REVIEW OF THE INTERNATIONAL
STATISTICAL INSTITUTE
Volume 33 : 1, 1965

ON THE FOUNDATIONS OF INFORMATION THEORY[1]

by

A. Rényi
Mathematical Institute of the Hungarian Academy of Sciences, Budapest

1. INTRODUCTION

In the present paper we shall give a review of certain investigations on the foundations of information theory. We shall deal exclusively with one question, which is however fundamental in information theory: *how should the amount of information be measured?* By other words we shall give a review of results concerning the *definition of information.*

This question can be approached from two essential different points of view. The first is the *axiomatic* (or postulational) *approach*; starting from the intuitive notion of information one formulates certain properties which a reasonable measure of information has to satisfy; after this the purely mathematical question has to be solved to find all the expressions which possess the postulated properties. The first to adopt this point of view was *C. Shannon* who in his fundamental paper [1] deduced from certain postulates his formula

$$H(P) = \sum_{k=1}^{a} p_k \log_2 \frac{1}{p_k} \tag{1.1}$$

for the entropy[2] of a probability distribution $P = (p_1, p_2, \ldots, p_a)$ i.e. the amount of information contained in a single observation of a random variable ξ which takes on a different values $x_1, x_2, \ldots, x_a$ with the probabilities $p_k = P(\xi = x_k)$ $(k = 1, 2, \ldots, a)$.

The second point of view may be characterized as the *pragmatic approach*; this approach starts from certain particular problems of information theory and accepts as measures of the amount of information the quantities which present themselves in the solution. According to this point of view the real justification of some measure of information is that it does work. This second point of view has been recently emphasized by *J. Wolfowitz* in his book [2], who however points out that the view that the main reason for introducing the quantity (1) is its role in coding theory, was expressed already by *Shannon.*

These two points of view are according to the opinion of the author of this paper not as opposed to each other as they seem to be; they are compatible and even complement each other and therefore both deserve attention. Both of the mentioned approaches may and should be used as a control to the other. As a matter of fact, if certain quantities are deduced from some natural postulates (from "first principles") these certainly need for their final justification the control whether they can be effectively used in solving concrete problems. On the other hand if one encounters a

---

[1] A review paper prepared for the 34th Session of the International Statistical Institute, Ottawa, Canada, August 1963.

[2] The quantity $H(P)$ defined by (1) is interpreted either as a measure of entropy (i.e. of uncertainty) or as a measure of information. Both interpretations are justified. As a matter of fact the difference between these two interpretations consists only in that whether we imagine ourselves in a moment *before* carrying out an experiment whose a possible results have the probabilities $p_1, p_2, \ldots, p_a$ (in which case $H(P)$ measures our uncertainty concerning the result of the experiment) or we imagine ourselves in a moment *after* the experiment has been carried out (in which case $H(P)$) measures the amount of information we got from the experiment.

certain quantity in course of the solution of a particular problem, this in itself does not prove that the quantity in question is of general importance; only the fact that one encounters the same quantity in solving a large number of rather different problems (as is the case e.g. with *Shannon*'s measure of information) convinces us that it is a fundamental notion. In this case however it is quite natural to try to find out the reason why the same quantity occurs in different contexts, i.e. to find which are its properties which make it so useful; and exactly this is the aim of the axiomatic approach.

In §§ 2–5 we shall deal with the pragmatic approach and in §§ 6–9 with the axiomatic approach, while in § 10 we add some further remarks.

2. A PRAGMATIC APPROACH

The entropy $H(P)$ of a finite discrete probability distribution $P = (p_1, p_2, \ldots, p_a)$ presents itself in the solution of the following simple coding problem. Let us consider a sequence of independent identically distributed random variables $\xi_1, \xi_2, \ldots, \xi_n$ each of which takes on the different values $x_1, x_2, \ldots, x_a$ with the corresponding probabilities $p_1, p_2, \ldots, p_a$ i.e.

$$P(\xi_j = x_k) = p_k \qquad (1 \leqq k \leqq a;\ 1 \leqq j \leqq n) \tag{2.1}$$

where $p_k \geqq 0$ $(1 \leqq k \leqq a)$ and $\sum_{k=1}^{a} p_k = 1$. The sequence $\xi_1, \xi_2, \ldots, \xi_n$ may be interpreted as produced by an information source emitting stationary and independent signals. Let Ω_n be the set of all ordered sequences of length n of the symbols $x_1, x_2, \ldots, x_a$. Let be given a fixed number ε $(0 < \varepsilon < 1)$ and let us consider those subsets E of Ω_n for which $P_n(E) \geqq 1 - \varepsilon$ where $P_n(E)$ is the probability that the observed sequence $\xi_1, \xi_2, \ldots, \xi_n$ belongs to the set E. Let $b(n, \varepsilon)$ denote the minimum of the number of elements of such sets E; by other words if $N(E)$ denotes the number of elements of a set E, we put

$$b(n, \varepsilon) = \min_{P_n(E) \geqq 1-\varepsilon} N(E). \tag{2.2}$$

Now it is easy to show that the limit

$$\lim_{n \to +\infty} \frac{\log_2 b(n, \varepsilon)}{n} = H(P) \tag{2.3}$$

exists, it is independent of ε, and it depends only on the distribution P, namely one has

$$H(P) = \sum_{k=1}^{a} p_k \log_2 \frac{1}{p_k}. \tag{2.4}$$

Still more is known; it has been shown recently by *W. Strassen* [3] (improving a previous result by *Jushkewich* [4]) that – if the numbers p_k are not all equal[3] one has

$$\log_2 b(n, \varepsilon) = n H(P) + \sqrt{n}\,.\,\lambda\, D - \frac{\log n}{2 \log 2} + O(1) \tag{2.5}$$

where

$$D = \left[\sum_{k=1}^{a} \left(\log_2 \frac{1}{p_k} - H(P)\right)^2 p_k\right]^{\frac{1}{2}}$$

[3] If $p_1 = p_2 = \ldots = p_a = \frac{1}{a}$ then clearly $b(n, \varepsilon)$ is equal to the least integer $\geqq a^n (1 - \varepsilon)$ and thus, as in this case $H(P) = \log_2 a$, we have $\log_2 b(n, \varepsilon) = n H(P) + O(1)$.

and λ is defined by $\Phi(\lambda) = 1 - \varepsilon$, where $\Phi(x)$ denotes the standard normal distribution function $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-u^2/2}\, du$; in (2.3) $O(1)$ denotes a remainder term[4] which remains bounded for $n \to +\infty$.

Thus we may consider (1.3) as a definition of the entropy $H(P)$. The result can be interpreted as follows: if we want to express (code) the sequence of signals $\xi_1, \xi_2, \ldots, \xi_n$ by a sequence of 0-s and 1-s so that the correspondence should be one to one, except for certain rather improbable sequences which we are ready to neglect if the total probability of occurrence of these neglected sequences does not exceed ε, then this can be accomplished by using sequences of 0-s and 1-s of length $n\,H(P) + O(\sqrt{n})$.

Thus if we accept as the unity of the amount of information the maximal amount of information which a signal capable of only two values (e.g. 0 and 1) can carry, then $H(P)$ can be interpreted as the amount of information per signal produced by a stationary source of independent signals, if the probability distribution of the possible values of the signals is $P = (p_1, p_2, \ldots, p_a)$.

As it is easy to see

$$H(P) \leqq \log_2 a \tag{2.6}$$

with equality standing in (2.6) if and only if $p_1 = p_2 = \ldots = p_a = \frac{1}{a}$.

An important point in favour of accepting (2.3) as the definition of the entropy $H(P)$ is that the amount of information per signal furnished by a more general channel (e.g. a stationary Markovian source, see *Jushkewich* [4]) further the definition of the capacity of a channel can be defined in a quite analogous way; as regards the latter question we refer to the coding theorem and its strong converse (see [2], [3]). Here and in what follows we restrict ourselves to discuss the entropy of a stationary source of independent signals.

3. ANOTHER PRAGMATIC APPROACH

While the simple coding problem discussed in § 1, the solution of which leads us to *Shannon*'s entropy $H(P)$, is certainly of fundamental importance, nevertheless there are other problems which are equally of fundamental character and which may also be used as an alternative for introducing in a pragmatic manner the entropy $H(P)$. In this § we shall discuss another problem of this kind, which may be called the *problem of random search* (see [5], [6], [7], [8]).

Let us consider the following situation: we want to find an unknown element x of a set S_n having n elements. The information available is of the following nature: we can carry out independent experiments, each experiment consisting in dividing at random the set S_n into a classes $C_1, C_2, \ldots, C_a$ so that for any element of S_n the probability that it will belong to the class C_k is equal to p_k ($k = 1, 2, \ldots, a$) independently from what happens to the other elements. By other words we suppose further that if the elements of S_n are $x_1, x_2, \ldots, x_n$ and λ_j denotes the index of the class into which x_j belongs, then the random variables $\lambda_1, \lambda_2, \ldots, \lambda_n$ are independent and $P(\lambda_j = k) = p_k$ ($1 \leqq j \leqq n$; $1 \leqq k \leqq a$). The result of the experiment consists in that we are informed which one of the classes $C_1, C_2, \ldots, C_a$ contains the unknown element x. Let be given a number ε ($0 < \varepsilon < 1$) and let $d(n, \varepsilon)$ denote the

[4] *Strassen* has analyzed this term further, but we do not go into these details here.

least such value of m that the probability that the unknown element can be uniquely determined from m experiments of the above mentioned type, exceeds $1-\varepsilon$. It is easy to see that $d(n, \varepsilon)$ is the least number m such that

$$(3.1)\qquad \sum_{\substack{\sum_{i=1}^{a} m_i = m}} \frac{m!}{m_1!\, m_2! \ldots m_a!} p_1^{m_1} \cdot p_2^{m_2} \ldots p_a^{m_a} (1 - p_1^{m_1} \cdot p_2^{m_2} \ldots p_a^{m_a})^{n-1} \geqq 1 - \varepsilon.$$

It follows by an easy calculation that putting

$$(3.2)\qquad D = \left[\sum_{k=1}^{a} p_k \left(\log_2 \frac{1}{p_k} - H(P) \right)^2 \right]^{1/2}$$

we have in case $D > 0$

$$(3.3)\qquad d(n, \varepsilon) = \frac{\log_2 n}{H(P)} + \frac{D \lambda \sqrt{\log_2 n}}{H(P)^{3/2}} + O(\sqrt{\log_2 n})$$

where λ is determined by the equation $\Phi(\lambda) = 1 - \varepsilon$.

Thus we have

$$(3.4)\qquad H(P) = \lim_{n \to +\infty} \frac{\log_2 n}{d(n, \varepsilon)}.$$

In case $D = 0$, i.e. if $p_1 = p_2 = \ldots = p_a = \frac{1}{a}$ we have for $0 < \varepsilon < 1 - \frac{1}{e}$

$$(3.5)\qquad d(n, \varepsilon) = \left\{ \frac{\log_2 n + \log_2 \frac{1}{\log \frac{1}{1-\varepsilon}}}{H(P)} \right\}$$

where $\{x\}$ denotes the least integer $\geqq x$ (and of course $H(P) = \log_2 a$).

The result obtained can be interpreted as follows: to determine an unknown element of a set S_n, having n elements, one needs $\log_2 n$ bits of information. If S_n is divided at random into a classes having approximately $p_1 n, p_2 n, \ldots, p_a n$ elements, and we are informed which of these classes contains the unknown element which we try to find, then each such experiment gives us in the average $H(P)$ bits of information, and thus approximately $\frac{\log_2 n}{H(P)}$ such experiments have to be carried out in order to get a sufficient amount of information. (The reason why slightly more than $\frac{\log_2 n}{H(P)}$ experiments are needed if we want to determine the unknown x with a probability $\geqq 1 - \varepsilon$, is that there is some overlapping between the partial informations obtained.)

The number $d(n, \varepsilon)$ can be interpreted also in the following alternative way: Let us consider again a stationary information source producing independent signals, with values $x_1, x_2, \ldots, x_a$ having the corresponding probabilities $p_1, p_2, \ldots, p_a$. Suppose that the source produces n sequences of signals, each sequence consisting of m signals; let us denote these sequences by

$$s_1 = \{\xi_{11}, \ldots, \xi_{1m}\}, s_2 = \{\xi_{21}, \ldots, \xi_{2m}\}, \ldots, s_n = \{\xi_{n1}, \ldots, \xi_{nm}\}.$$

Then $d(n, \varepsilon)$ is the least value of m such that the probability that the sequences $s_2, s_3, \ldots, s_n$ are all different from the sequence s_1, exceeds $1 - \varepsilon$.

4. AN OTHER MEASURE OF INFORMATION

We shall deal now with a modification of the problem considered in § 3 (see [7]) which leads to a measure of information, different from that of *Shannon*. The problem discussed in § 3 can be characterized as follows: we divide a finite set S_n into a classes at random several times, and we are interested in the number m of such subdivisions which is necessary in order that the m divisions together should separate a fixed (but unknown) element x of S_n from all other elements of S_n with a prescribed probability. (We shall say that a subdivision of S_n into $a \geqq 2$ classes separates two elements x and y of S_n if x and y belong to different classes of the subdivision.) Now we modify the problem as follows: how many independent subdivisions of the same type are needed in order that these subdivisions should separate any two element of S_n from each other, with a prescribed probability.

We suppose again that the subdivisions in question are independent from each other and are such that each element of S_n will belong to the k-th class with probability p_k independently of what happens to the other elements (i.e. if $x_1, \ldots, x_n$ are the elements of S_n and λ_j denotes the index of the class into which x_j belongs, then the random variables $\lambda_1, \ldots, \lambda_n$ are independent and $P(\lambda_j = k) = p_k$ for $1 \leqq j \leqq n$, $1 \leqq k \leqq a$). Let P_{nm} denote the probability that m such subdivisions separate any two elements of S_n. P_{nm} can be expressed as follows. Let us consider all possible ordered products consisting of m factors, each of which is one of the numbers $p_1, p_2, \ldots, p_a$, i.e. the products of the form $p_{i_1} p_{i_2} \ldots p_{i_m}$. There are clearly a^m such products (two products containing the same factors but in a different order are considered as different). Let us label these products in any order by the numbers $1, 2, \ldots, a^m$ and denote them by $w_1, w_2, \ldots, w_{a^m}$. Then we have

$$P_{nm} = \Sigma' w_{j_1} w_{j_2} \ldots w_{j_n} \tag{4.1}$$

where the summation is to be extended over all positive ordered n-tuples $j_1, j_2, \ldots, j_n$ which consist of different numbers chosen among the numbers $1, 2, \ldots, a^m$; thus the sum on the right contains $\binom{a^m}{n} n!$ terms. It can be proved that if $e_2(n, \varepsilon)$ denotes the least integer m such that $P_{nm} \geqq 1 - \varepsilon$ then

$$e_2(n, \varepsilon) = \left\{ \frac{2 \log_2 n + \log_2 \dfrac{1}{\log \dfrac{1}{1-\varepsilon}}}{I_2(P)} \right\} \tag{4.2}$$

where

$$I_2(P) = \log_2 \frac{1}{\left(\sum_{k=1}^{a} p_k^2 \right)}. \tag{4.3}$$

The quantity $I_2(P)$ which in formula (4.2) plays the same role as the entropy $H(P)$ in formula (3.3), can also be considered as a measure of the amount of information contained in the value of a discrete random variable ξ which has the distribution $P = (p_1, \ldots, p_a)$.

The relation between $H(P)$ and $I_2(P)$ can be characterized as follows. Let $P = (p_1, \ldots, p_a)$ be any finite probability distribution; let λ be a random variable which takes on the values $1, 2, \ldots, a$ with the corresponding probabilities $p_1, p_2, \ldots, p_a$ and let π be the random variable defined by $\pi = p_\lambda$; by other words π is always equal to the probability of that value of λ which λ takes on. Thus we have

$$P(\pi = p_k) = p_k \qquad (k = 1, 2, \ldots, a). \tag{4.4}$$

Now let us put

$$L(x) = \log_2 \frac{1}{x} \tag{4.5}$$

and let us denote by $E(\eta)$ the expectation of the random variable η. With these notations we have

$$H(P) = E(L(\pi)) \tag{4.6a}$$

and

$$I_2(P) = L(E(\pi)) . \tag{4.6b}$$

Thus $H(P)$ is obtained by transforming π by the function L and then calculating the expectation of this random variable, while $I_2(P)$ is obtained by carrying out the same operations in the reversed order.

$I_2(P)$ has certain properties in common with $H(P)$. For instance both are additive with respect to taking the direct product of probability distributions; more exactly if $P = (p_1, \ldots, p_a)$ and $Q = (q_1, \ldots, q_b)$ are two arbitrary probability distributions, and if we denote by $P * Q$ the distribution consisting of the terms $P_i q_j$ $(1 \leqq i \leqq a;\ 1 \leqq j \leqq b)$ then

$$H(P * Q) = H(P) + H(Q) \tag{4.7}$$

and similarly

$$I_2(P * Q) = I_2(P) + I_2(Q) \tag{4.8}$$

It follows from Jensen's inequality and the convexity of the function $\log_2 \frac{1}{x}$ that the values of $H(P)$ and $I_2(P)$ are equal if and only if $p_1 = p_2 = \ldots = p_a = \frac{1}{a}$; in every other case one has $I_2(P) < H(P)$.

5. A CLASS OF MEASURES OF INFORMATION

Both $H(P)$ and $I_2(P)$ considered in the previous §-s belong to the same class of measures of information, defined as follows (see [9] and [10]): Let α be any fixed positive number, $\alpha \neq 1$; we shall call the quantity

$$I_\alpha(P) = \frac{1}{1-\alpha} \log_2 \left(\sum_{k=1}^{a} p_k^\alpha \right) \tag{5.1}$$

the *measure of order* α *of the amount of information contained* in a value of a random variable having the finite discrete probability distribution $P = (p_1, \ldots, p_a)$. For $\alpha = 1$ we define $I_1(P)$ as the limit of $I_\alpha(P)$ for α tending to 1; we have evidently

$$I_1(P) = \lim_{\alpha \to 1} I_\alpha(P) = H(P) \tag{5.2}$$

Clearly for $\alpha = 2$ we get the same quantity $I_2(P)$ which we encountered in (3.2).

The problem considered in the previous § can be generalized so that we encounter instead of $I_2(P)$ the measure $I_r(P)$ of (integral) order r of information. As a matter of fact, let us consider again m independent random subdivisions $\Delta_1, \ldots, \Delta_m$ of the finite set S_n having n elements, so that each subdivision splits the set S_n into a classes, so that each element of S_n should be contained in the k-th class of any of the subdivisions with probability p_k $(k = 1, 2, \ldots, a)$ independently of what happens to the other elements. The m subdivisions together define a subdivision Δ of the set S_n which

may be called the product of the subdivisions Δ_j ($j = 1, \ldots, m$) and denoted by $\Delta = \Delta_1 . \Delta_2 \ldots \Delta_m$. The subdivision Δ is defined as follows: two elements of S_n belong to the same class of Δ if and only if they belong to the same class with respect to each subdivision $\Delta_1, \Delta_2, \ldots, \Delta_m$. Now let $e_r(n, \varepsilon)$ ($r \geqq 2, 0 < \varepsilon < 1$) denote the least value of m such that with probability $\geqq 1 - \varepsilon$ each class of $\Delta = \Delta_1 . \Delta_2 \ldots \Delta_m$ contains less than r elements. Clearly $e_r(n, \varepsilon)$ is for $r = 2$ equal to $e_2(n, \varepsilon)$ defined in the previous §. Now it can be proved (though the proof is rather complicated) that for $r = 2, 3, \ldots$ we have

$$\lim_{n \to \infty} \frac{\log_2 n}{e_r(n, \varepsilon)} = (1 - \frac{1}{r}) I_r(P) = \frac{1}{r} \log_2 \frac{1}{\sum_{k=1}^{a} p_k^r} . \tag{5.3}$$

As it is easy to see the right hand side of (5.3) is an increasing function of r.

Another closely related relation containing $I_r(P)$ which however is easy to prove, is as follows: let $f_r(n)$ denote the least value of m such that the expected number of such r-tuples of elements of S_n which belong to the same class of Δ does not exceed 1; then we have

$$f_r(n) = \frac{\log_2 n}{(1 - \frac{1}{r}) I_r(P)} + O(1) \tag{5.4}$$

As a matter of fact, if $v_r(m, n)$ denotes the number of r tuples of elements of S_n which belong to the same class of $\Delta = \Delta_1 \ldots \Delta_m$, then we have evidently

$$E(v_r(m, n)) = \binom{n}{r} \left(\sum_{k=1}^{a} p_k^r \right)^m \tag{5.5}$$

and (5.5) clearly implies (5.4).

The quantity $I_\alpha(P)$ has the same additivity property as $H(P)$ for each positive value of α; that is for any two probability distribution P and Q we have

$$I_\alpha(P * Q) = I_\alpha(P) + I_\alpha(Q) \tag{5.6}$$

If $p_1 = p_2 = \ldots = p_a = \frac{1}{a}$ then $I_\alpha(P) = \log_2 a$ for each $\alpha > 0$. For other distributions $I_\alpha(P)$ is a decreasing function of α.

As regards other properties of $I_\alpha(P)$ these will be discussed in connection with the axiomatic characterization of measures of information.

6. AXIOMATIC CHARACTERIZATION OF SHANNON'S MEASURE OF INFORMATION

Different sets of postulates have been found which characterize *Shannon*'s entropy $H(P)$ (defined by (1.1) uniquely). The original set of postulates given by *Shannon* [1] himself was later somewhat simplified by *Chintschin* and *Faddeew* [11], [12]. In this last form these postulates are as follows: Let Π denote the set of all finite, discrete probability distributions $P = (p_1, \ldots, p_a)$ ($p_k \geqq 0$ ($k = 1, \ldots, a$) and $\sum_{k=1}^{a} p_k = 1$). Let us suppose that a function $I(P) = I[p_1, \ldots, p_a]$ is defined for all $P \in \Pi$ which satisfies the following conditions:

A) $I[p, 1 - p]$ is continuous for $0 \leqq p \leqq 1$ and $I[\frac{1}{2}, \frac{1}{2}] = 1$.

B) $I[p_1, \ldots, p_a]$ is a symmetrical function of its variables.

C) If $0 \leqq \vartheta < 1$, we have

$$I[p_1, \ldots, p_{a-1}, \vartheta p_a, (1-\vartheta) p_a] = I[p_1, p_2, \ldots, p_a] + p_a I[\vartheta, 1-\vartheta]$$

Then $I(P) = H(P)$ where $H(P)$ is the entropy of the distribution P defined by (1.1). (For a simple proof see [12]).

The proof depends mainly on the following number theoretical lemma due to *P. Erdös* [13]: If $f(n)$ is an additive number theoretical function (i.e. $f(n\,m) = f(n) + f(m)$) further if $\lim_{n \to +\infty} (f(n+1) - f(n)) = 0$ then $f(n) = c \log n$ where c is a constant. (For simple proofs see [14] and [15]).

Note that postulate C) implies that $I(P)$ has the additivity property

$$I(P * Q) = I(P) + I(Q) \tag{6.1}$$

but C) is not implied by (6.1). As a matter of fact C) is not valid for $I_\alpha(P)$ with $\alpha \neq 1$. However C) can be replaced by (6.1) and by some other supplementary condition. For instance *Chaundy* and *McLeod* [16] have remarked that if we suppose that $I(P)$ is of the form $I(P) = \sum_{k=1}^{a} F(p_k)$ where $F(x)$ is continuous in the interval [0,1] and $F(\frac{1}{2}) = \frac{1}{2}$ and if $I(P)$ satisfies (6.1) then $F(p) = p \log_2 \frac{1}{p}$ and thus $I(P) = H(P)$ (see also [17] where it is shown that it is sufficient to suppose the validity of (6.1) for the case when P and Q have an equal number of terms).

Another such condition has been given in [9], for functions $I(P)$ defined for the wider class of *generalized probability distributions*, including also incomplete distributions.

Let Π^* denote the set of all finite sequences $P = (p_1, \ldots, p_a)$ of nonnegative numbers such that $0 < w(P) = \sum_{k=1}^{a} p_k \leqq 1$. We shall call every $P \in \Pi^*$ a generalized distribution and $w(P) = \sum_{k=1}^{a} p_k$ the *weight* of the generalized distribution P. A distribution $P \in \Pi^*$ will be called a complete probability distribution if $w(P) = 1$ and an incomplete probability distribution if $w(P) < 1$. The natural extension of Shannon's formula for $P \in \Pi^*$ is

$$H(P) = \frac{\sum_{k=1}^{a} p_k \log_2 \frac{1}{p_k}}{\sum_{k=1}^{a} p_k} \tag{6.2}$$

The direct product $P * Q$ is defined in the same way for generalized probability distribution as for ordinary (complete) probability distributions. If $P \in \Pi^*$ and $Q \in \Pi^*$ where $P = (p_1, \ldots, p_a)$, $Q = (q_1, \ldots, q_b)$ and $w(P) + w(Q) \leqq 1$ we put $P \cup Q = (p_1, \ldots, p_a, q_1, \ldots, q_b)$. (If $w(P) + w(Q) > 1$ then $P \cup Q$ is not defined.)

Now we have shown that if $I(P) = I[p_1, \ldots, p_a]$ is defined for $P \in \Pi^*$ such that

A*) $I[p]$ is continuous for $0 < p \leqq 1$, and $I[\frac{1}{2}] = 1$

B*) $I[p_1, \ldots, p_a)$ is a symmetric function of its variables,

C_1*) $I[P * Q) = I(P) + I(Q)$, further

C_2*) if $w(P) + w(Q) \leqq 1$ then

$$I(P \cup Q) = \frac{w(P)\, I(P) + w(Q)\, I(Q)}{w(P) + w(Q)}$$

then $I(P) = H(P)$ where $H(P)$ is defined by (6.2).

An incomplete probability distribution P can be interpreted as the probability distribution of a random variable whose value can not be observed always only with probability $w(P)$.

Note that the extension of the notion of information (entropy) to incomplete distributions has among others the advantage that it has a sense to speak about the entropy of a single event with probability p, this being equal to $\log_2 \frac{1}{p}$.

Thus introducing this extension of the notion of entropy one can say that the entropy of a probability distribution is equal to the weighted mean of the entropies of its terms, the weights being the probabilities themselves.

7. AXIOMATIC CHARACTERIZATION OF MEASURES OF INFORMATION OF ORDER α

By weakening to some extent the axiom C) one can obtain a characterization of the measures of information of order α. These investigations show that in a certain sense there are no other measures of information besides those of order $\alpha \neq 1$ which have similar properties as *Shannon*'s measure of information. This was conjectured by the author in [9], and his conjecture was proved by *J. Aczél* and *Z. Daróczy* in [17]. Their result has been recently sharpened by *Z. Daróczy* [18].

We reproduce here only this last result which is as follows: If $I(P)$ is defined for $P \in \Pi$, satisfies the condition (6.1) further there exists a function $f(x)$ such that

$$I(P) = \log_2 \frac{1}{M(P)} \tag{7.1}$$

with

$$M(P) = f^{-1}\left(\sum_{k=1}^{a} p_k f(p_k)\right) \tag{7.2}$$

where $f(x)$ is a strictly monotonic function, $\lim_{x \to 0} x f(x) = 0$ and $f(x)$ is continuous for $0 < x \leqq 1$, then $f(x)$ is either a linear function or a linear function of an exponential function and correspondingly either $I(P) = H(P) = I_1(P)$ or $I(P) = I_\alpha(P)$ with $\alpha \neq 1$, $\alpha > 0$.

Note that the continuity of $I(P)$ follows from the continuity of $x f(x)$ while the symmetry of $I(p_1, \ldots, p_n)$ from the symmetric form of $M(P)$.

The expression $M(P)$ is called by *J. Aczél* and *Z. Daróczy* a mean value of the probability distribution P.

The mentioned theorem can be stated also in the following form: a mean value $M(P)$ of the distribution $P = (p_1, \ldots, p_a)$ which is of the form (7.2) where $f(x)$ is strictly monotonic and $x f(x)$ continuous for $0 \leqq x \leqq 1$ and which has the multiplicative property

$$M(P * Q) = M(P) \cdot M(Q) \tag{7.3}$$

is necessarily either of the form

$$M(P) = \prod_{k=1}^{a} p_k^{p_k} \tag{7.4}$$

or of the form

$$M(P) = \left(\sum_{k=1}^{a} p_k^{\alpha}\right)^{\frac{1}{\alpha - 1}} \text{ with } \alpha > 0,\ \alpha \neq 1 \tag{7.5}$$

Shannon's entropy can be uniquely characterized as follows: Let $R = \{r_{jk}\}$ $(1 \leqq j \leqq a;\ 1 \leqq k \leqq b)$ be any distribution whose projections are the distribution $P = \{p_1, \ldots, p_a\}$ and $Q = (q_1, \ldots, q_b)$, i.e.

$$\sum_{k=1}^{b} r_{jk} = p_j \ (1 \leqq j \leqq a) \text{ and } \sum_{j=1}^{a} r_{jk} = q_k \ (1 \leqq k \leqq b).$$

We shall write $R = [P, Q]$

If $r_{jk} = p_j \cdot q_k$, i.e. if $R = P * Q$ then one has for every

$$\alpha > 0 \quad I_\alpha(R) = I_\alpha(P) + I_\alpha(Q).$$

However the inequality

$$I_\alpha(R) \leqq I_\alpha(P) + I_\alpha(Q) \tag{7.6}$$

holds for any $R = [P, Q]$, if and only if $\alpha = 1$.

8. CHARACTERIZATION OF THE MEASURES OF INFORMATION OF ORDER α OF GENERALIZED PROBABILITY DISTRIBUTIONS

In analogy with (6.2) the measure of information $I_\alpha(P)$ of order α $(\alpha > 0, \alpha \neq 1)$ of a generalized distribution $P \in \Pi^*$ $(P = \{p_1, \ldots, p_a\})$ is defined by

$$I_\alpha(P) = \frac{1}{1-\alpha} \log_2 \frac{\sum_{k=1}^{a} p_k^{\alpha}}{\sum_{k=1}^{a} p_k}. \tag{8.1}$$

Note that if $P \in \Pi^*$ consists of a single term i.e. $P = \{p\}$ we have for each value of α

$$I_\alpha(\{p\}) = \log_2 \frac{1}{p} \tag{8.2}$$

Z. Daróczy [19] has shown that the quantities (6.2) and (8.1) can be characterized as follows: Let $I(P) = I[p_1, \ldots, p_a]$ be defined for $P \in \Pi^*$ and suppose that $I(P)$ satisfies the following postulates:

A*) $I[p]$ is a continuous function of p in the interval $0 < p \leqq 1$, and $I[\frac{1}{2}] = 1$

C_1*) $I(P * Q) = I(P) + I(Q)$

C_3*) There exists a strictly monotonic and continuous function $g(x)$ such that if $P \in \Pi^*$, $Q \in \Pi^*$ further $w(p) + w(Q) \leqq 1$ then

$$I(P \cup Q) = g^{-1}\left[\frac{w(P)\, g(I(P)) + w(Q)\, g(I(Q))}{w(P) + w(Q)}\right]$$

Then $g(x)$ is either a linear function or a linear function of an exponential function and correspondingly either $I(P) = H(P)$ or $I(P) = I_\alpha(P)$ with some *real* $\alpha \neq 1$. *J. Aczél* [20] has simplified the proof of this theorem. He remarked also that the undesirable functions $I_\alpha(P)$ with $\alpha < 0$ can be excluded by the additional postulate:

D*) $$\lim_{q \to 0} I[p, q] = I[p].$$

9. AXIOMATIC CHARACTERIZATION OF INFORMATION WITHOUT USING PROBABILITIES

An interesting attempt has been made recently by *R. S. Ingarden* and *K. Urbanik* (see [21], [22]). They gave an axiomatic characterization of *Shannon*'s measure of infor-

mation, which does not presuppose the notion of probability. Information is defined as a real valued function on a set X of finite Boolean rings. Each ring $R \in X$ is interpreted as an experiment, the elements of the ring as events. As regards X it is supposed that if R_1 is a ring belonging to X and R_2 a nontrivial subring of R_1 then R_2 belongs to X also, further if $R_1 \in X$ there exists an other ring $R_2 \in X$ such that R_1 is a proper subring of R_2. With respect to any real valued function F defined on X a distance function $\varrho_F(R_1, R_2)$ is defined as follows: $\varrho_F(R_1, R_2) = 1$ if R_1 and R_2 are non-isomorphic, while $\varrho_F(R_1, R_2) = \delta_F(R_1, R_2) / 1 + \delta_F(R_1, R_2)$ if R_1 and R_2 are isomorphic, where $\delta_F(R_1, R_2) = \min_{\varphi} \max_{R} | F(R) - F(\varphi(R)) |$ where R runs over all subrings of R_1 and φ over all isomorphisms of R_1 onto R_2. X is a pseudometric space with respect to ϱ_F. A ring $R \in X$ is called F-homogeneous if for every automorphism ψ of R and for every subring R_1 of R we have $F(R_1) = F(\psi(R_1))$. Let X_F denote the class of all F-homogeneous rings from X and their subrings. The function F is called regular if X_F is dense in X with respect to the metric ϱ_F. Let as usual $\cup$ resp. $\cap$ denote union (joint) resp. intersection (meet) and $-$ difference of elements of a Boolean ring. For any $A \in R$ let us put $\bar{A} = 1(R) - A$ where $1(R)$ is the unity element of R.

If R is a ring and $A \in R$ let $A\,R$ denote the subring of R consisting of all $B \in R$ which can be written in the form $B = A \cap C$ with $C \in R$. If $R \in X$ and $A \in R$ let $R \mid A$ denote the least Boolean ring containing A and all elements of $\bar{A}\,R$ where $\bar{A} = 1(R) - A$. ($R \mid A$ can be interpreted as the experiment which differs from the experiment R only in that the outcomes belonging to A are "pooled".)

A real valued function $H = H(R)$ defined for $R \in X$ is called an information if it satisfies five axioms of which the first is as follows:

I) $(H(R) - H(\tilde{R}_3))\,H(R_1)\,H(R_2) = (H(R) - H(\tilde{R}_1))\,H(R_2)\,H(R_3) + (H(R) - H(\tilde{R}_2))\,H(R_1)\,H(R_3)$

Here R is any element of X, $R_1 = A\,R$, $R_2 = B\,R$ where A and B are any disjoint elements $\neq 0$ of R, $R_3 = (A \cup B)\,R$, $\tilde{R}_1 = R \mid A$, $\tilde{R}_2 = R \mid B$, $\tilde{R}_3 = R \mid A \cup B$.

The second axiom can be formulated as follows:

II) If $R_1 \in X$, $R_2 \in X$, $A \in R_1$, $B \in R_2$ further if $\varrho_H(A\,R_1, B\,R_2) = 0$ and $\varrho_H(R_1 \mid \bar{A}, R_2 \mid \bar{B}) = 0$ (here of course $\bar{B}$ stands for $1(R_2) - B$) then

$$H(R_1) - H(R_1 \mid A) = H(R_2) - H(R_2 \mid B).$$

The next two axioms postulate the monotonicity of information $H(R_2) < H(R_1)$ if R_2 is a proper subring of R_1, further that isomorphic H-homogeneous rings have ϱ_H distance 0. The last axiom normalizes information by postulating that if R is an H-homogeneous ring having two atoms then $H(R) = 1$.

Ingarden and *Urbanik* proved the following theorem:

If H is an information on X then for every $R \in X$ there exists a unique strictly positive probability measure $P(A \mid R)$ defined for $A \in R$ such that for every subring R_1 of R and for every $B \in R_1$ one has $P(B \mid R_1) = \dfrac{P(B \mid R)}{P(1(R_1) \mid R)}$ where $1(R_1)$ is the unity element of R_1, and $H(R) = \Sigma\, P(A \mid R) \log_2 \dfrac{1}{P(A \mid R)}$ where the base of the logarithm is 2 and A runs over all atoms of R.

By other words the information as defined by the axioms of the authors (which do

not presuppose the notion of probability) can be expressed by *Shannon*'s well known formula by means of a uniquely defined (conditional) probability measure, the existence of which follows from the existence of information. Thus the axiomatic approach of the authors shows that though the usual logical order, according to which information is defined by means of probability can be reversed, and one can introduce information first, without using probabilities, probabilities come in however inevitably at a later stage.

It seems to us that the fact that a theory which starts with the aim to define information without probability leads to the proof of the existence of (conditional) probabilities, supports the view that the notion of information can not be separated from that of probability. The situation becomes particularly clear if we adopt the point of view of § 8, i.e. we consider also the information corresponding to an incomplete distribution (i.e. to an experiment whose result is not always observable.)

In this case to each event A there correspond two numbers: its probability $P(A)$ and its information content $I(A)$, which are connected by the formulae

$$I(A) = \log_2 \frac{1}{P(A)}, \quad P(A) = 2^{-I(A)} \tag{9.1}$$

Thus it does not matter which of $P(A)$ or $I(A)$ is taken as fundamental, the other can be expressed by it.

An interesting feature of the result is that the probabilities whose existence is deduced from that of information are not ordinary but conditional probabilities in the sense of the author's paper [23].

Let us add some remarks concerning the meaning of the axioms of *Ingarden* and *Urbanik*.

Clearly the axiom C) of *Faddeew* can be expressed with the present notations as follows:

$$H(R) = H(R \mid A) + P(A)\, H(A\, R) \tag{9.2}$$

Thus if $H(A\, R) > 0$, we have

$$P(A) = \frac{H(R) - H(R \mid A)}{H(A\, R)} \tag{9.3}$$

Now in the light of the formula (9.3) the meaning of the axioms becomes clear. As a matter of fact if we define the probability $P(A)$ by (9.3) then evidently the first axiom expresses the additivity of probability. As regards the second axiom it expresses the fact, that while in the definition (9.3) of $P(A)$ there occurs the ring R which can be chosen arbitrarily with the single restriction that it should contain the event A, nevertheless the value of $P(A)$ does not depend on the choice of R. Thus axioms I) and II) together imply axiom C). In view of this it is now clear why the axioms of *Ingarden* and *Urbanik* characterize *Shannon's* measure of information.

The axioms of *Ingarden* and *Urbanik* can be formulated also in term of finite algebras of sets instead of Boolean rings. We do not go into the details here.

10. CONCLUDING REMARKS

In addition to what has been said in the previous § about the relation of the notions of probability and information, one remark has to be added. While when speaking about a simple event A, both its probability $P(A)$ and its information content $I(A)$

are positive numbers[5], connected by formula (9.1) further in this case the measures of information of order α have all the same value, the situation changes when we are dealing with a system of (disjoint) events $(A_1, A_2, \ldots, A_a)$. In this case we have a corresponding probability distribution $P = (P(A_1), P(A_2), \ldots, P(A)_a)$ being a set of numbers (or to put it otherwise, the probability distribution is vector-valued).

On the other hand we have different measures of information $I_\alpha(P)$ corresponding to the distribution P, all of which are *number-valued*. All these measures of information are obtained from the *information-distribution* $J = (I(A_1), I(A_2), \ldots, I(A_a))$ corresponding to the events $A_1, A_2, \ldots, A_a$ by some method of averaging, i.e. they are all mean values. This explains why different measures of information are possible: the situation is the same as taking different (arithmetic, geometric, harmonic etc.) mean values of a set of numbers. Shannon's measure of information is clearly the simplest among them, and plays a role analogous to that of the ordinary (linear) mean value. Nevertheless in certain situations the other non-linear mean values of information (i.e. the informations of order $\alpha > 0$, $\alpha \neq 1$) may also be useful.

REFERENCES

[1] Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27, 379–423, 623–653.

[2] Wolfowitz, J. (1961). Coding theorems of information theory. *Ergebnisse der Mathematik und ihre Grenzgebiete*, Springer, Berlin-Göttingen-Heidelberg, 1–125.

3] Strassen, W. (1962). Asymptotic expansions in information theory. *Colloquium on combinatorial methods in probability theory*, Aarhus, 74–77.

[4] Jushkewich, A. (1958). On a limit theorem connected with the notion of the entropy of a Markov chain. *Uspechi Mat. Nauk*, 8, p. 5 (in Russian).

[5] Rényi, A. (1961). On random generating elements of a finite Boolean algebra. *Acta Scientarium Mathematica Szeged*, 22, 75–81.

[6] Rényi, A. (1961). Statistical laws of accumulation of information. *Bulletin de l'Institut International de Statistique, 33rd Session*, Paris, 1–7.

[7] Rényi, A. (1962). On statistical laws of the accumulation of information. *Magyar Tudományos Akadémia III. Osztélyának Közleményei*, 12, 15–33.

[8] Rényi, A. (1961). On a problem of information theory. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, 6, 505–516. (in Hungarian with English and Russian summaries)

[9] Rényi, A. (1961). On measures of entropy and information. *Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability*. University of California Press, Vol. I, 547–561.

[10] Rényi, A. (1962). Wahrscheinlichkeitsrechnung, mit einem Anhang über Informationstheorie. *Deutscher Verlag der Wissenschaften*, Berlin, 435–498.

[11] Chintschin, A. J. (1957). Der Begriff der Entropie in der Wahrscheinlichkeitsrechnung, Arbeiten zur Informationstheorie, I. *Deutscher Verlag der Wissenschaften*, Berlin, 7–25.

[12] Faddeew, D. K. (1957). Zum Begriff der Entropie eines endlichen Wahrscheinlichkeitsschemas, Arbeiten zur Informationstheorie. *Deutscher Verlag der Wissenschaften*, Berlin, 88–90.

[13] Erdös, P. (1946). On the distribution function of additive functions. *Annals of Mathematics*, 47, 1–20.

[14] Rényi, A. (1960). On a theorem of P. Erdös and its application in information theory. *Mathematica*, 1, 341–344.

[15] Besicovitch, A. S. (1962). On additive functions of a positive integer. *Studies in Mathematical Analysis and related topics*, Stanford University Press, Stanford, 38–41.

[16] Chaundy, T. W. and McLeod, J. B. (1960). On a functional equation. *Edinburgh Mathematical Notes*, 43, 7–8.

[17] Aczél, J. and Daróczy, Z. (1963). Charakterisierung der Entropien positiver Ordnung und der Shannonschen Entropie. *Acta Mathematica Academiae Scientiarium Hungaricae*, 14, 95–121.

[5] We exclude here for the sake of simplicity both impossible and sure events.

[18] Daróczy, Z. (1964). Über Mittelwerte und Entropien vollständiger Wahrscheinlichkeitsverteilungen. *Acta Mathematica Academiae Scientiarium Hungaricae*, 15, 203–210.
[19] Daróczy, Z. (1963). Über die gemeinsame Charakterisierung der zu den nicht vollständigen Verteilungen gehörigen Entropien von Shannon und von Rényi. *Zeitschrift für Wahrscheinlichkeitstheorie*, 1, 381–388.
[20] Aczél, J. Zur gemeinsamen Charakterisierung der Entropien α-ter Ordnung und der Shannonschen Entropie nicht unbedingt vollständiger Verteilungen. *Zeitschrift für Wahrscheinlichkeitstheorie* (in print).
[21] Ingarden, R. S. and Urbanik, K. (1961). Information as a fundamental notion of statistical physics. *Bulletin de L'Académie des Sciences Polonaise, Ser. sci. math.*, 9, 313–316.
[22] Ingarden, R. S. and Urbanik, K. (1962). Information without probability. *Colloquium Mathematicum*, 9, 131–150.
[23] Rényi, A. (1955). A new axiomatic theory of probability. *Acta Mathematica Academiae Scientiarium Hungaricae*, 6, 285–335.

RESUME

L'auteur expose les différentes définitions de la quantité d'information. On peut distinguer deux types de définitions. Les définitions *axiomatiques* partent de certaines propriétés plausibles qu'une mesure d'information raisonnable doit posséder; ensuite intervient le problème d'ordre purement mathématique de trouver les expressions ayant les propriétés postulées. Les autres définitions peuvent êtres appellées des définitions *pragmatiques*: on prend comme point de départ certains problèmes concrèts de la théorie de l'information; en resolvant ces problèmes, on constate que certaines expressions figurent dans la solution, et par conséquent on considère ces expressions comme des quantités d'information. L'auteur montre que les deux points de vue ne sont pas antagonistes; bien plus, ils se complètent l'un l'autre, et l'on peut arriver à la quantité d'information de Shannon, ainsi qu'aux quantités d'information d'ordre α introduites par l'auteur, en partant aussi bien du point de vue axiomatique, que du point de vue pragmatique.

DISCUSSION

C. R. RAO: Is there any special advantage in choosing $\alpha = 1/2$ in the measure of information?

$$P_\alpha(p) = \frac{1}{1-\alpha} \log_2 \sum_1^k p_i{}^\alpha .$$

It was found useful in theory of statistical inference as a measure of distance between $(p_1, \ldots p_k)$ and the uniform distribution $(\frac{1}{k} \ldots \frac{1}{k})$.

A. RÉNYI: In answer to the question of Professor Rao, I should like to mention that in a previous paper I made reference to the work of Indian mathematicians concerning the case $\alpha = 1/2$ in connection with the information – theoretical distance of two probability distributions. In the present paper, however, I dealt only with the amount of information connected with a single distribution, and it seems to me that in this case the information of order 1/2 does not play the same exceptional role.

D. G. KENDALL: In § 6 of his paper Prof. Rényi recalls the theorem of Fadeev, that Shannon's function is uniquely characterised by the conditions A, B, and C. The assumption that $h(t) = I(t, 1-t)$ is continuous for $0 < t < 1$ is very natural, but the more severe assumption (used by Fadeev) that $h(.)$ is continuous for $0 \leqslant t \leqslant 1$ is not so natural; it excludes the possibility of solutions for which $h(t) \to -\infty$ when $t \downarrow 0$. A different argument (due to H. Tverberg) replaced the continuity assumption of Fadeev by the requirement that $h \in L(0, 1)$, but this is not at all natural, and is still undesirably restrictive. I have shown that Shannon's function is uniquely characterised by *B* and *C*, the condition that $h(1/2) = 1$, and the condition that $h(.)$ is to be non-decreasing for $0 < t \leqslant 1/2$. In a sequel to this work Mr. P. M. Lee has shown that Shannon's function is uniquely characterised by *B* and *C*, the condition that $h(1/2) = 1$, and the condition that $h(.)$ is to be *Lebesgue measurable* on (0.1). It seems likely that Lee's theorem is incapable of further improvement, in the sense that there may exist non-measurable solutions to the functional equations, but attempts to construct such non-measurable entropies have so far proved unsuccessful. It is not, of course, suggested that the question of their existence has any practical relevance.