Probability methods in number theory.*)

By ALFRÉD RÉNYI.

§ 1. During the last years the calculus of probability has developed into an exact mathematical theory, founded on the theory of additive set functions in abstract spaces and satisfying the same standards of rigour as any other chapter of Analysis. The most simple and at the same time the most powerfull axiomatic theory of probability has been given in 1933 by A. KOLMOGOROFF[1]) in his fundamental work "Grundbegriffe der Wahrscheinlichkeitsrechnung". Since that time the theory has been developed further by many authors, among whom we mention (without aiming at completeness) H. CRAMÉR, W. FELLER, A. KHINTCHIN, B. GNEDENKO, P. LÉVY, M. FRÉCHET, E. SLUTSKY, M. KAC, H. STEINHAUS and J. L. DOOB, for references we refer to a lecture of H. CRAMÉR[2]) held at the Princeton Bicentennial Conference on "The Problems of Mathematics", in 1946. As every abstract axiomatic theory, the theory of probability as given by KOLMOGOROFF admits of infinitely many interpretations, i. e. realizations by different models. This is the reason why the theory can be applied not only in the usual fields of application of probability theory (including the most recent physical applications) but also in some chapters of mathematics, which have, at the first sight, nothing to do with the everyday concept of probability. We mention — for instance — the application of probability methods in the theory of orthogonal functions[3]), in the ergodic theory[4]), or the proof of the well known theorem of Weierstrass concerning the uniform approximation of continuous functions as given by S. BERNSTEIN[5]), deducing this theorem by using the elementary inequality of CHEBYSEV. Thus it is by no means surprising that the methods and results of probability theory can be also applied in number theory. Such applications have been given by E. TORNIER[6]), by P. ERDŐS and M. KAC[7]) and others. Other authors have used probability arguments in number theory

*) Inaugural lecture, for attaining venia legendi at the University of Budapest, held January 20, 1949.

only as heuristic methods, and obtained frequently results which were definitely false. As an example how probability arguments when they are applied freely may lead to incorrect results, we mention the following example: According to the prime number theorem we may say that the probability that an integer chosen at random between N and N^2 should be prime, is asymtotically equal to $\frac{1}{2\log N}$. On the other hand an integer between N and N^2 is prime if it is not divisible by any prime $< N$. The probability that a number shall not be divisible by the prime p is evidently $1-\frac{1}{p}$ and as the events that an integer n should be divisible by two different primes can be considered as independent, we obtain that the required probability is equal to $\prod\limits_{p<N}\left(1-\frac{1}{p}\right)$. Thus we obtained

$$\prod_{p<N}\left(1-\frac{1}{p}\right)\sim\frac{1}{2\log N} \tag{1. 1}$$

which is evidently false as by a well known theorem of MERTENS[8]) we have

$$\prod_{p<N}\left(1-\frac{1}{p}\right)\sim\frac{e^{-c}}{\log N} \tag{1. 2}$$

where c denotes EULER's constant.

§ 2. The mistake in the above sketched argument is contained in the assumption that the distributions of integers in residue classes with respect to two different primes are independent. This assumption is contained in most examples where a false result in number theory is deduced by probabilistic arguments. Nevertheless this assumption is in some sense true, but can not be applied as in has been made above. The situation can be made clear as follows: Let us choose for the set E of elementary events the set of all positive integers. The field F of random events shall be chosen as the set of finite or infinite sequences of integers $A=\{n_1, n_2, \ldots, n_k, \ldots\}$ for which putting

$$A(x)=\sum_{n_k\leq x}1 \tag{2. 1}$$

the limit

$$\lim_{x\to\infty}\frac{A(x)}{x}=P(A) \tag{2. 2}$$

exists. The field F together with the additive function $P(A)$ defined by (2. 2) satisfies the first five axioms of KOLMOGOROFF (l. c. [1])) but the difficulty lies in the fact, that the sixth axiom, the so-called axiom of continuity, is not satisfied. As a matter of fact, this axiom states that if A_n, $n=1, 2, 3, \ldots$ is a sequence of sets belonging to the field F, each A_n being contained in the preceding set A_{n-1}, and if the product of all sets A_n is empty, then

$$\lim_{n\to\infty} P(A_n)=0.$$

If we choose for A_n the set of all integers $\geq n$ (which clearly belongs to F) we see that this axiom is not satisfied in our case. This difficulty can be settled if we choose for the set E the finite set of integers $1, 2, \ldots, N$. In this case F can be chosen as the family of all subsets $A=\{n_1, n_2, \ldots, n_k\}$ of $1, 2, \ldots, N$ and put

$$P(A)=\frac{k}{N}. \tag{2. 4}$$

But in this case the distributions of the sequence $1, 2, \ldots, N$ in residue classes with respect to two different primes (or with respect to two relatively prime moduli) are not longer independent except in the special case if both moduli are divisors of N. The mistake in the example mentioned above consits in confusing these two points of view. The correct method consists in taking the second point of view described above (i. e. to confine ourselves to a finite set of integers) and take into account the fact that in this case the distributions of integers in residue classes with respect to two relatively prime integers are only "almost" independent in some sense to be precised.

§ 3. In what follows we shall prove a new theorem of probability theory, which has important applications in number theory. One of its applications is a generalization of the "large sieve" of JU. V. LINNIK[9]), which served as a starting point in my proof of the theorem that every integer $N\geq 3$ can be represented in the form $N=p+P$ where p is a prime and the number of prime divisors (counted with their multiplicity) of P does not exceed a universal constant K.[10]) In what follows we shall confine ourselves to proving only the mentioned theorem of probability theory[11]), the applications shall be published elsewhere[12]).

§ 4. Let E denote an arbitrary set which shall be called in what follows the space E, the elements of E shall be called points and denoted by the letter ξ. Let F denote an additive class of subsets A of E. We mean by this that: *a)* the empty set belongs to F, *b)* when a set belongs to F so does its complement, *c)* the sum of a finite or enumerable sequence of sets belonging to F is also contained in F. It follows from a, b, c clearly that the space E itself belongs to F further that if A and B belong to F so does their product, and their difference. Let us suppose further that $P(A)$ is an absolutely additive set function defined for every set A belongig to F, $P(A)\geq 0$ and $P(E)=1$. The class F of sets and the additive set function $P(A)$ form together a probability field in the sense of KOLMOGOROFF (l. c.). Let $u = u(\xi)$ denote a random variable defined on E, i. e. a real-valued point-function in E for which the set A_x of those points ξ for which $u(\xi) < x$ belongs for every real x to F. Let $V(x)$ denote the distribution (d. f.) of u, defined by $V(x) = P(A_x)$. Let G denote any set belonging to F; $P(G) \neq 0$. We define the d. f. $V^G(x)$ of u with respect to G by putting

$$V^G(x) = \frac{P(A_x . G)}{P(G)} . \tag{4. 1}$$

Let $U(x)$ and $V(x)$ denote two distribution functions and let us suppose that $U(x)$ is constant in every interval in which $V(x)$ is constant. We define the distance of $U(x)$ from $V(x)$ — denoted by (U, V) and supposed to be non-negative — by putting

$$(U, V)^2 = \int_{-\infty}^{+\infty} \frac{d^2 (U - V)}{d V} \tag{4. 2}$$

where the integral figuring in (4. 2) is to be taken in the sense of HELLINGER[13]). The distance (U, V) can be expressed also by a Burkill integral. Let I denote an interval (a, b) let us put $U[I] = U(b) - U(a)$, $V[I] = V(b) - V(a)$ and $F[I] = \frac{(U[I] - V[I])^2}{V[I]}$, it follows that

$$(U, V)^2 = \int_{-\infty}^{+\infty} F[I] \tag{4. 3}$$

where the integral of the interval function $F[I]$ is to be taken in the sense of BURKILL[14]). As further $F[I]$ is a subadditive function of interval, i. e. $F[I_1 + I_2] \leq F[I_1] + F[I_2]$ (which follows by the elementary inequality $\frac{(a+b)^2}{c+d} \leq \frac{a^2}{c} + \frac{b^2}{d}$ for a, b, c, d real, $c > 0$, $d > 0$ from the fact that $U[I]$ and $V[I]$ are additive functions of interval) we can define $(U, V)^2$ also as the least upper bound of sums

$$\sum_{k=1}^{n} F[I_k] \tag{4. 4}$$

for every decomposition $R = I_1 + I_2 + \ldots + I_n$ of the real axis R. As we supposed $U[I] = 0$ if $V[I] = 0$, i. e. the interval function $U[I]$ is absolutely continuous with respect to $V[I]$, according to the theorem of RADON—NIKODYM[15]) there exists a function $f(x)$ for which

$$U[I] = \int_I f(x)\, d V(x). \tag{4. 5}$$

By using this function we obtain a third formula for $(U, V)^2$

$$(4.6) \qquad (U, V)^2 = \int_{-\infty}^{+\infty} (f(x)-1)^2 \, d V(x)$$

where the integral is a LEBESGUE—STIELTJES integral.*)

Now let $V(x)$ denote the d. f. of the random variable u and let us choose for $U(x)$ the d. f. $V^G(x)$ of u with respect to the set G (or we may say with respect to the random "event" G). We denote

$$(4.7) \qquad D_G(u) = (V^G, V) . \sqrt{\frac{P(G)}{1-P(G)}}$$

and shall call $D_G(u)$ *the discrepance of u on the set G*, (or with respect to the event G). To make clear the meaning of $D_G(u)$ we mention some special cases: let us suppose that the random variable u takes only a finite number of different values: let us put $E = H_1 + H_2 + \ldots + H_n$ with $H_i . H_j = 0$ for $i \neq j$, and let us suppose that $u(\xi) = u_k$ for $\xi \in H_k$, $k = 1, 2, \ldots, n$. In this case it is easy to see that

$$(4.8) \qquad D_G(u) = \left(\sum_{k=1}^{n} \frac{(P(G.H_k) - P(G)P(H_k))^2}{P(H_k)P(G)(1-P(G))} \right)^{\frac{1}{2}}$$

i. e. our $D_G(u)$ is a generalization of the deviation-measure χ^2 of K. PEARSON[16]). If in the above example $n=2$, i. e. $H_1 = E - H_2 = H$, we have

$$D_G(u) = \frac{|P(GH) - P(G)P(H)|}{\sqrt{P(G)(1-P(G))P(H)(1-P(H))}}$$

i. e. $D_G(u)$ is the square of the correlation coefficient of the two events: $\xi \in H$ and $\xi \in G$.

After these preliminary remarks we are in the position to formulate our main

Theorem. *Let us consider a sequence $u_1, u_2, \ldots, u_n, \ldots$ of pairwise independent random variables defined on E. Let G denote a subset of E belonging to the class F and let us suppose* $0 < P(G) < 1$. *If $D_G(u_n)$ denotes the discrepance of u_n on the set G, we have*

$$(5.1) \qquad \sum_{n=1}^{\infty} D_G^2(u_n) \leq \frac{1}{1-P(G)} .$$

It may be mentioned, that this theorem is in some sense a best possible result. As a matter of fact, let us choose for u_1 the characteristic function of the set G. In this case as a consequence of the independence of u_1 and u_n we have $D_G(u_n) = 0$ for $n = 2, 3, \ldots$ further evidently by (4.9) $D_G(u_1) = 1$. As $P(G)$ can be made as small as we please, it

*) As regards the expression of a HELLINGER integral by means of a LEBESGUE-integral vide H. HAHN, *Monatshefte f. Math. u. Phys.* **23** (1912) pp. 161—224.

can be seen that (5. 1) can not be true for any $\lambda < 1$ with $\frac{\lambda}{1-P(G)}$ instead of $\frac{1}{1-P(G)}$ on the right of (5. 1).

The proof of (5. 1) shall be based on the following

Lemma: Let $\varphi_1(\xi), \varphi_2(\xi), \ldots, \varphi_n(\xi), \ldots$ denote a *sequence of quasi-orthonormal functions,* belonging to the class L^2 on E i. e. we suppose

$$\left|\sum_{i=1}^{\infty}\sum_{k=1}^{\infty} c_{ik}x_i x_k\right| \leq C\sum_{k=1}^{\infty} x_k^2 \tag{5. 2}$$

with an absolute constant C for any sequence x_n of real numbers for which the right-hand side of (5. 2) is finite, where c_{ik} in (5. 2) is defined by

$$c_{ik} = \int_E \varphi_i(\xi)\,\varphi_k(\xi)\,P(dE) \quad {}^{*)} \tag{5. 3}$$

If $g(\xi)$ is any function belonging to L^2 on E, let us put

$$(g, \varphi_k) = \int_E g(\xi)\,\varphi_k(\xi)\,P(dE). \tag{5. 4}$$

It follows

$$\sum_{k=1}^{\infty}(g, \varphi_k)^2 \leq c\int_E g^2(\xi)\,P(dE). \tag{5. 5}$$

The content of our Lemma can be characterized by saying that the inequality of Bessel can be generalized for quasi-orthonormal sequences of functions. This has been proved first by R. P. BOAS, Jr.[17]) the proof follows in two lines, from the obvious inequality

$$\int_E \left[f(\xi) - \sum_{k=1}^{N}\varphi_k(\xi)\cdot(f, \varphi_k)\right]^2 P(dE) \geq 0 \tag{5. 6}$$

by effecting the multiplication, integrating term by term and applying (5. 2) with $x_k = (f, \varphi_k)$ for $k \leq N$, and after that effecting $N \to \infty$.

§ 6. Let us turn now to the proof of (5. 1). According to the remark made in § 4. with respect to the evaluation of the Burkill integrall $\int F[I]$, it suffices to prove the inequality

$$\frac{P(G)}{1-P(G)}\sum_{n=1}^{\infty}\sum_{k=1}^{N_n}\frac{(V_n^G[I_{nk}] - V_n[I_{nk}])^2}{V_n[I_{nk}]} \leq \frac{1}{1-P(G)} \tag{6. 1}$$

where $V_n[I] = V_n(b) - V_n(a)$ if $I = (a, b)$ and $V_n(x)$ denotes the d. f. of u_n

*) Here and in what follows the integral $\int_E f(\xi)\,P(dE)$ is to be understood as the generalized LEBESGUE integral with respect to the measure $P(A)$ in the abstract space E. For particulars see S. SAKS [15]) l. c.

further $V_n^G[I]=V_n^G(b)-V_n^G(a)$ and $V_n^G(x)$ denotes the d. f. of u_n with respect to G for every sequence of subdivisions $I_{n1}+I_{n2}+\ldots+I_{nN_n}=R$ of the real axis R. Evidently we may confine ourselves to subdivisions $\{I_{nk}\}$ for which $V_n[I_{nk}]>0$ for every k and n. To every subdivision $\{I_{nk}\}$ of R there corresponds a subdivision $\{E_{nk}\}$ of the space E where E_{nk} is defined as the set of those points ξ for which $u_n(\xi)\in I_{nk}$. The sets E_{nk} belong evidently to F. Let us put $P(E_{nk})=V_n[I_{nk}]=p_{nk}>0$, let $F_{nk}(\xi)$ denote the characteristic function of the set E_{nk}, and let us define

$$\varphi_{nk}(\xi)=\frac{F_{nk}(\xi)-p_{nk}}{\sqrt{p_{nk}}} \quad \text{for } k=1,2,\ldots,N_n \text{ and } n=1,2,3,\ldots \tag{6. 2}$$

We have by a simple calculation:

$$\begin{aligned} &\int_E \varphi_{nk}^2(\xi)\,P(dE)=1-p_{nk} \\ &\int_E \varphi_{nk}(\xi)\,\varphi_{ml}(\xi)\,P(dE)=0 \quad \text{for} \quad n\neq m \\ &\int_E \varphi_{nk}(\xi)\,\varphi_{nk'}(\xi)\,P(dE)=-\sqrt{p_{nk}p_{nk'}} \quad \text{for} \quad k\neq k'. \end{aligned} \tag{6. 3}$$

The sequence of functions $\varphi_{nk}(\xi)$ $k=1,2,\ldots,N_n$; $n=1,2,3,\ldots$ is a quasi-orthonormal system as defined in the above Lemma, the value of the constant C being equal to unity.

As a matter of fact, let us put

$$C_{nkml}=\int_E \varphi_{nk}(\xi)\,\varphi_{ml}(\xi)\,P(dE) \tag{6. 4}$$

We have by (6. 3)

$$\sum_{n=1}^{\infty}\sum_{m=1}^{\infty}\sum_{k=1}^{N_n}\sum_{l=1}^{N_m} C_{nkml}x_{nk}x_{ml}=\sum_{n=1}^{\infty}\sum_{k=1}^{N_n}x_{nk}^2-\sum_{n=1}^{\infty}\left(\sum_{k=1}^{N_n}x_{nk}\sqrt{p_{nk}}\right)^2. \tag{6. 5}$$

As by the inequality of SCHWARZ we obtain

$$\left(\sum_{k=1}^{N_n}x_{nk}\sqrt{p_{nk}}\right)^2\leq\left(\sum_{k=1}^{N_n}p_{nk}\right)\left(\sum_{k=1}^{N_n}x_{nk}^2\right)=\sum_{k=1}^{N_n}x_{nk}^2 \tag{6. 6}$$

and thus it follows, that the left-haud side of (6. 5) does not exceed

$$\sum_{n=1}^{\infty}\sum_{k=1}^{N_n}x_{nk}^2 \tag{6. 7}$$

in absolute value, and therefore we have $C=1$.

Now we can apply our Lemma, with $g(t)$ being the characteristic function of the set G. As in this case

$$(g,\varphi_{nk})=P(G)\,.\,\frac{V_n^G[I_{nk}]-V_n[I_{nk}]}{\sqrt{V_n[I_{nk}]}} \tag{6. 8}$$

further

$$\int_E g^2(\xi)\, P(dE) = P(G) \tag{6. 9}$$

we obtain from (5. 5) the inequality (6. 1), which, according to what has been said above, proves our theorem.

§ 7. Our theorem can be generalized for the case in which the random variables u_n are not independent, only "almost-independent" in the following sense: If the sequence of random variables $u_1, u_2, \ldots, u_n$ satisfies the condition

$$\left| \frac{P(A_{ab}^{(n)} . A_{cd}^{(m)})}{P(A_{ab}^{(n)})\, P(A_{cd}^{(m)})} - 1 \right| \leq \delta_n \delta_m \tag{7. 1}$$

for any two pairs of real numbers (a, b) and (c, d), where $A_{a,b}^{(n)}$ resp. $A_{c,d}^{(m)}$ denotes the set of points ξ of E for which $a < u_n(\xi) < b$ resp. $c < u_m(\xi) < d$ holds, and the positive constants δ_n decrease rapidly so that the series

$$\sum_{n=1}^{\infty} \delta_n^2 = \delta \tag{7. 2}$$

converges, we shall call the sequence $u_1, u_2, \ldots, u_n, \ldots$ a pairwise almost independent sequence of random variables, and we shall call δ the coefficient of dependence of the sequence. Using this definition we obtain by the same method of proof the following generalization of our theorem:

If the sequence of random variables $u_1, u_2, \ldots, u_n, \ldots$ is pairwise almost independent in the sense precised above, having its coefficient of dependence $\delta < 1$, and if G is any set belonging to F with $0 < P(G) < 1$, we have

$$\sum_{n=1}^{\infty} D_G^2(u_n) \leq \frac{1}{(1-\delta)(1-P(G))} . \tag{7. 3}$$

As the idea of the proof is the same, it may be left to the reader. In this form our theorem can be immediately applied to the proof of the large sieve of Linnik.

I wish to express my sincere thanks to Prof. F. Riesz for his valuable remarks.

Budapest, May 26, 1949.

References:

[1] Ergebnisse der Mathematik, II. 3, Berlin, 1933, pp. 1—62.

[2] Annals of Math. Statistics XVIII, 2, 1947, pp. 165—193.

[3] see for instance St. Kaczmarz and H. Steinhaus, Theorie der Orthogonalreihen. *Monografje Matematyczne* VI, Warszawa—Lwow 1935, p. 126, 453.

[4] see E. Hopf, Ergodentheorie, Ergebnisse der Mathematik V. 2, Berlin 1937.

[5] Communic. Soc. Math. Charkow, Ser. 2, 13, 1912, p. 1–2.

[6] Journal f. r. u. ang. Mathematik, 160, 1929, pp. 177–198.

[7] Proc. Nat. Acad. Sciences, 25, 4, 1939, pp. 206–207.

[8] see p. e. G. H. Hardy and E. M. Wright, An introduction to the theory of numbers, Oxford 1945 2nd ed. p. 349.

[9] Comptes Rendus de l'académie des Sciences de l'URSS, 30. 4, 1941, pp. 290–292.

[10] A. Rényi, Bulletin de l'Academie des Sciences de l'URSS, Ser. Math. Vol. 12, No. 1, 1948, pp. 57–78.

[11] A special case of this theorem shall be published in a paper of the author to be appear in the Journal de Mathématique.

[12] To be appear in the Compositio Mathematica.

[13] E. Hellinger, Journal f. d. r. u. ang. Mathematik, 136, 1909, pp. 210–271.

[14] see p. e. St. Kempisty, Fonctions d'intervalle non additives, Act. Sci. et Ind. Paris, 1937.

[15] See S. Saks, Theory of the Integral, *Monografie Matematyczne*, VII, Warszava—Lwow, 1937, p. 36.

[16] K. Pearson, Phil. Mag. VI, 50, 1901, p. 157.

[17] Amer. Journal Math. 63, 1941, pp. 361–370.