{A}_\xi$-measurable); its expectation $\mathbf{E}(\mathbf{H}(\theta|\xi))$ is interpreted as *the average amount of information still missing about θ after having observed ξ*. We shall call this quantity for the sake of brevity when there is no danger of misunderstanding simply „the amount of missing information", and denote it by $R(\xi, \theta)$; i. e., we put[3]

$$R(\theta, \xi)=\mathbf{E}(\mathbf{H}(\theta|\xi)). \tag{1. 6}$$

The *amount of information* $I(\theta, \xi)$ in the observed sample ξ with respect to the (unknown) parameter θ is defined as the average decrease of the entropy of θ by observing ξ; that is, we put

$$I(\theta, \xi)=\mathbf{H}(\theta)-R(\theta, \xi). \tag{1. 7}$$

Evidently the conditional (posterior) distribution $\{p_1(\xi), \ldots, p_r(\xi)\}$ of θ is identical with its prior distribution $\{p_1, \ldots, p_r\}$ if and only if ξ and θ are independent. In this case $R(\theta, \xi)=\mathbf{H}(\theta)$, i. e., $I(\theta, \xi)=0$, that is the observation of the sample ξ does not give us any information on θ. In every other case one has $R(\theta, \xi)<\mathbf{H}(\theta)$

[3] Here and in what follows $\mathbf{E}(\eta)$ denotes the expectation of the random variable η.

and thus $I(\theta, \xi) > 0$. This can be shown by Jensen's inequality as follows. As the function $x \log_2 \frac{1}{x}$ is concave in (0, 1) and by Jensen's inequality for any concave function $f(x)$ and any random variable η the values of which are lying in the domain of definition of $f(x)$ one has

$$\mathbf{E}(f(\eta)) \leqq f(\mathbf{E}(\eta)) \tag{1.8}$$

it follows

$$R(\theta, \xi) = \sum_{k=1}^{r} \int_{\Omega} p_k(\xi) \log_2 \frac{1}{p_k(\xi)} dP \leqq \tag{1.9}$$

$$\leqq \sum_{k=1}^{r} \left(\int_{\Omega} p_k(\xi) dp \right) \frac{1}{\left(\int_{\Omega} p_k(\xi) d\xi \right)} = \mathbf{H}(\theta)$$

because by (1. 4)

$$\int_{\Omega} p_k(\xi) dP = \mathbf{P}(H_k) = p_k. \tag{1.10}$$

Evidently there is equality in (1. 9) if and only if the distribution $\{p_1(\xi), \ldots, p_r(\xi)\}$ is (with probability 1) identical to the distribution $\{p_1, \ldots, p_r\}$, i. e., if ξ and θ are independent.

Let $g(x)$ $(x \in E_n)$ be any k-dimensional vector valued Borel measurable function defined on the n-dimensional space E_n. We shall call the random variable $g(\xi)$ a *statistic*. If after observing ξ we consider the value of the statistic $g(\xi)$ only, and disregard every information (on θ) contained in the observation of ξ and not contained in $g(\xi)$, we usually loose some amount of information, i. e.,

$$I(g(\xi), \theta) \leqq I(\xi, \theta). \tag{1.11}$$

The inequality (1. 11) is clearly equivalent to

$$R(\xi, \theta) \leqq R(g(\xi), \theta). \tag{1.12}$$

To prove (1. 12) we need the following Lemma 1 which is an immediate consequence of the definition of conditional probability.

LEMMA 1. *If $f(x)$ is any Borel measurable function and $A \in \mathcal{A}$ any event, we have*

$$\mathbf{E}(f(g(\xi))\mathbf{P}(A|g(\xi))) = \mathbf{E}(f(g(\xi))\mathbf{P}(A|\xi)). \tag{1.13}$$

Using Lemma 1, we obtain

$$R(\theta, g(\xi)) - R(\theta, \xi) = \mathbf{E}\left(\sum_{k=1}^{r} \mathbf{P}(H_k|\xi) \log_2 \frac{\mathbf{P}(H_k|\xi)}{\mathbf{P}(H_k|g(\xi))} \right). \tag{1.14}$$

Now we need the following simple

LEMMA 2. *If $\{q_1, q_2, \ldots, q_r\}$ and $\{Q_1, Q_2, \ldots, Q_r\}$ are arbitrary probability distributions consisting of the same number r of terms, we have*

$$\sum_{k=1}^{r} q_k \log_2 \frac{q_k}{Q_k} \geqq 0 \tag{1.15}$$

with equality standing in (1. 15) *if and only if $q_k = Q_k$ for $k = 1, 2, \ldots, r$.*

Applying Lemma 2, we obtain from (1. 14) that (1. 12) holds and there is equality in (1. 12) if and only if with probability 1, one has

$$\mathbf{P}(H_k|\xi)=\mathbf{P}(H_k|g(\xi)) \qquad (k=1, 2, \ldots, r). \tag{1. 16}$$

If (1. 16) holds (with probability 1) we call $g(\xi)$ a *sufficient function of* ξ *for* θ (or a *sufficient statistic*). Thus a function of the observations is called sufficient for a parameter if and only if it contains all information in the observation which is relevant to the parameter, in the sense that there is equality in (1. 11).

Note that if (1. 16) holds and the random vector ξ has the conditional density $\varphi_k(x)$ under condition H_k, and $g(\varphi)$ has the density $\psi_k(g(x))$, then

$$\varphi_k(x)=\psi_k(g(x))\chi(x)$$

where the function $\chi(x)$ does not depend on k; as clearly $\varphi_k(x)$, $\psi_k(g(x))$ and $\chi(x)$ are all independent from the prior distribution $\{p_1, \ldots, p_r\}$ of θ, it follows that our definition of sufficiency is equivalent with the usual definition of a sufficient statistic in case both definitions are applicable. An advantage of our definition is that it does not depend on the existence of densities; besides it has a clear information-theoretical meaning.

Before proceeding further we prove the following

THEOREM 1. *The conditional distribution* $\Pi(\xi)=(p_1(\xi), \ldots, p_r(\xi))$ *of* θ *given* ξ, *considered as a statistic, is sufficient with respect to* θ.

To prove our theorem it is clearly enough to show that

$$p(H_k|\Pi(\xi))=p_k(\xi) \qquad (k=1, 2, \ldots, n). \tag{1. 17}$$

But (1. 17) is evidently true as $p_k(\xi)$ is $\mathscr{A}_{\Pi(\xi)}$-measurable ($p_k(\xi)$ being the k-th component of the vector $\Pi(\xi)$, we get $p_k(\xi)$ by projecting the vector $\Pi(\xi)$ to the x_k-axis.)

The statement of Theorem 1 can be expressed by saying that the *conditional distribution of* θ *given* ξ *contains all information relevant on* θ *which is present in the sample* ξ.

§ 2. A Bayesian Version of the Fundamental Lemma of Neyman and Pearson

If we have to make a *decision* concerning the parameter θ, on the basis of the observed value of the sample ξ, i. e., after observing ξ we have to select one of the possible values of θ, this decision can be described by a Borel measurable function $D(\xi)$ of ξ, the set of values of which is the set $\{\theta_1, \theta_2, \ldots, \theta_r\}$ of possible values of θ. The *error* e of such a decision is simply the probability of the decision being false, that is

$$e=\mathbf{P}(D(\xi)\neq\theta). \tag{2. 1}$$

We define the *standard decision* $\Delta(\xi)$ as follows: we decide always in favor of that hypothesis H_k (that value θ_k of θ) which has the largest conditional probability given the value of ξ; in case there is more than one value k such that $p_k(\xi)=\max_{1\leq j\leq r} p_j(\xi)$, we select in some way one among those values—say the least such value of k. If another rule is applied we call the corresponding decision a *variant of the standard decision.*

It is easy to see that it does not matter much which one of these values of k we choose (i. e., whether we use the standard decision or one of its variants) as the error of the decision is independent from this selection. As a matter of fact if ε denotes the *error of the standard decision*, we obtain by the definition (1. 4) of conditional probabilities

$$\varepsilon = \mathbf{P}(\Delta(\xi) \neq \theta) = 1 - \mathbf{P}(\Delta(\xi) = \theta) = 1 - \mathbf{E}(\mathbf{P}(\theta = \Delta(\xi)|\xi)). \tag{2. 2}$$

Clearly if we change the definition of the standard decision for some value of ξ from $\Delta(\xi) = \theta_{k_1}$ to $\Delta(\xi) = \theta_{k_2}$ where $p_{k_1}(\xi) = p_{k_2}(\xi)$, then ε remains unchanged, because $\mathbf{P}(\theta = \Delta(\xi)|\xi) = p_{\Delta(\xi)}(\xi)$ is by definition not affected by such a change.

Now let $D(\xi)$ be any other decision, and e its error. Then we get, similarly to (2. 2)

$$e = 1 - \mathbf{E}(\mathbf{P}(\theta = D(\xi)|\xi)). \tag{2. 3}$$

Thus we have

$$e - \varepsilon = \mathbf{E}(\mathbf{P}(\theta = \Delta(\xi)|\xi) - \mathbf{P}(\theta = D(\xi)|\xi)). \tag{2. 4}$$

The random variable, the expectation of which gives the difference $e - \varepsilon$, is clearly always non-negative, because for each value of ξ we have for some value of k (namely $k = D(\xi)$)

$$\mathbf{P}(\theta = \Delta(\xi)|\xi) - \mathbf{P}(\theta = D(\xi)|\xi) = \max_{1 \leqq j \leqq r} p_j(\xi) - p_k(\xi) \geqq 0. \tag{2. 5}$$

Thus we have proved the following

THEOREM 2. *No decision can have a smaller error than the standard decision.*

Clearly if the decision $D(\xi)$ is such that $\mathbf{P}(\theta = D(\xi)|\xi) \neq \mathbf{P}(\theta = \Delta(\xi)|\xi)$ with positive probability, then $e > \varepsilon$. However, if $\mathbf{P}(\theta = D(\xi)|\xi) = \mathbf{P}(\theta = \Delta(\xi)|\xi)$ with probability 1, this means that the decision $D(\xi)$ differs from the decision $\Delta(\xi)$ only in that in case a tie presents itself, i. e., if the value of k for which $p_k(\xi)$ is maximal is not unique, the decision $D(\xi)$ prescribes another choice among those values k for which $p_k(\xi)$ is maximal as $\Delta(\xi)$; thus *except for variants of the standard decision every other decision has a definitely larger error than the standard decision* (or any of its variants).

Note that the difference between Theorem 2 and the usual form of the Neyman—Pearson lemma consists in that we have supposed that the parameter θ is a random variable, i. e. we have taken the Bayesian point of view. Thus we do not distinguish between errors of the first and second kind: only one sort of error is possible. A decision is namely either correct, or wrong, and the error of a decision is the probability of it being wrong. A formal difference of minor importance is that by using the general notion of a conditional probability we did not need any supposition concerning the existence of densities.

Note that it follows from Theorem 2 that the error of the standard decision is $\leqq 1/2$ in the case $r = 2$, because if $\bar{\Delta}$ means the decision which is the opposite of Δ, $\bar{\Delta}$ has the error $1 - \varepsilon$ and thus by Theorem 2, $\varepsilon \leqq 1 - \varepsilon$.

As regards the standard decision $\Delta(\xi)$, we may compute the amount of information contained in the value of $\Delta(\xi)$ with respect to θ, i. e., the quantity $I(\Delta(\xi), \theta)$. Clearly one has $I(\Delta(\xi), \theta) \leqq I(\xi, \theta)$ with strict inequality except when $\Delta(\xi)$ is a sufficient function of ξ concerning θ; thus *even when the best possible decision is adopted some information is lost.* The explanation of this somewhat paradoxically

sounding statement is that usually the information on θ contained in the observed value of ξ is not enough to decide with certainty which is the value of the parameter, it only gives us a (conditional) probability distribution on the possible values. If, nevertheless, we insist on choosing one of the possible values and rejecting all the others, we naturally lose by this a certain amount of information.

§ 3. Estimating the Error of the Standard Decision by the Amount of Missing Information

We prove in this section the following [4]

THEOREM 3. *Let ε denote the error of the standard decision and $R=R(\theta, \xi)$ the amount of missing information, then the following inequality holds*

$$\log_2 \frac{1}{1-\varepsilon} \leqq R, \tag{3.1}$$

or expressed otherwise

$$\varepsilon \leqq 1 - \frac{1}{2^R}. \tag{3.2}$$

PROOF OF THEOREM 3. Let us denote for the sake of brevity the event $\Delta(\xi)=\theta_j$ by A_j $(j=1, 2, \ldots)$. Then we have clearly

$$R = \mathbf{E}(\mathbf{H}(\theta|\xi)) = \sum_{j=1}^{r} \mathbf{P}(A_j)\mathbf{E}(\mathbf{H}(\theta|\xi)|A_j). \tag{3.3}$$

(Here and in what follows $\mathbf{E}(\eta|B)$ denotes the conditional expectation of the random variable η with respect to the condition B, when B is an event such that $\mathbf{P}(B)>0$.) Now by definition under condition A_j we have $p_k(\xi)\leqq p_j(\xi)$ for $k=1, 2, \ldots, r$; in view of (1. 5) we get that

$$R \geqq \sum_{j=1}^{r} \mathbf{P}(A_j)\mathbf{E}\left(\log_2 \frac{1}{p_j(\xi)} \,\middle|\, A_j\right). \tag{3.4}$$

Applying now Jensen's inequality to the convex function $\log_2 \frac{1}{x}$ $(0 \leqq x \leqq 1)$, it follows that

$$R \geqq \sum_{j=1}^{r} \mathbf{P}(A_j) \log_2 \frac{1}{\mathbf{E}(p_j(\xi)|A_j)}. \tag{3.5}$$

Now it follows from (1. 4) that

$$\mathbf{E}(p_j(\xi)|A_j) = \frac{\int_{A_j} \mathbf{P}(\theta = \theta_j|\xi)\, dP}{\mathbf{P}(A_j)} = \mathbf{P}(\Delta(\xi) = \theta|A_j). \tag{3.6}$$

[4] In our previous paper [3] we have proved only the weaker estimate $\varepsilon \leqq R$. Clearly (3. 1) implies not only $\varepsilon \leqq R$ but also $\frac{\varepsilon}{\ln 2} \leqq R$.

Thus we obtain from (3. 5)

$$R \geqq \sum_{j=1}^{r} \mathbf{P}(A_j) \log_2 \frac{1}{\mathbf{P}(\Delta(\xi) = \theta | A_j)}. \tag{3. 7}$$

We need now Jensen's inequality, in the form that if $f(x)$ is a convex function, $x_1, \ldots, x_r$ any values in the domain of definition of $f(x)$ and $w_1, \ldots, w_r$ non-negative numbers with sum equal to one, then

$$\sum_{j=1}^{r} w_j f(x_j) \geqq f\left(\sum_{j=1}^{r} w_i x_j \right). \tag{3. 8}$$

Applying (3. 8) it follows from (3. 7) that

$$R \geqq \log_2 \frac{1}{\sum_{j=1}^{r} \mathbf{P}(A_j) \mathbf{P}(\Delta(\xi) = \theta | A_j)} = \log_2 \frac{1}{\mathbf{P}(\Delta(\xi) = \theta)} = \log_2 \frac{1}{1-\varepsilon} \tag{3. 9}$$

and this proves (3. 1).

In our previous paper [4] we have shown for the special case $r=2$ that the inequality $2\varepsilon \leqq R$ holds; for this special case this is slightly better than (3. 1). We reproduce here the proof of this inequality as it requires only a few lines. Let the possible values of θ be θ_0 and θ_1, the corresponding hypotheses $\theta = \theta_0$ and $\theta = \theta_1$ shall be denoted by H_0 and H_1 respectively. Put

$$h(x) = x \log_2 \frac{1}{x} + (1-x) \log_2 \frac{1}{1-x}. \tag{3. 10}$$

Then we have evidently $h(x) = h(1-x)$ and $h(x) \geqq 2x$ for $0 \leqq x \leqq 1/2$. Let us put

$$p^*(\xi) = \begin{cases} p_0(\xi) & \text{if } p_0(\xi) \leqq \frac{1}{2} \text{ i. e. if } \Delta(\xi) = \theta_1 \\ p_1(\xi) & \text{if } p_0(\xi) \geqq \frac{1}{2} \text{ i. e. if } \Delta(\xi) = \theta_0. \end{cases} \tag{3. 11}$$

Then we have clearly $p^*(\xi) \leqq \frac{1}{2}$ further

$$R = \mathbf{E}(h(p^*(\xi))) \geqq 2\mathbf{E}(p^*(\xi)). \tag{3. 12}$$

Denoting the event $\Delta(\xi) = \theta_0$ by B_0 and the event $\Delta(\xi) = \theta_1$ by B_1 we obtain

$$R \geqq 2\left(\int_{B_1} p_0(\xi) dP + \int_{B_0} p_1(\xi) dP \right). \tag{3. 13}$$

As by (1. 4) we have

$$\int_{B_1} p_0(\xi) dP = \mathbf{P}(H_0 B_1) \quad \text{and} \quad \int_{B_0} p_1(\xi) dP = \mathbf{P}(H_1 B_0) \tag{3. 14}$$

it follows that

$$R \geqq 2(\mathbf{P}(H_0 B_1) + \mathbf{P}(H_1 B_0)) = 2\varepsilon, \tag{3. 15}$$

which was to be proved.

Returning to the general case, we mention that one can also get an upper bound for the amount of missing information by means of the error of the standard decision. In this direction the following theorem is known (see [6] p. 35.).

THEOREM 4. *One has*

$$R \leqq h(\varepsilon) + \varepsilon \log_2(r-1) \tag{3. 16}$$

where $h(x)$ *is defined by* (3. 10).

My thanks are due to G. KATONA, who called my attention to the fact that the estimation (3.16), proved first by R. M. FANO [7], is slightly sharper than a similar estimate which I have found previously.

§ 4. Conclusion

It follows from Theorems 3 and 4 that if we have an infinite sequence of observations $\xi_1, \xi_2, \ldots, \xi_n, \ldots$ each ξ_n being a random variable on the probability space S (it is not a restriction to suppose that each ξ_n is real valued), and $\xi^{(n)}$ denotes the sample $(\xi_1, \xi_2, \ldots, \xi_n)$ further Δ_n the standard decision concerning the true value of θ taken on the basis of observing the sample $\xi^{(n)}$ and ε_n the error of the decision Δ_n, and if finally R_n denotes the average amount of information on θ still missing after having observed the sample $\xi^{(n)}$, then $\lim_{n\to\infty} \varepsilon_n = 0$ if and only if $\lim_{n\to\infty} R_n = 0$. This shows that to get in the limit all information on θ which is needed, is equivalent with having the possibility to make decisions on the true value of θ the probability of correctness of which is in the limit equal to 1. By other words the information-theoretical point of view is in accordance with the usual point of view of statistics.

REFERENCES

[1] RÉNYI, A.: On the amount of information concerning an unknown parameter in a sequence of observations, *Magyar Tud. Akad. Mat. Kutató Int. Közl.* **9** (1964) 617—625.

[2] RÉNYI, A.: On the amount of information in a frequency count, *35th Session of the International Statistical Institute,* Beograd, 1965, 1—8.

[3] RÉNYI, A.: On the amount of missing information and the Neyman-Pearson lemma, *Festschrift for J. Neyman,* Wiley, London, 1966, 281—288.

[4] RÉNYI, A.: On the amount of information in a random variable concerning an event, *Journal of Mathematical Sciences (Delhi)* **1** (1966) 30—33.

[5] RÉNYI, A.: On some basic problems of statistics from the point of view of information theory, *Proceedings of the 5th Berkeley Symposium* (in print).

[6] FEINSTEIN, A.: *Foundations of Information Theory*, McGraw-Hill, New York, 1958.

[7] FANO, R. M.: *Statistical Theory of Communication*, MIT, Cambridge, Mass., 1954.

MATHEMATICAL INSTITUTE OF THE HUNGARIAN ACADEMY OF SCIENCES, BUDAPEST

(Received October 8, 1966.)