INFORMATION ON AN UNKNOWN PROBABILITY CONTAINED IN RELATIVE FREQUENCY

By

J. REIMANN

Department of Civil Engineering Mathematics, Technical University, Budapest, H-1521.

(Received: July 14, 1981)

Summary

Two methods are presented for computing the information content of the relative frequency of an event on its unknown probability. The first method relies on the Shannon information theory in Bayes-notion and is based on a method described by Rényi. The second method starts from Kalmár's "qualitative information theory" where the amount of information given by the relative frequency computed from a statistical sample is measured by the shortening ratio of the confidence interval of probability. It is demonstrated that the amount of information in the elements of a sample is not uniformly distributed but decreases in geometrical proportion in the successive experiments.

RÉNYI described [2] how to use the *Shannon* information theory to solve a given statistical decision problem.

The classical methods of parametrical hypothesis analysis in mathematical statistics (parametrical tests) are known to generally concern the decision between two hypotheses. If X is a continuous random variable with distribution function $F(x; \Theta)$ where Θ is a parameter of the distribution (e.g. expected value) then generally the null hypothesis $H_0 : \Theta = \theta_0$ on the value of parameter Θ is tested against an alternative hypothesis $H_1 : \Theta = \theta_1$.

In many practical problems it has to be decided between several simultaneously possible hypotheses, for instance the possible values of parameter Θ are supposed to be numbers $\theta_1, \theta_2, \ldots, \theta_k$, that is, the hypotheses $H_1 : \Theta = \theta_1$; $H_2 : \Theta = \theta_2; \ldots; H_k : \Theta = \theta_k$ are set and it has to be decided for one of them. The decision is based on a statistical sample of random variable X such that

(I) $\quad \vec{\mathbf{X}} : x_1, x_2, \ldots, x_n$ ($\vec{\mathbf{X}}$ is a vector of n numbers).

Sample (I) contains all the information to decide which of the values $\theta_1, \theta_2, \ldots, \theta_k$ is the real value of Θ. In order to apply the information theory (at least the *Shannon* theory) for this multi-value decision problem, the standpoint of *Bayes* has to be adopted, namely to take parameter Θ for random variable with possible values $\theta_1, \theta_2, \ldots, \theta_k$ assumed by some $p_1, p_2, \ldots, p_k$ where $\sum_{i=1}^{k} p_i = 1$.

Consequently let

$$P(H_i) = P(\Theta = \theta_i) = p_i \qquad (i = 1, 2, \ldots, k).$$

The finite scheme

$$\Theta : \begin{pmatrix} \theta_1 & \theta_2 & \ldots & \theta_k \\ p_1 & p_2 & \ldots & p_k \end{pmatrix} \tag{1}$$

is called the *a priori* distribution of Θ.

The entropy of the finite

$$H(\Theta) = -\sum_{i=1}^{k} P(\Theta = \theta_i) \log^2 P(\Theta = \theta_i) = -\sum_{i=1}^{k} p_i \log^2 p_i \tag{2}$$

is the measure of the uncertainty concerning the distribution of Θ.

Knowledge of the statistical sample $\vec{X}$ permits to calculate the conditional probabilities

$$P(\Theta = \theta_1 | \vec{\mathbf{X}}),\ P(\Theta = \theta_2 | \vec{\mathbf{X}}),\ \ldots,\ P(\Theta = \theta_k | \vec{\mathbf{X}}) \tag{3}$$

and with their help the conditional entropy

$$H(\Theta | \vec{\mathbf{X}}) = -\sum_{i=1}^{k} P(\Theta = \theta_i | \vec{\mathbf{X}}) \log^2 P(\Theta = \theta_i | \vec{\mathbf{X}}). \tag{4}$$

The conditional probability distribution (3) is called *a posteriori* distribution of Θ based on the statistical sample $\vec{\mathbf{X}}$ of n observations of the random variable X.

Using the *Jensen* inequality it can be proved that

$$H(\Theta | \vec{\mathbf{X}}) \leq H(\Theta) \tag{5}$$

i.e. the uncertainty of the distribution of Θ is by no means increased by any knowledge of random variable X. If X is independent of Θ then its value does not inform on the distribution of Θ. In this case in (5) the sign of equality is effective, in any other case unequality is valid.

The conditional entropy $H(\Theta | \vec{\mathbf{X}})$ can be interpreted as the amount of still missing information on Θ after the observation of X.

Let us form the difference

$$I(\Theta; \vec{\mathbf{X}}) = H(\Theta) - H(\Theta | \vec{\mathbf{X}}) \tag{6}$$

representing the actual amount of information on Θ given by the observed value of X.

The conditional entropy $H(\Theta | X)$ is a random variable itself, as its value depends on the statistical sample $\vec{X} = (x_1, x_2, \ldots, x_n)$.

The expected value

$$R(\Theta | \vec{X}) = E[H(\Theta | \vec{X})] \tag{7}$$

can be calculated, and then the difference

$$\bar{I}(\Theta; \vec{X}) = H(\Theta) - R(\Theta|\vec{X}) \tag{8}$$

is the measure of the average information on Θ contained in the statistical sample of random variable X.

In statistical practice it is often advisable to calculate a statistics $= t(x_1, x_2, \ldots, x_n)$ of the sample $\vec{X} = (x_1, x_2, \ldots, x_n)$ that is a sufficient statistics for Θ (i.e. statistics t contains all the information in the sample on Θ). By means of this statistics t the magnitudes $H(\Theta|t)$ and $R(\Theta|t) = E[H(\Theta|t)]$ and the average information gain

$$\bar{I}(\Theta; t) = H(\Theta) - R(\Theta|t) \tag{9}$$

can be calculated and as *Rényi* proved it, if using sufficient statistics,

$$I(\Theta; \vec{X}) = I(\Theta; t). \tag{10}$$

Provided random variable X is of binomial distribution with parameter Θ, then in sample (I), $x_i = 1$ or $x_i = 0$ depending on whether event A occurs or not in the i^{th} trial ($i = 1, 2, \ldots, n$). In this case, relative frequency $\frac{k}{n}$ of event A is a sufficient statistics for Θ, and (9) yields the measure of information on the unknown probability $P(A) = \Theta$ given by the relative frequency of an event. These statements will be illustrated in practice by a hydrological example.

In course of the last 30 floods of the river *Tisza* at *Szeged* it happened 5 times in the second quarter of the year (1st April to 30th June) that the peak value exceeded the level $c = 650$ cm by more than 180 cm. The difference between the peak value and the c-level is called "excess". Let random variable X express the excess value. Now, what is the probability of the event $X \geq 180$ cm at the gauge station of Szeged in the second quarter of the year? During the last 70 years 30 excesses over the level $c = 650$ cm were recorded in the second quarters, the arithmetic mean of the excesses being $\overline{X} = 90$ cm.

Let us consider $\overline{X} = 90$ cm as the expected value of random variable X. According to the *Markov* inequality:

$$P\big(X \geq \lambda E(X)\big) \leq \frac{1}{\lambda}. \tag{11}$$

Choosing $\lambda = 2$

$$P\big(X \geq 2E(X)\big) = P(X \geq 180 \text{ cm}) \leq \frac{1}{2}. \tag{12}$$

Thus, considering Eq. (12) as starting information,

$$\Theta = P(X \geq 180 \text{ cm}) \leq \frac{1}{2}$$

may be taken as granted. Establishing hypotheses $H_1 : \Theta = \theta_1 = 0.1$; $H_2 : \Theta = \theta_2 = 0.2$; $H_3 : \Theta = \theta_3 = 0.3$; $H_4 : \Theta = \theta_4 = 0.4$ and deciding among them, the probable value of unknown Θ is approximated at about 10%.

Let us regard the finite scheme

$$\Theta : \begin{pmatrix} \theta_1 & \theta_2 & \theta_3 & \theta_4 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{pmatrix} \tag{13}$$

as *a priori* distribution of Θ (Θ is considered as random variable according to *Bayes*) and let's compute the entropy

$$H(\Theta) = -4 \cdot \frac{1}{4} \log^2 \frac{1}{4} = -\log^2 \frac{1}{4} = 2 \tag{14}$$

that is, the uncertainty concerning the distribution of Θ. In the last 30 floods (during the period in question) the relative frequency of event $\{X \geq 180 \text{ cm}\}$

$$t = \frac{k}{n} = \frac{5}{30} = \frac{1}{6} = 0.166 \ldots \approx 0.17$$

obviously contains information on the unknown random variable Θ, information to be quantified.

As in this case the relative frequency approaches $\Theta = \theta_2 = 0.2$, the assumption on the distribution of Θ is changed. It is not presumed any more that it takes any value with the same probability. Let us compute the conditional distribution of Θ under the condition $t = \frac{k}{n} = 0.17$, that is, determine the conditional probability

$$P\left(\Theta = \theta_1 \middle| \frac{k}{n}\right), \quad P\left(\Theta = \theta_2 \middle| \frac{k}{n}\right), \quad P\left(\Theta = \theta_3 \middle| \frac{k}{n}\right), \quad P\left(\Theta = \theta_4 \middle| \frac{k}{n}\right). \tag{15}$$

Since

$$P\left(\Theta = \theta_i \middle| t = \frac{k}{n}\right) P\left(t = \frac{k}{n}\right) = P\left(t = \frac{k}{n} \middle| \Theta = \theta_i\right) P(\Theta = \theta_i) \qquad (i = 1, 2, 3, 4)$$

according to the *Bayes*-formula:

$$P\left(\Theta = \theta_i \middle| t = \frac{k}{n}\right) = \frac{P\left(t = \frac{k}{n} \middle| \Theta = \theta_i\right) P(\Theta = \theta_i)}{\sum_{j=1}^{4} P\left(t = \frac{k}{n} \middle| \Theta = \theta_j\right) P(\Theta = \theta_j)}$$

where

$$P\left(t = \frac{k}{n} \middle| \Theta = \theta_i\right) = \binom{n}{k} \theta_i^k (1 - \theta_i)^{n-k}; \quad P(\Theta = \theta_i) = \frac{1}{4}, \quad i = 1, 2, 3, 4.$$

From the table of binomial distribution:

$$P\left(t = \frac{5}{30} \middle| \Theta = 0.1\right) = \binom{30}{5} 0.1^5 \cdot 0.9^{25} = 0.1023$$

$$P\left(t = \frac{5}{30} \middle| \Theta = 0.2\right) = \binom{30}{5} 0.2^5 \cdot 0.8^{25} = 0.1723$$

$$P\left(t = \frac{5}{30} \middle| \Theta = 0.3\right) = \binom{30}{5} 0.3^5 \cdot 0.7^{25} = 0.0464$$

$$P\left(t = \frac{5}{30} \middle| \Theta = 0.4\right) = \binom{30}{5} 0.4^5 \cdot 0.6^{25} = \frac{0.0042}{0.3252}.$$

Taking into consideration, that

$$P\left(\Theta = \theta_i \middle| t = \frac{k}{n}\right) = \frac{P\left(t = \frac{k}{n} \middle| \Theta = \theta_i\right)}{\sum_{j=1}^{4} P\left(t = \frac{k}{n} \middle| \Theta = \theta_j\right)}$$

consequently

$$P\left(\Theta = 0.1 \middle| t = \frac{5}{30}\right) = 0.3146$$

$$P\left(\Theta = 0.2 \middle| t = \frac{5}{30}\right) = 0.5228$$

$$P\left(\Theta = 0.3 \middle| t = \frac{5}{30}\right) = 0.1427$$

$$P\left(\Theta = 0.4 \middle| t = \frac{5}{30}\right) = \frac{0.0013}{0.9814 \approx 1}.$$

Let us compute the conditional entropy

$$(16)\quad H\left(\Theta \middle| t = \frac{k}{n}\right) = -\sum_{i=1}^{4} P\left(\Theta = \theta_i \middle| t = \frac{k}{n}\right) \log^2 P\left(\Theta = \theta_i \middle| t = \frac{k}{n}\right) = 1.43.$$

The gain of information

$$(17)\qquad I(\Theta; t) = H(\Theta) - H\left(\Theta \middle| t = \frac{k}{n}\right) = 2 - 1.43 = 0.57$$

shows, how much of information on the distribution of Θ was given by the relative frequency $t = \frac{5}{90}$. Knowledge of the relative frequency reduced the uncertainty about the actual value of Θ (at least for the rounded value above) by more than a quarter. The relative gain of information is:

$$\frac{I(\Theta;t)}{H(\Theta)} = 1 - \frac{1.43}{2} = 0.285. \tag{18}$$

The conditional entropy $H\left(\Theta \middle| t = \frac{k}{n}\right)$ takes ever different values depending on the k value, that is, $H\left(\Theta \middle| t = \frac{k}{n}\right)$ is a random variable itself.

Hence to answer the question, how much information is given by the relative frequency $\frac{k}{n}$ for the probability of parameter Θ, the event, then the expected value of conditional entropy $H\left(\Theta \middle| t = \frac{k}{n}\right)$

$$R(\Theta) = \sum_{k=0}^{n} H\left(\Theta \middle| t = \frac{k}{n}\right) P\left(t = \frac{k}{n}\right) \tag{19}$$

is to be computed, and the difference

$$I(\Theta) = H(\Theta) - R(\Theta) \tag{20}$$

yields the information on the probability of the event in question, generally given by the relative frequency.

To compute the expected value $R(\Theta)$ is rather cumbersome, not to be discussed here.

As for the decision problem described by *Rényi*, among hypotheses $H_1 : \Theta = \theta_1$; $H_2 : \Theta = \theta_2, \ldots$; $H_k : \Theta = \theta_k$ it is advisable to accept hypothesis H_i with the maximal conditional probability $P\left(\Theta = \Theta \middle| t = \frac{k}{n}\right)$.

Rényi called this decision standard decision and showed not to be other decision of smaller error than the standard decision.

According to *Bayes*, no error of first or second kind of the decision can be spoken of. A decision is wrong if not the true hypothesis H_i is accepted. In the example, the hypothesis $H_2 : \Theta = \theta_2 = 0.2$ is accepted. In this example, the probability of a standard decision to be wrong is:

$$\varepsilon = P(\Theta \neq \theta_2) = 0.3142 + 0.1427 + 0.0013 \approx 0.46.$$

It is a surprisingly high probability of error, as nearly every second case could be erroneous. In statistical practice, in testing hypotheses an error of 5% is accepted in general. The question arises, in case of a standard decision relying on information theory what a number of elements in a sample would ensure a decision error of about 5%. The question can be simply answered in the case of only two hypotheses: $H_1 : \Theta = \theta_1$ and $H_2 : \Theta = \theta_2$.

In this case either

$$P(\Theta = \theta_1 \mid \vec{x}) \geq 0.95 \quad \text{and} \quad P(\Theta = \theta_2 \mid \vec{x}) \leq 0.05 \tag{21}$$

or

$$P(\Theta = \theta_1 \mid \vec{x}) \leq 0.05 \quad \text{and} \quad P(\Theta = \theta_2 \mid \vec{x}) \geq 0.95 \tag{22}$$

must be true. Any of (21) or (22) is true:

$$H(\Theta \mid \vec{x}) \leq -(0.95 \log^2 0.95 + 0.05 \log^2 0.05) \approx 0.29. \tag{23}$$

Considering distribution $\Theta : \begin{pmatrix} \theta_1 & \theta_2 \\ 1/2 & 1/2 \end{pmatrix}$ as the *a priori* distribution of Θ, in case (21) is true

$$\frac{P(\Theta = \theta_1 | \vec{x})}{P(\Theta = \theta_2 | \vec{x})} = \frac{P(\vec{x} | \theta_1) P(\Theta = \theta_1)}{P(\vec{x} | \theta_2) P(\Theta = \theta_2)} = \frac{P(\vec{x} | \theta_1)}{P(\vec{x} | \theta_2)} \approx 19 \tag{24}$$

(the same holds if (22) is true), that is, sampling has to be continued until the conditional probability of the sample (or the satisfactory statistics computed of it) for one hypothesis is about 19 times that of the other one. This result is known from the sequential probability ratio test. From this point, Eqs (23) and (24) can be regarded as equivalent.

The result shows that it is possible to decide between two hypotheses with 95% confidence even if nearly one third of the starting (maximal) uncertainty persists. In case $k > 2$ the answer is complicated, at least according to the considerations described above. This problem will be referred to later.

Another question arises, whether the amount of information on the unknown probability contained in the statistical sample can only be measured by the *Shannon*-entropy or by other means, too.

L. KALMÁR described another possible way of setting up the information theory — he called it qualitative information theory —, unfortunately his work got discontinued, so it awaits to be completed. To set up qualitative information theory, *Kalmár* started from analysing the information offered by one symbol when writing down a number with the digits 0, 1, 2, 3, 4, 5, 6, 7, 8, 9 (each digit is a symbol).

First of all, some starting information is needed, namely what kind of number is in question. Generally this starting information is at disposal, trivially. If e.g. the probability $P(A) = p$ of an event A is to be given as decimal fraction then it is known to be one point in the interval $I_0 = [0.1]$, the integer part of the number is 0, it can be written down immediately without any other symbol. If the first decimal digit of the probability $P(A) = p$ in a question is 2, then the point is in the interval $I_1 = [0.2; 0.3)$ closed from the left side. As I_1 is a real subset of I_0, by giving the digit 2 (or any other digit), information is obtained, as point p is to be found in a shorter interval,

a smaller set. Now the pair of sets $[I_0, I_1)$ represents the information gain from digit 2. If the following digit of the decimal fraction is given, let it be e.g. 7, then point p is known to be in the interval $I_2 = [0.27; 0.28)$ closed from the left. I_2 is seen to be a real subset of I_1, consequently there is a gain of information obtained by being given the next symbol, digit 7, and borne by the pair of sets $[I_1, I_2)$. The same is true for giving subsequent digits. Upon giving each digit (symbol) the interval containing point p becomes ten times smaller. The number of symbols for indicating a point must exceed that of the digits 0, 1, 2, . . ., 9 imposing to introduce symbol $\triangledown$ indicating that no more digits are needed to give the decimal fraction wanted. Without further precisions of KALMÁR's considerations [1], only the idea is used here that the information on a wanted point of interval [0.1] is borne by a pair of sets $[I_{n-1}, I_n)$ when a symbol is given (it may but needs not be a digit).

No publication has been found on how to measure the amount of information in this pair of sets if to measure at all. One possible way is to introduce a set-function μ in the interval [0.1] to be valued $\mu(I) = 1$ for $I_n \subset I_{n-1}$ then obviously $\mu(I_n) \leq \mu(I_{n-1})$ and the information in the interval $[I_{n-1}, I_n)$ is expressed by the quotient $\frac{\mu(I_n)}{\mu(I_{n-1})}$ showing how many times the interval including point p is smaller.

If this measure μ is the *Lebesgue* measure then the measure of interval I_k equals its length. In this case the quotient $\frac{\mu(I_n)}{\mu(I_{n-1})}$ represents how many times the interval on the number axis including point p becomes shorter with the n^{th} digit given than it was previously, when only digits $0, a_1, a_2, \ldots, a_{n-1}$ were known. This quotient seems to be a natural measure for the information gain. In certain cases it may be advisable to choose a measure different from that by *Lebesgue* in the interval (0.1) for measuring information in the way described above.

Further in this paper an attempt is made to analyse the outlined problem in a similar way as *Kalmár*'s idea.

In the presented hydrologic example the starting information is that — according to the *Markov* inequality — the probability $P(X \geq 180 \text{ cm})$ does not exceed $\frac{1}{2}$, that is, Θ is a point in the interval $\left[0, \frac{1}{2}\right)$. The relative frequency of the event $\{X \geq 180 \text{ cm}\}$ was found to be $\frac{k}{n} = \frac{5}{30} \approx 0.17$. According to the *Moivre—Laplace* limit theorem, binomial distribution can be approximated by normal distribution; as $\sqrt{\Theta(1-\Theta)} \leq 1/2$, choosing $\lambda = 2$ yields:

$$P\left(\left|\frac{k}{n} - \Theta\right| \geq \frac{1}{\sqrt{n}}\right) \approx 0.05. \tag{25}$$

In conformity with (25):

$$\left(\frac{k}{n}-\frac{1}{\sqrt{n}};\ \frac{k}{n}+\frac{1}{\sqrt{n}}\right)=\left(\frac{5}{30}-\frac{1}{\sqrt{30}};\ \frac{5}{30}+\frac{1}{\sqrt{30}}\right)\approx$$

$$\approx(-0.0125;\ 0.3525)\simeq(0;\ 0.35)$$

is a confidence interval of about 95% for the unknown probability Θ.

The starting information located the unknown point Θ in interval $I_0=[0;\ 0.5]$; knowledge of the relative frequency $\frac{k}{n}=\frac{5}{30}$ located it in the shorter interval $I_1=(0;\ 0.35)$ namely $I_1\subset I_0$.

Consequently the pair of sets (I_0, I_1) contains information on Θ. Now the question is how to measure it. It seems to be useful to measure the amount of information with the factor of shortening the interval containing the unknown point Θ. In our example this factor is

$$\frac{|I_1|}{|I_0|}=\frac{0.35}{0.5}=0.7. \tag{26}$$

In this way the relative gain of information:

$$\frac{|I_0|-|I_1|}{|I_0|}=1-\frac{|I_1|}{|I_0|}=0.3 \tag{27}$$

is about the same (a little bit more) than the relative information obtained by *Shannon*-entropy. It would be useful to measure the amount of information in identical units, i.e. bits. (The numerical value of the *Shannon* entropy is in bits.)

The value

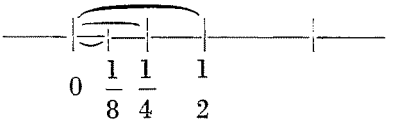
$$I(\Theta;t)=H(\Theta)-H(\Theta|t)$$

in (17) amounts to 0.57 bit. Now the question is, how many bits of information correspond to shortening factor $\frac{|I_1|}{|I_0|}=0.7$.

It can be answered by thinking over the following. If a point Θ of interval $(0;\ 1)$ is given in diadic decimal fraction form:

$$\Theta=0.0011101\ldots=0+0\cdot\frac{1}{2}+0\cdot\frac{1}{2^2}+1\cdot\frac{1}{2^3}+1\cdot\frac{1}{2^4}+\ldots$$

then giving each digit (0 or 1) halves the length of the interval containing the Θ value:



If the digit following the decimal point is 0, then Θ is in the interval $\left(0, \frac{1}{2}\right)$; if the next digit is 0 again, then it is in the interval $\left(0, \frac{1}{4}\right)$. If the next digit is 1, then Θ is in the interval $\left(\frac{1}{8}, \frac{1}{4}\right)$, etc. Going from one digit to the next one, the interval is halved; in case of digit 0 the point is in the next left-side half-interval, in case of digit 1, it is in the right-side half-interval. It means that giving one bit (binary signal) halves the interval containing the point in question, that is, by giving one bit: $\frac{(I_1)}{(I_0)} = \frac{1}{2} = 0.5$.

Obviously, the more the information, the smaller the ratio of the next to the preceding interval $\frac{(I_1)}{(I_0)}$. In our example $\frac{(I_1)}{(I_0)} = 0.7$ means that the relative frequency $\frac{5}{30}$ contains less than 1 bit of information on the unknown probability Θ. Its exact amount can be computed by inverse proportionality from the proportion:

$$x : 1 = 0.5 : 0.7; \quad x = \frac{0.5}{0.7} = 0.71 \text{ bit.}$$

This numerical result shows that somewhat more information is obtained on the unknown probability by using the method described above, than by computing conditional entropy, likely to result from not having rounded the possible values of Θ. Applying the "qualitative" information measurement yields a somewhat more detailed insight into the dependence of the amount of information on the probability of an event contained in a statistical sample on the number of elements in the sample. Let A be a certain event of an unknown probability Θ to be approximated through n independent experiments. A random variable X_i is assigned to each experiment so that $X_i = 1$ or $X_i = 0$ depending on whether A happened in the i^{th} experiment or not. (X_i is the indicator variable of event A in the i^{th} experiment.) A number n of experiments yield the statistical sample $X_1, X_2, \ldots, X_n$ to compute the relative frequency

$$\overline{X} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{k}{n}.$$

As a starting information, Θ is a point in the interval (0;1).

According to (25):

$$P\left(\left|\frac{k}{n} - \Theta\right| \geq \frac{1}{\sqrt{n}}\right) \approx 0.05$$

that is, in case $n = 16$, interval $\left(\frac{k}{n} - \frac{1}{4}; \frac{k}{n} + \frac{1}{4}\right)$ is the confidence interval of about 95% for the unknown probability. Thus, a relative frequency computed

after $n = 16$ experiments permits to find Θ in an interval half that before the experiments, that is, 16 experiments give 1 bit of information on Θ. If $n = 64$ then $\left(\frac{k}{n} - \frac{1}{8}; \frac{k}{n} + \frac{1}{8}\right)$ is a confidence interval of 95%, that is, further 48 experiments give one more bit of information on the value of Θ. If $n = 256$ then interval $\left(\frac{k}{n} - \frac{1}{16}; \frac{k}{n} + \frac{1}{16}\right)$ contains Θ with a confidence of 95%, that is, 192 new experiments are necessary to one more bit of information. To halve this interval, $1024 = 32^2$ experiments, that is, further 768 experiments are needed. Thus, to successively halve the intervals, samples of

$$\begin{array}{ccccccc} 16 & 64 & 256 & 1024 & 4096 & \dots \\ 4^2 & 8^2 & 16^2 & 32^2 & 64^2 & \dots \end{array}$$

elements are needed. The series of differences, that is, the number of further experiments necessary for one more bit of information:

$$48 \qquad 192 \qquad 768 \qquad 3072 \qquad \dots$$

These differences form a geometrical progression of quotient 4. It means that the information on the unknown probability Θ decreases in geometrical proportion in the elements of the sample. The "closer" the value of Θ is known, the more experiments are necessary to further refinements. The number of elements of the sample needed for one further bit of information increases very fast. New information is given very ungenerously, to get some more knowledge costs high in terms of number of elements in the sample.

References

1. KALMÁR, I.: Problems of the Qualitative Information Theory. (In Hungarian) MTA. III. Oszt. Közl. III/4. 1962
2. RÉNYI, A.: Statistics and Information Theory. Studia Scientiarum Math. Hung. 2 (1967) 249—256

Prof. Dr. József REIMANN, H-1521, Budapest.