On random generating elements of a finite Boolean algebra

By A. RÉNYI in Budapest

Dedicated to Professor L. Rédei on his 60th birthday

We consider finite Boolean algebras. As it is well known the number of elements of a finite Boolean algebra is equal to an integral power of two, and if $\mathcal{A}_n$ is a Boolean algebra having 2^n elements ($n=0, 1, 2, \ldots$) then $\mathcal{A}_n$ is isomorphic with the set of all subsets of a set S_n containing exactly n elements. In the present paper we consider the following problem to which the author was led by some problems in information theory: let us choose at random k elements of the Boolean algebra $\mathcal{A}_n$, and let $\mathcal{A}'$ denote the least Boolean subalgebra of $\mathcal{A}_n$ which contains these elements; calculate the probability that $\mathcal{A}'=\mathcal{A}_n$. By other words the question is: what is the probability that k elements of $\mathcal{A}_n$ chosen at random should generate $\mathcal{A}_n$?

We shall calculate this probability for every k, however we are interested in the first place in the question how large k should be in order that the k elements of $\mathcal{A}_n$ selected at random should generate the whole Boolean algebra $\mathcal{A}_n$ with a prescribed probability p where $0<p<1$.

To make the question determined, one has to define what should be understood by the random choice of the elements of $\mathcal{A}_n$. We shall solve our problem under two different definitions of random choice.

Definition 1. We suppose that at every choice every element of $\mathcal{A}_n$ has the same probability to be chosen, and that the subsequent choices are independent. This implies that if $A_1, A_2, \ldots, A_k$ is an arbitrary ordered sequence of elements of $\mathcal{A}_n$ (the same element of $\mathcal{A}_n$ may occur more than once in the sequence $A_1, A_2, \ldots, A_k$) then the probability that exactly these sets will be chosen (in the given order) is equal to $\frac{1}{2^{nk}}$.

Definition 2. We suppose that the first element A_1 is selected so that each element of $\mathcal{A}_n$ has the same probability to be chosen and that at

subsequent choices all those elements which have not yet been selected have the same probability to be chosen as the next. This implies that the randomly chosen elements $A_1, A_2, \ldots, A_k$ are all different and that for any ordered k-tuple $A_1, A_2, \ldots, A_k$, consisting of different elements of $\mathcal{A}_n$ the probability that exactly this k-tuple will be chosen (in the given order) is equal to $\frac{1}{2^n(2^n-1)\ldots(2^n-k+1)}$.

In § 1 and § 2 we solve our problem when Definition 1 or Definition 2 is adopted, respectively. In § 3 we generalize the question considered in § 1.

Before going into details I should like to say a few words about the connection of the problem considered in this paper with information theory. Let x be an unknown element of a set S_n which has n elements. We get information about x in the form that we are informed whether x belongs or not to the subsets $A_1, A_2, \ldots, A_k$ of S_n. Each such answer contains at most one unit of information; thus to determine x uniquely we need at least $\{\log_2 n\}$ such answers ($\{x\}$ denotes the least integer $\geqq x$) as the uncertainty concerning x is equal to $\log_2 n$. Now as well known, there can be in fact chosen $\{\log_2 n\}$ such subsets $A_1^*, A_2^*, \ldots, A_{\{\log_2 n\}}^*$ that every element x of S_n is uniquely determined by the information to which of the sets $A_1^*, A_1^*, \ldots, A_{\{\log_2 n\}}^*$ it belongs and to which not. For instance let the elements of S_n be labelled with the numbers $0, 1, \ldots, n-1$ and let A_j^* $(j = 1, 2, \ldots, \{\log_2 n\})$ denote the subset of those elements of S_n which are labelled by such a number m which when written in the binary system has its j-th digit equal to 1, i. e. is of the form

$$m = \sum_{h=1}^{\{\log_2 n\}} b_h \cdot 2^{h-1} \quad \text{(where } b_h = 1 \text{ or } b_h = 0\text{) with } b_j = 1.$$

Then clearly if it is known whether x belogs to A_j^* or not for $j = 1, 2, \ldots, \{\log_2 n\}$, then the binary expansion of x and thus x itself is uniquely determined. This can be expressed also in the following way: to any two elements x and $y \neq x$ of S_n there is at least one among the sets $A_1^*, A_2^*, \ldots, A_{\{\log_2 n\}}^*$ which separates the elements x and y, that is one of them is contained in this set and the other not.

We shall call such a system of sets which may be used to separate any two elements of a set a *separating set of subsets*. Evidently a separating set of subsets of S_n has at least $\{\log_2 n\}$ elements. We shall call a separating system of subsets of S_n consisting of exactly $\{\log_2 n\}$ sets an *optimal separating system*.

Now the question arises that if we do not choose the subsets $A_1, A_2, \ldots, A_k$

in such an optimal and systematic way, but choose them at random, how many subsets have to be chosen in order that the system of sets obtained should be a separating system, with a prescribed probability p. Clearly this is equivalent to demanding that the least algebra $\mathcal{A}'$ of subsets of S_n containing the sets $A_1, A_2, \ldots, A_k$ should be the set of all subsets of S_n. As a matter of fact the requirement that the sets $A_1, A_2, \ldots, A_k$ should form a separating system for the set S_n is equivalent with the assertion that the atoms of the least algebra $\mathcal{A}'$ containing $A_1, A_2, \ldots, A_k$ should consist each of only one element of S_n, and this is equivalent with saying that the subsets $A_1, A_2, \ldots, A_k$ generate the set of all subsets of S_n.

Of course this is possible only if $k \geqq \{\log_2 n\}$, thus at least $\{\log_2 n\}$ sets are needed for this purpose, and the question is exactly this: how much larger k should be than $\{\log_2 n\}$? Theorem 1 gives an answer to this question.

§ 1. Random choice of subsets according to Definition 1

In this § we suppose that the random choice of the elements $A_1, A_2, \ldots, A_k$ of the Boolean algebra $\mathcal{A}_n$ is subject to Definition 1, i. e. these elements are chosen independently of each other and each may be equal with the same probability $\left(\text{i. e. with probability } \frac{1}{2^n}\right)$ to each of the 2^n elements of $\mathcal{A}_n$. Let $P(\ldots)$ denote the probability of the event in the brackets.

We prove the following

Theorem 1. *Let $E_{n,k}$ denote the event that the random elements $A_1, A_2, \ldots, A_k$ of the finite Boolean algebra $\mathcal{A}_n$ (having 2^n elements) generate the whole algebra $\mathcal{A}_n$, supposing that these elements are chosen independently and each element of $\mathcal{A}_n$ has the same probability to be selected at every choice. Then we have*

$$P(E_{n,k}) = \prod_{j=1}^{n-1}\left(1-\frac{j}{2^k}\right). \tag{1}$$

Proof. Without restricting the generality we may suppose that the Boolean algebra $\mathcal{A}_n$ in question is the set of all subsets of a set S_n having n elements, which we denote by $a_1, a_2, \ldots, a_n$. Every subset A of S_n can be characterized completely by the sequence of numbers $\varepsilon_h(A)$ $(h = 1, 2, \ldots, n)$ where $\varepsilon_h(A) = 1$ or $= 0$ according to whether A contains a_h or not. If A is selected at random so that each of the 2^n subsets of S_n has the same probability to be chosen, then the $\varepsilon_h(A)$ $(h = 1, 2, \ldots, n)$ are independent random variables each taking on the values 1 and 0 with probability $\frac{1}{2}$. As a

matter of fact if $\delta_1, \delta_2, \ldots, \delta_n$ is an arbitrary sequence of zeros and ones there is exactly one subset A of S_n for which $\varepsilon_h(A) = \delta_h$ for $h = 1, 2, \ldots, n$ and thus

$$P(\varepsilon_1(A) = \delta_1, \varepsilon_2(A) = \delta_2, \ldots, \varepsilon_n(A) = \delta_n) = \frac{1}{2^n} \tag{2}$$

for each such sequence $\delta_1, \delta_2, \ldots, \delta_n$.

Let us consider now the random variables $\varepsilon_h(A_j)$ $(h = 1, 2, \ldots, n; j = 1, 2, \ldots, k)$. According to what has been said and to Definition 1 the random variables $\varepsilon_h(A_j)$ $(h = 1, 2, \ldots, n; j = 1, 2, \ldots, k)$ are all independent and each takes on the values 1 and 0 with probability $\frac{1}{2}$. Consequently the random k-dimensional vectors $\mathfrak{V}_h = (\varepsilon_h(A_1), \varepsilon_h(A_2), \ldots, \varepsilon_h(A_k))$ $(h = 1, 2, \ldots, n)$ are also independent, and each takes on any of its possible 2^k values with probability $\frac{1}{2^k}$. Now clearly the event $E_{n,k}$ is equivalent with the statement that the vectors $\mathfrak{V}_1, \mathfrak{V}_2, \ldots, \mathfrak{V}_n$ are all different. Thus we have

$$P(E_{n,k}) = \frac{\prod_{j=0}^{n-1}(2^k - j)}{2^{kn}} = \prod_{j=1}^{n-1}\left(1 - \frac{j}{2^k}\right),$$

which proves Theorem 1.

We deduce from Theorem 1 by some easy calculations the following

Corollary. *If n_j and k_j are two sequences of positive integers such that the limit*

$$\lim_{j \to +\infty} (k_j - 2 \log_2 n_j) = c \tag{3}$$

exists, then

$$\lim_{j \to +\infty} P(E_{n_j, k_j}) = \begin{cases} 1 & \text{if } c = +\infty, \\ e^{-1/2^{c+1}} & \text{if } c \text{ is finite}, \\ 0 & \text{if } c = -\infty. \end{cases} \tag{4}$$

Theorem 1 shows that if we choose the subsets $A_1, A_2, \ldots, A_k$ of the set S_n at random then the number of such sets which are required, in order that these sets should with considerable probability generate the full Boolean algebra of subsets of S_n, is roughly twice as large as the minimal number of systematically selected subsets which have this property. Especially if $k \sim 2 \log_2 n$ then choosing k subsets of S_n at random these will generate the full Boolean algebra with a probability tending to $e^{-1/2} = 0{,}6065\ldots$.

An other formulation of Theorem 1 is as follows: let us choose at random (according to Definition 1) elements of $\mathfrak{A}_n$ and denote them by

$A_1, A_2, \ldots$. Let $\mathcal{A}^{(k)}$ be the least subalgebra of $\mathcal{A}_n$ containing $A_1, A_2, \ldots, A_k$. Let ν_n denote the least integer for which $\mathcal{A}^{(\nu_n)} = \mathcal{A}_n$. Then ν_n is a random variable which has the probability distribution

$$P(\nu_n \leqq k) = \prod_{j=1}^{n-1}\left(1-\frac{j}{2^k}\right). \tag{5}$$

It follows that

$$\lim_{n\to+\infty} P(\nu_n - 2\log_2 n \leqq c) = e^{-\frac{1}{2^{c+1}}}. \tag{6}$$

Note that it follows from (1) that $P(E_{nk})$ vanishes (as it should) for $2^k < n$ while in the case $2^k = n$ we have

$$P(E_{2^k,k}) = \frac{2^k!}{2^{k\cdot 2^k}}. \tag{7}$$

Thus the probability to find by random choice an optimal separating system of subsets is fairly small. The numerator $2^k!$ of the fraction on the right of (7) is clearly nothing else than the number of different optimal separating systems for S_{2^k}. That this number is $2^k!$ can be proved also directly as follows: each set in an optimal separating system for the set S_{2^k} has to consist clearly of 2^{k-1} elements, the second set has to dissect the first set as well as its complementary set into two subsets of 2^{k-2} elements each, the third set has to dissect all of these four subsets into two subsets of 2^{k-3} elements each, etc. Thus we obtain for the number $O(k)$ of optimal separating systems

$$O(k) = \binom{2^k}{2^{k-1}}\cdot\binom{2^{k-1}}{2^{k-2}}^2\cdot\binom{2^{k-2}}{2^{k-3}}^4\cdots\binom{2}{1}^{2^{k-1}} = 2^k! \tag{8}$$

which is equivalent with (7).

Of course in the number $2^k!$ each optimal separating system is counted in every possible order of its elements. Thus the number $O^*(k)$ of essentially different optimal separating system for S_{2^k} is

$$O^*(k) = \frac{2^k!}{k!}. \tag{9}$$

The result of the Corollary of Theorem 1 is rather surprising as one would have expected that with random selection a much larger number of sets is necessary in the average to generate the whole algebra. Let us remember that in principle among the sets $A_1, A_2, \ldots, A_k$ the same set may occur more than once (though the probability of this is rather small). In the next § we shall show that the result remains valid if this is excluded, i. e. if we adopt Definition 2 for the random selection of sets.

§ 2. Random choice of subsets according to Definition 2

Let us denote by $E^*_{n,k}$ the event that if the selection of sets is made according to Definition 2, the selected sets $A_1, A_2, \ldots, A_k$ generate the Boolean algebra of all subsets of S_n. Clearly we have

$$P(E^*_{n,k}) = \frac{M(n,k)}{2^{nk}} \tag{10}$$

where $M(n,k)$ denotes the number of such matrices having k rows and n columns each element of which is equal to 0 or 1, which have the property that all its row vectors are different and all its column vectors are different.

The exact formula for $M(n,k)$ is rather complicated. We shall consider here only the asymptotic behaviour of $P(E^*_{n,k})$ and prove that the Corollary of Theorem 1 holds for $P(E^*_{n_j,k_j})$ instead of $P(E_{n_j,k_j})$ too. This can be shown as follows: we have clearly

$$P(E^*_{n,k}) = \frac{P(E_{n,k}B_{n,k})}{P(B_{n,k})} \tag{11}$$

where the product of two sets denotes their intersection and $B_{n,k}$ denotes the event that if the subsets $A_1, A_2, \ldots, A_k$ of S_n are chosen at random according to Definition 1 they turn out to be all different. If $\bar{B}_{n,k}$ denotes the event contrary to $B_{n,k}$, it follows that

$$\frac{P(E_{n,k}) - P(\bar{B}_{n,k})}{1 - P(\bar{B}_{n,k})} \leqq P(E^*_{n,k}) \leqq \frac{P(E_{n,k})}{1 - P(\bar{B}_{n,k})}. \tag{12}$$

Now clearly

$$P(\bar{B}_{n,k}) \leqq \frac{\binom{k}{2}}{2^n} = o(1) \quad \text{if} \quad k = o(2^{\frac{n}{2}}). \tag{13}$$

It follows from (12), (13) and (4) that if

$$\lim_{j \to +\infty} (k_j - 2\log_2 n_j) = c \tag{14}$$

then

$$\lim_{j \to +\infty} P(E^*_{n_j,k_j}) = \begin{cases} 1 & \text{if} \quad c = +\infty, \\ e^{-1/2^{c+1}} & \text{if} \quad c \quad \text{is finite}, \\ 0 & \text{if} \quad c = -\infty. \end{cases} \tag{15}$$

Thus the same asymptotic results hold for $P(E^*_{nk})$ as for $P(E_{nk})$.

§ 3. Generalizations

We may ask the following question. If we choose n and k so that the probability of E_{nk} should be essentially less than 1, what can be said about the structure of the least subalgebra $\mathcal{A}'$ of $\mathcal{A}_n$ which contains the random sets $A_1, A_2, \ldots, A_k$. Let $B_1, B_2, \ldots, B_r$ be the atoms of $\mathcal{A}'$, then $B_1, B_2, \ldots, B_r$ are disjoint subsets of S_n whose union is equal to S_n. If there are r_s atoms B_i which consist of exactly s elements of S_n then

$$\sum_{s=1}^{n} s r_s = n \quad \text{and} \quad \sum_{s=1}^{n} r_s = r \leqq 2^k \tag{16}$$

and the sequence $(r_1, r_2, \ldots, r_n)$ may be called the *signature* of $\mathcal{A}'$. Clearly $\mathcal{A}' = \mathcal{A}_n$ if and only if the signature of $\mathcal{A}'$ is $(n, 0, 0, \ldots, 0)$.

Now we may ask what is the probability that $\mathcal{A}'$ should have a prescribed signature $(r_1, r_2, \ldots, r_n)$ (we suppose (16) to be satisfied). According to what has been said in § 1 this problem is equivalent with the following one: put n balls into 2^k urns independently from each other; what is the probability that there will be among the 2^k urns exactly r_s urns which contain exactly s balls $(s = 1, 2, \ldots, n)$? The answer to this question may be easily given by elementary combinatorial considerations: the probability in question is

$$\frac{2^k! \cdot n!}{2^{kn} \cdot \left(\prod_{s=1}^{n} r_s! \, (s!)^{r_s} \right) \left(2^k - \sum_{s=1}^{n} r_s \right)!} \tag{17}$$

Especially one gets from (17) if $r_1 = n$, $r_s = 0$ for $s > 1$ Theorem 1 as a special case.

Other results of a similar type on systems of random subsets of a finite set will be given elsewhere ([1], [2]). The theory of systems of random subsets of a finite set can be developed along similar lines as the theory of random graphs as worked out by P. ERDŐS and the author of the present paper (see [3]).

References

[1] RÉNYI A., Egy általános módszer valószínűségszámítási tételek bizonyítására és annak néhány alkalmazása, *Magyar Tud. Akad. Mat. Fiz. Oszt. Közl.* (in print).

[2] A. RÉNYI, On random subsets of a finite set, *Mathematica* (Cluj) (in print).

[3] P. ERDŐS and A. RÉNYI, On the evolution of random graphs, *Magyar Tud. Akad. Mat. Kutató Int. Közl.*, **5** (1960), 17—61.

(Received January 18, 1961)