ON THE AMOUNT OF INFORMATION CONCERNING AN UNKNOWN PARAMETER IN A SEQUENCE OF OBSERVATIONS

by

ALFRED RÉNYI

Introduction

Let \( \xi_1, \xi_2, \ldots, \xi_n, \ldots \) be a sequence of identically distributed random variables. We suppose that the common distribution of the random variables \( \xi_n \) depends on a parameter \( \Theta \), which itself is also a random variable. In the present paper we restrict ourselves to the case when \( \Theta \) is a discrete random variable; we shall denote the possible different values of \( \Theta \) by \( t_1, t_2, \ldots, t_r \). We shall suppose further that the random variables \( \xi_n \) \((n = 1, 2, \ldots)\) are independent and identically distributed under each of the conditions \( \Theta = t_j \) \((j = 1, 2, \ldots, r)\) and denoting by \( F_j(x) \) the conditional distribution function of the variables \( \xi_n \) under the condition \( \Theta = t_j \), we suppose that \( F_j(x) \) and \( F_h(x) \) are not identical if \( j \neq h \). Let \( P(\Theta = t_j) = w_j \) \((j = 1, \ldots, r)\) denote the prior probability of \( \Theta \) taking on the value \( t_j \). Without restricting the generality we may suppose \( w_j > 0 \) for \( j = 1, 2, \ldots, r \). We denote by \( I_n = I_n((\xi_1, \ldots, \xi_n), \Theta) \) the amount of information concerning the (unknown) value of \( \Theta \) contained in the sequence of random variables \( \xi_1, \xi_2, \ldots, \xi_n \). As well known (see [1], [2], [3]) \( I_n \) is defined by

\[
I_n = I_n((\xi_1, \ldots, \xi_n), \Theta) = H(\Theta) - M(H(\Theta | \xi_1, \ldots, \xi_n)) .
\]

Here and in what follows for any (discrete) random variable \( \eta \) we denote by \( H(\eta) \) the entropy of \( \eta \), i.e. if \( y_1, \ldots, y_r \) are the values taken on by \( \eta \) and \( P(\eta = y_j) = p_j \) \((j = 1, 2, \ldots, r)\) we put

\[
H(\eta) = \sum_{j=1}^{r} p_j \log \frac{1}{p_j} .
\]

(Here and in what follows \( P(A) \) denotes the probability of the event \( A \) and \( \log \) denotes the logarithm with base 2.) If \( \zeta \) is an other random variable we denote by \( H(\eta | \zeta) \) the conditional entropy of \( \eta \) given the value of \( \zeta \), i.e. we put

\[
H(\eta | \zeta) = \sum_{j=1}^{r} P(\eta = y_j | \zeta) \log \frac{1}{P(\eta = y_j | \zeta)}
\]

where \( P(A | \zeta) \) denotes the conditional probability of the event \( A \) given the value of \( \zeta \). \( H(\eta | \zeta) \) is of course itself a random variable, as its value depends on the value taken on by \( \zeta \). We denote by \( M(\ldots) \) the mean value of the random variable in the brackets. Thus \( M(H(\Theta | \xi_1, \ldots, \xi_n)) \) denotes the average
uncertainty which remains with respect to \( \Theta \) after observing the values \( \xi_1, \xi_2, \ldots, \xi_n \).

The first author suggesting the use of the quantity \( I_n \) in statistical problems was — according to our knowledge — D. V. LINDLEY [4].

The aim of the present paper is to prove that for \( n \to \infty \) \( I_n \) tends to \( H(\Theta) \) in such a way that there exist constants \( A > 0 \) and \( q \) (\( 0 < q < 1 \)) so that

\[
0 \leq H(\Theta) - I_n \leq A q^n
\]

and to consider some consequences of this fact.

§ 1. Estimation of the amount of information obtained from observations

We shall restrict ourselves to the case when the random variables \( \xi_k \) have a finite discrete distribution. For the sake of the simplicity of notation we suppose that the possible values of the variables \( \xi_k \) are the integers \( 1, 2, \ldots, s \).

This is no essential restriction, because if the distribution of the variables \( \xi_k \) under condition \( \Theta = t_j \) is different for each \( j \), then one can find a Borel-measurable function \( g(x) \) taking on only a finite number of different values, such that the conditional distribution of the variables \( g(\xi_k) = \xi_k^* \) under condition \( \Theta = t_j \) is different for each \( j \) and one has evidently

\[
I_n((\xi_1, \ldots, \xi_n), \Theta) \geq I_n((\xi_1^*, \ldots, \xi_n^*), \Theta).
\]

Thus if we prove that

\[
0 \leq H(\Theta) - I_n((\xi_1^*, \ldots, \xi_n^*), \Theta) \leq A q^n
\]

this implies

\[
0 \leq H(\Theta) - I_n((\xi_1, \ldots, \xi_n), \Theta) \leq A q^n
\]

Moreover one has

\[
I_n((\xi_1, \ldots, \xi_n), \Theta) = \sup_g I_n((\xi_1^*, \ldots, \xi_n^*), \Theta),
\]

where the supremum has to be taken over all possible choices of the function \( g \).

By the way, this remark shows, that if we observe the random variables \( \xi_k \) with the single aim of collecting information with respect to \( \Theta \), we may round off in an appropriate way the observed values \( \xi_n \) with no serious loss of information.

We shall need an elementary inequality, contained in the following

**Lemma.** There exists a universal constant \( C > 0 \) such that for any sequence \( p_1, \ldots, p_N \) of positive numbers forming a probability distribution (i.e. for which \( p_1 + p_2 + \ldots + p_N = 1 \)) we have

\[
\sum_{k=1}^{N} p_k \log \frac{1}{p_k} \leq C \sum_{k=2}^{N} \sqrt{p_k}.
\]
Proof of the lemma. Clearly both \( x \log \frac{1}{x} \) and \( (1 - x) \log \frac{1}{1 - x} \) are continuous in the closed interval \([0, 1]\). Putting

\[ (1.6) \quad C_1 = \max_{0 \leq x \leq 1} \frac{x \log \frac{1}{x}}{x} \]

and

\[ (1.7) \quad C_2 = \max_{0 \leq x \leq 1} \frac{(1 - x) \log \frac{1}{1 - x}}{x} \]

we have

\[ (1.8) \quad \sum_{k=2}^{N} p_k \log \frac{1}{p_k} \leq C_1 \sum_{k=2}^{N} \sqrt{p_k} \]

and

\[ (1.9) \quad p_1 \log \frac{1}{p_1} \leq C_2 \sqrt{\sum_{k=2}^{N} p_k} \leq C_2 \sum_{k=2}^{N} \sqrt{p_k} . \]

Thus (1.5) follows with \( C = C_1 + C_2 . \)

Now we are in the position to prove (4). We have

\[ (1.10) \quad H(\Theta) - I_n = \sum_{h=1}^{r} w_h \mathcal{M}(H(\Theta|\xi_1, \ldots, \xi_n)\mid \Omega_h) \]

where \( \Omega_h \) denotes the subset of the full probability space \( \Omega \) on which \( \Theta = t_h \) \((h = 1, 2, \ldots, r)\) and \( \mathcal{M}(\eta|B) \) denotes conditional expectation with respect to \( B \), i.e.

\[ (1.11) \quad \mathcal{M}(\eta|B) = \frac{1}{\mathcal{P}(B)} \int_B \eta \, d\mathcal{P}. \]

On the other hand by supposition, putting

\[ (1.12) \quad \mathcal{P}(\xi_m = x|\Theta = t_j) = P_j(x) \quad (x = 1, 2, \ldots, s) \]

we have by Bayes theorem

\[ (1.13) \quad \mathcal{P}(\Theta = t_j|\xi_1, \ldots, \xi_n) = \frac{w_j \prod_{k=1}^{n} P_j(\xi_k)}{\sum_{l=1}^{r} w_l \prod_{k=1}^{n} P_l(\xi_k)} \leq \frac{w_j \prod_{k=1}^{n} P_j(\xi_k)}{w_h \prod_{k=1}^{n} P_h(\xi_k)} . \]

It follows by our lemma

\[ (1.14) \quad H(\Theta|\xi_j, \ldots, \xi_n) \leq C \sum_{j \neq h}^{r} \left[ \frac{w_j}{w_h} \prod_{k=1}^{n} \frac{P_j(\xi_k)}{P_h(\xi_k)} \right] . \]
and thus in view of the conditional independence of the variables $\xi_k$ on $\Omega_h$

$$M(H(\Theta \mid \xi_1, \ldots, \xi_n) \mid \Omega_h) \leq C \sum_{j=1}^{r} \sqrt{\frac{w_j}{w_h}} \left[ M\left(\left|\frac{P_j(\xi_1)}{P_h(\xi_1)}\right| \mid \Omega_h\right)\right]^n.$$  

Now clearly

$$M\left(\left|\frac{P_j(\xi_1)}{P_h(\xi_1)}\right| \mid \Omega_h\right) = \sum_{i=1}^{s} \sqrt{P_j(l) P_h(l)} = q_{jh}.$$  

By Cauchy's inequality and in view of the supposition that the distributions $\{P_j(l)\}$ and $\{P_h(l)\}$ are different for $j \neq h$ we obtain $0 \leq q_{jh} < 1$. Thus putting

$$q = \max_{1 \leq j < h \leq r} q_{jh}$$  

we have $q < 1$, and it follows from (1.10), (1.15), (1.16) and (1.17), further from (1.14) that

$$M(H(\Theta \mid \xi_1, \ldots, \xi_n)) \leq Aq^n$$  

with

$$A = C \sum_{j=1}^{r} \sum_{h \neq j}^{r} \sqrt{w_j w_h} \leq C(r - 1).$$

Thus we have proved the following

**Theorem 1.** Let $\Theta$ be a discrete random variable taking on $r$ different values $t_1, t_2, \ldots, t_r$ with positive probabilities $w_j = P(\Theta = t_j)$ ($j = 1, 2, \ldots, r$). Let us suppose that the discrete random variables $\xi_1, \xi_2, \ldots, \xi_n, \ldots$ are for each $j$ ($1 \leq j \leq r$) independent and identically distributed under the condition that $\Theta = t_j$ and let us suppose that the conditional distribution of the variables $\xi_n$ under condition $\Theta = t_j$ is different for each $j$. Then, denoting by $I_n$ the amount of information on $\Theta$ obtained from observing the values of $\xi_1, \xi_2, \ldots, \xi_n$, we have

$$\lim_{n \to \infty} I_n = H(\Theta) = \sum_{j=1}^{r} w_j \log \frac{1}{w_j};$$

moreover the rate of convergence in (1.20) is exponential, i.e. there exist positive constants $A$ and $q < 1$ such that

$$0 \leq H(\Theta) - I_n \leq Aq^n \quad (n = 1, 2, \ldots).$$

As $H(\Theta)$ is the full information missing on $\Theta$ [i.e. the total uncertainty concerning $\Theta$ before observing the values of $\xi_k$ ($k = 1, 2, \ldots$)] the result (1.20) expresses that the sequence of observations $\xi_1, \xi_2, \ldots, \xi_n, \ldots$ gives us — at least in the limit — total information on the unknown value of the parameter. This leads one to guess that there exists a decision function depending only on the observed values $\xi_1, \ldots, \xi_n, \ldots$ which leads *almost surely* to a correct decision concerning the value of $\Theta$. This will be shown in the next §.
§ 2. Construction of an almost surely reliable decision function

If we have to decide after observing $\xi_1, \xi_2, \ldots, \xi_n$ what is the most plausible value $t_j$ of $\Theta$, the most natural decision is to decide for that value $t_j$ of $\Theta$ for which $P(\Theta = t_j | \xi_1, \ldots, \xi_n)$ is maximal. Note that this is not quite the same as the decision suggested by the maximum likelihood principle which suggests to choose that value $t_j$ of $\Theta$ for which $\prod_{k=1}^{n} P_j(\xi_k)$ is the largest, though for large values of $n$ the two decisions coincide with probability near to 1. Let us put

$$A_n = t_j \text{ if } \max_h P(\Theta = t_h | \xi_1, \ldots, \xi_n) = P(\Theta = t_j | \xi_1, \ldots, \xi_n).$$

(Of course the decision $A_n$ is not necessarily unambiguous but as will be seen below it will be such with probability 1 if $n$ is sufficiently large.) Thus if we take a sequence of observations $\xi_1, \xi_2, \ldots, \xi_n, \ldots$ we get a sequence of decisions $A_1, A_2, \ldots, A_n, \ldots$ Clearly to each sequence of observations $\xi_1, \ldots, \xi_n, \ldots$ there corresponds a "true" value of $\Theta$. Now we prove

**Theorem 2.** With probability 1 all but a finite number of the decisions $A_n$ are unambiguous and equal to the true value of the parameter $\Theta$.

**Proof of Theorem 2.** Evidently

$$P(A_n \neq \Theta) \leq \sum_{h=1}^{r} \sum_{j \neq h} w_h P\left( \frac{P(\Theta = t_j | \xi_1, \ldots, \xi_n)}{P(\Theta = t_h | \xi_1, \ldots, \xi_n)} \geq 1 \right) \Omega_h.$$ 

By Markoff's inequality:

$$P(\eta \geq \lambda M(\eta)) \leq \frac{1}{\lambda}$$

valid for any nonnegative random variable $\eta$ and any $\lambda > 1$, and in view of

$$P(\Theta = t_j | \xi_1, \ldots, \xi_n) = \frac{w_j}{w_h} \prod_{k=1}^{n} \frac{P_j(\xi_k)}{P_h(\xi_k)}$$

and taking (1.16) into account, we obtain

$$P(A_n \neq \Theta) \leq \sum_{h=1}^{r} w_h \sum_{j \neq h} \sqrt{\frac{w_j}{w_h} q_{jn}^n} \leq B q^n$$

where

$$B = \sum_{h=1}^{r} \sum_{j \neq h} \sqrt{w_j w_h} = \left( \sum_{j=1}^{r} \sqrt{w_j} \right)^2 - 1 \leq r - 1$$

and $q$ is defined by (1.17).

Thus it follows that the series

$$\sum_{n=1}^{\infty} P(A_n \neq \Theta)$$
is convergent. Applying the Borel–Cantelli lemma it follows that with probability 1 $A_n = \Theta$ for sufficiently large values of $n$. This proves Theorem 2.

Of course if $v$ denotes the least integer such that $A_n = \Theta$ for $n \geq v$ then we get from Theorem 2 that the random variable $v$ is almost everywhere finite, but this does not mean that $v$ is bounded. However it is easy to prove that the expectation of $v$ is finite. As a matter of fact

$$P(v \geq m) \leq \sum_{n=m}^{\infty} P(A_n \neq \Theta) \leq \frac{Bq^n}{1 - q}$$

and thus the series

$$M(v) = \sum_{m=1}^{\infty} P(v \geq m)$$

is convergent; from (2.7) we even get the estimate

$$M(v) \leq \frac{Bq}{(1 - q)^2}.$$  

Of course this does not mean that one can determine the true value of $\Theta$ from a finite number of observations with probability 1, because even if $A_n$ is equal to the same value $t_j$ for $n \geq M$, if we stop observing the $\xi_n$'s with the observation $\xi_{M+N}$ we can not be quite sure whether $A_n$ would have remained equal to $t_j$ if we would have continued the observations. Of course the probability of this event will be arbitrarily small if $M + N$ is sufficiently large.

For the discussion of the case when $\Theta$ has a continuous distribution other methods are needed. We intend to return to this problem in an other paper.

§ 3. Some further remarks

In the present paper we have taken the Bayesian point of view, which — as has been pointed out already by Lindley (see [4]) — is the natural point of view if one wants to compute the amount of information on a parameter furnished by some observations, this quantity being dependent on the prior distribution of this parameter. In this connection we want to emphasize that as the inequalities

$$0 \leq H(\Theta) - I_n \leq C(r - 1)q^n$$

and

$$P(A_n \neq \Theta) \leq (r - 1)q^n$$

hold uniformly for all possible prior distributions $\{w_j\}$, our results are meaningful even in the case when nothing is known about the prior distribution.

It should be added finally that the random variables $\xi_1, \xi_2, \ldots, \xi_n, \ldots$ considered in this paper, while being independent on each subspace $\Omega_h$ of the total probability space $\Omega$, are dependent on $\Omega$, but in a particularly simple way: they are equivalent (symmetrically dependent); as a matter of fact if $n_1 < n_2 < \ldots < n_k$ we have

$$P(\xi_{n_1} = x_1, \xi_{n_2} = x_2, \ldots, \xi_{n_k} = x_k) = \sum_{h=1}^{r} w_h \prod_{l=1}^{k} P_h(x_l)$$
and thus this probability does not depend on the choice of the indices \( n_1, n_2, \ldots, n_k \).

Thus the values \( \xi_1, \ldots, \xi_n, \ldots \) may be considered as a sequence of signals emitted by a stationary source of a particularly simple type which may be called a \textit{source with equivalent signals}. We shall return to the corresponding problem for an arbitrary stationary source in another paper.

Finally we want to call attention to the information-theoretical meaning of the quantities \( q_j \) defined by (1.16). Some years ago we have introduced (see [5], [6], [7]) the notion of the entropy of order \( a \) \((0 < a < +\infty)\) of a probability distribution \( \mathcal{F} = (p_1, \ldots, p_r) \) which is defined for \( a \neq 1 \) by the formula

\[
H_a(\mathcal{F}) = \frac{1}{1-a} \log \left( \sum_{k=1}^{r} p_k^a \right).
\]

We define \( H_1(\mathcal{F}) \) by passing to the limit in (3.1), i.e. we put

\[
H_1(\mathcal{F}) = \lim_{a \to 1} H_a(\mathcal{F}) = \sum_{k=1}^{r} p_k \log \frac{1}{p_k}.
\]

Thus \( H_1(\mathcal{F}) \) is equal to Shannon's entropy of the distribution \( \mathcal{F} \).

Similarly we have introduced the notion of "information gain" of order \( a \) obtained if the prior distribution \( \mathcal{F} = (p_1, \ldots, p_r) \) is replaced by the posterior distribution \( Q = (q_1, \ldots, q_r) \) defined for \( a \neq 1 \) by the formula

\[
I_a(Q \mid \mathcal{F}) = \frac{1}{a-1} \log \left( \sum_{k=1}^{r} q_k^a p_k^{1-a} \right).
\]

If we define \( I_1(Q \mid \mathcal{F}) \) again by passing to the limit in (3.3), we obtain

\[
I_1(Q \mid \mathcal{F}) = \lim_{a \to 1} I_a(Q \mid \mathcal{F}) = \sum_{k=1}^{r} q_k \log \frac{q_k}{p_k}
\]

i.e. \( I_1(Q \mid \mathcal{F}) \) is given by the familiar formula for information-gain.

Now clearly

\[
q_{jk} = 2^{-I_1(Q \mid \mathcal{F})}
\]

where \( P_j = \{P_j(1), \ldots, P_j(s)\} \) is the conditional distribution of the variables \( \xi_n \) under the condition \( \Theta = t_l \) \((l = 1, 2, \ldots, r)\).

Similarly the constant \( B \) figuring in (2.5) can be expressed by the formula

\[
B = 2^{H_1(W)} - 1
\]

where \( W = (w_1, \ldots, w_r) \) is the prior distribution of the random variable \( \Theta \).

Let us mention that the quantity \( I_{1/2}(Q \mid \mathcal{F}) \) as a measure of the discrepancy between the distributions \( Q \) and \( \mathcal{F} \) has been considered earlier by A. Bhattacharyya [8] and H. Jeffreys [9].

Thus the situation can be characterized as follows: though we accepted as a measure of the amount of information in the sample \((\xi_1, \ldots, \xi_n)\) Shannon's usual measure (of order \( a = 1 \)) nevertheless in course of the proof the
measure of order $a = \frac{1}{2}$ of information gain and of entropy resp. entered our
considerations for purely technical reasons. (See also [10], [11] and [12].)

Note that instead of $a = 1/2$ any value of $a$ such that $0 < a < 1$ could
have been used. The value of $a$ which leads to the least value of $q$ depends in
general on the conditional distributions $P_l (l = 1, 2, \ldots, r)$. However, in the
present paper we did not aim at finding the least possible value of $q$ only to
prove that such a value less than 1 exists, therefore for the sake of simplicity
we have chosen $a = \frac{1}{2}$.

(Received November 20, 1964.)

REFERENCES

[1] RÖLMGÖROFF, A. N.: “Theorie der Nachrichtenübermittlung.” Arbeiten zur Infor-

tion der Information.” Arbeiten zur Informationstheorie II. 57—60. Deutscher

Informationstheorie II. 61—64. Deutscher Verlag der Wissenschaften, Berlin, 1958.


Köz. 10 (1960) 251—282.

Conference on Information theory (1960) 545—556.


on the sum of observations.” Bull. Amer. Math. Soc. 23 (1952) 493—507.


О КОЛИЧЕСТВЕ ИНФОРМАЦИИ ОТНОСИТЕЛЬНО НЕИЗВЕСТНОГО
ПАРАМЕТРА В ПОСЛЕДОВАТЕЛЬНОСТИ НАБЛЮДЕНИЙ

А. РЕНЬИ

Резюме

Пусть $ξ_1, ξ_2, \ldots, ξ_n, \ldots$ последовательность случайных величин, któ-
рые при всяком данном значении параметра $θ$ независимы и одинаково рас-
пределены. Предположим, что сама $θ$ случайная величина с дискретным
распределением $P(θ = t_j) = w_j (j = 1, 2, \ldots, r; \sum_{j=1}^{r} w_j = 1)$. В работе доказы-
вается следующая теорема: Существуют постоянные $A$ и $q (A > 0, 0 < q < 1)$,
так что, если \( I_n \) обозначает количество информации в выборке \((\xi_1, \xi_2, \ldots, \xi_n)\) относительно \( \Theta \) и \( H \) означает энтропию от \( \Theta \), тогда имеет место неравенство

\[
0 \leq H - I_n \leq Aq^n
\]

Из этого следует, что для достаточно большого числа \( n \) решение \( A_n \) правильно с вероятностью 1. Здесь \( A_n \) означает решение, что после наблюдения значений \( \xi_1, \xi_2, \ldots, \xi_n \) мы выберем как «истинное» значение от \( \Theta \) то значение \( t_j \), для которого условная вероятность \( P(\Theta = t_j \mid \xi_1, \ldots, \xi_n) \) максимальна.