PROBABILITY AND PRIME NUMBERS

BY S. DUCRAY

Received July 23, 1964

(Communicated by Sir C. V. Raman)

THIS note sets up a sample-space connected with the infinite succession of prime integers. The properties of this sample-space cast fresh light upon some fundamental problems of analytic number theory.

Definitions.—Let an arbitrary denumerable set of positive real numbers (not necessarily integers) be given: $0 < a_1 < a_2 \cdots$ with $a_n \to \infty$ as $n \to \infty$. For a fixed length $u > 0$, a *covering* of the real half-line $y > 0$ is given by the sequence of intervals $I_1, I_2, \cdots$ where I_n: $(n-1)\ u \leq y < nu$. By $s_n = s(u, n)$ is meant the number of points with co-ordinates $y = a_i$ contained in I_n. Thus $\{s_n\}$ provides a sample-sequence for the particular covering which begins from $y = 0$. Other sequences similarly obtained by beginning the covering from some other point of $y > 0$ or (what is the same thing) by subtracting the corresponding value from each a_i. Displacement of the initial point through an integral multiple ku of u gives the same sequence begun from the $(k+1)st$ term. We shall say that two sequences are essentially different if they do not coincide after a finite number of terms of one are omitted.

LEMMA 1.—*If the sequence* $\{a_i\}$ *has the properties*: (*a*) *that there are infinitely many gaps* $a_{r+1} - a_r > 2u$ *and* (*b*) *the* a_i *are Gleichverteilt modulo* u, *then the number of different sample-sequences obtained by displacement of the initial point through distances not exceeding* u *can be mapped in a* $1-1$ *manner upon the right-open unit interval* $0 \leq y < 1$.

Proof.—Suppose that, throughout some displacement $w < u$, the same sequence $\{s_n\}$ is obtained. Then the number of points $y = a_i$ lost to the left by a displaced interval must be precisely equal to that gained from the right, hence the same for all intervals. But the gaps ensure that no matter where the initial point be taken, there are always intervals with zero gain and loss. Hence the number gained or lost must always be zero throughout the displacement $w > 0$. This means a regular gap of length w in the numbers a_i as reduced modulo u, which contradicts hypothesis *b*. Therefore, $w = 0$

and there is a different sample-sequence for every point of $(0, u]$, which interval may then be projected upon $(0, 1]$.

LEMMA 2.—*If, in lemma 1, condition* b *be replaced by requiring only that the set of cluster-points of* $\{a_i\}$ *modulo* u *be of positive measure, it still follows that a sub-set of the distinct sample-sequences* $\{s_i\}$ *obtained by displacement of the initial point through not more than one* u*-interval may be mapped in a* $1 - 1$ *manner upon* $(0, 1]$.

Proof.—Now, it may be possible to obtain the same sequence for some positive displacement w, as gaps among the cluster points are permitted. Let w_1 be the limit superior of such displacements, beginning from $y = 0$. If, thereafter, the cluster-points modulo u are dense throughout some sub-interval $(w, w_1 + h)$, there will be a different sample-sequence for every point of this sub-interval, and the theorem is proved. If not, the cluster-points near w can be covered by an interval of arbitrary small length $\epsilon > 0$, and we proceed to the next cluster-point outside this small sub-interval, say w_2. Again, cover this with an interval $\epsilon/2$, then $\epsilon/4$ and so on, if at no stage is a finite interval of density attained for the cluster points. But then the set of cluster-points modulo u must be of measure zero, as their total measure cannot exceed 2ϵ, arbitrarily small. This proves the lemma by contradicting the hypothesis.

Definition.—In what follows, we take $1, 2, \ldots n, \ldots$ as the positive integers marked off at unit intervals on the half-line $x > 0$. Let $Li\,(x)$ be the integral $\int dt/\log t$ to the upper limit x.. Take $y = Li\,(x) - Li\,(x_0)$ for any $x_0 \geq 2$. Our set $\{a_i\}$ is the image-set on this y-line of the prime numbers $x = 2, 3, 5, \ldots p$.. Every covering is always to begin with $y = 0$, but the initial point may be varied by displacement of x_0. Thus, $s_n = s\,(x_0, u; n)$ is the number of primes included in the x-image of I_n: $(n - 1)\,u \leq y < nu$.

THEOREM 1.—*There exists at least one* $u > 0$ *such that a subset of the different sample-sequences (of prime images)* $\{s_n\}$ *obtained by displacement of the initial point continuously through the image of a single covering interval may be mapped in a* $1 - 1$ *manner upon* $(0, 1]$.

Proof.—Condition *a* of Lemma *1* is satisfied by the Erdös[1] gap-theorem. On the y-line, there are infinitely many gaps greater than $f\,(y)$ between the images of consecutive primes, where f tends rather slowly to infinity with y; hence the gaps are greater than any Au, for arbitrary constants A and u. Condition *b* of lemma *1* is apparently satisfied by a whole range of u-values, according to a result of G. Ricci.[2] However, P. Erdös[3] has pointed out that

the result actually proven shows only the existence of a positive measure for the set of cluster-points on the y-line (modulo any $u > 0$) for the images of the primes. This in any case satisfies the requirements of Lemma 2, which suffices.

Hereafter, take u to be one of the particular values under Theorem 1. Then make a *canonical mapping* onto (0, 1] of (almost all) covering sequences with initial points in some one fixed u-interval. Every sequence obtained by displacement of the initial point through any integral number of coverting intervals in either direction is to be mapped on the same point of (0, 1]. Every sequence thus mapped upon a given point of (0, 1] has then the same limiting frequency properties. That is, if the number of intervals of the sequence covering *0, 1, 2,..k..* images of primes reaches some limiting proportion f_k for one sequence associated with a point, it does so for every sequence mapped upon that point. *Random choice* of a sequence is defined as follows: First, all points of (0,1] have an equal chance of being chosen (uniform distribution on the map). Then, for a given point of the map, the actual sequence may be begun from any term whatever, counting that term as s_1, the next as s_2, and so on. This eliminates x_0 altogether from consideration, and we may speak only of properties of the sequences associated with points of the canonical map, using Labesgue measure on the map for probability.

In this situation, probability concepts apply to the $\{s_n\}$.

THEOREM 2.—*For all sequences* $\{s_n\}$ *defined as above, the mean value (expectation) is given by* $E(s_n) = u$ *for all* n.

Proof.—This is an immediate and obvious consequence of the prime number theorem, and of the method of choice of the sequences, seeing that s_n can be the number of primes covered by any interval of length u anywhere on the half-line $y > 0$. Here, the prime-number theorem is taken for granted, in the form $\pi(x) \sim Li(x)$, where $\pi(x)$ is the number of primes $p \leq x$.

There arise two cases, according to whether the random variables s_i are independent in the sense of probability theory or not.

THEOREM 3.—*Should the consecutive* s_r *of the same sequence (chosen at random, as above) be independent in probability, then the following results are true with unit probability* (i.e., *for almost all points of the canonical map*).

3.1. *The probability for any* s_i *assuming the value* k *is given by* $P(s_i = k) = e^{-u}u^k/k$ *for* $k = 0, 1, 2,..$

3.2. If $S_n = s_1 + s_2 + \cdots s_n$, *then for any arbitrary* $\epsilon > 0$,

$$-(1+\epsilon)\sqrt{2\mathrm{Nu}\ \log\log \mathrm{Nu}} < S_N - \mathrm{Nu} < (1+\epsilon)\sqrt{2\mathrm{Nu}\ \log\log \mathrm{Nu}}$$

for all except a finite number of values of N.

3.3. *If, in* 3.2, ϵ *be replaced by* $-\epsilon$, *then each of the two inequalities is false infinitely often as* $N \to \infty$.

Proof.—The first of these, namely 3.1, is equivalent to a result published by Kosambi[4] showing the primes on the y-line to be in a Poisson distribution with parameter u. The result is almost obvious under the given conditions.

With the Poisson distribution and complete independence in probability, textbook[5] methods lead immediately to the other two results. Of these, 3.2 is the *upper* law of the iterated logarithm, and 3.3 the *lower* law of the iterated logarithm.

However, independence in probability is not easy to prove for consecutive terms of our sample-sequences. Nevertheless, the sieve of Eratosthenes in its most elementary form enables the most important and useful part of the theorem, namely 3.2, to be carried over. This is best done in two stages:

LEMMA 3.—*For large* N *and each* k, $1 \leq k \leq N$, *the probability of* k *being the first index for which* $|S_k - ku| \geq \sqrt{2\mathrm{Nu}\ \log\log \mathrm{Nu}}$ *cannot exceed the same probability as calculated under the assumption of independence of the* s_i *as in Theorem 3.*

Proof.—The main idea is that the sieve of Eratosthenes prevents very large deviations from expectation from accumulating, if it has any effect at all upon independence of $(s_n - u)$ in the sense of probability theory.

If the position on the x-line were known, the primes in the image of k consecutive u-intervals would be completely determined. As it is, all that can be said is that there exists an unknown background parameter x such that $Nu \sim x/\log x$ for large N. The primes about x on the x-line are the numbers not deleted by the sieve, *i.e.*, the numbers not multiples of any primes $p \leq \sqrt{x}$. It is known that a connected stretch of length h on the x-line can contain at most $ch/\log h$ primes, where c is an absolute constant. A length ku on the y-line has an image $\sim ku \log x$. Therefore the probability in Lemma 3 is zero for $k \leq C\sqrt{x}\,(\log\log x/\log x)^{3/2}$, using the asymptotic values for N and x.

The question of independence now appears in the following manner. Given that r primes have in fact occurred in a specific number of consecutive

intervals; are the chances of some number m of primes occurring in a pre-assigned number of following intervals increased or decreased thereby, or remain unaffected—with no other information available. The answer is worked out as follows:

For every composite number that occurs, each prime factor $<\sqrt{x}$ cannot act as deleting prime for the corresponding distance on either side. The prime-number theorem and the existence of an expectation say that the probability for an integer in the x-image of a single interval being a prime is $1/\log x$, in order. If an unusually large number of primes turn up in a given stretch, this means unusually few composite numbers, unusually few $p \leq \sqrt{x}$ inactivated as deleting primes, and so, if the probability for primality is affected at all, a slight decrease therein. In the opposite direction, unusually few primes may mean more than a fair share of deleting primes dropping out of action, hence possible enhancement of the probability for primality in adjacent stretches. Nothing more can be said, provided of course, that the stretch where the known number of primes have turned up is of x-length less than $\sqrt{x}$ in order. For greater x-lengths, all deleting primes will delete in the stretch. The most that can then be said is that unusually many of these multiply each other when the number of primes left in the stretch is well above expectation; and the opposite when the number of primes covered by the stretch is far below expectation. In neither case can the same phenomenon be expected to continue over the next stretch. So, *the effect of dependence, if any, upon* $S_n - nu$ *may be compensatory, but never cumulative.* This proves the lemma.

THEOREM 4.—*If* π (x) *be defined as the number of primes* $p \leq x$, *then*

$$\pi(x) - Li(x) = O(\sqrt{x \log \log x / \log x}).$$

Proof.—With independence in probability, the result 3.2 and the asymptotic values for N and x prove the result immediately. With dependence, the estimates of lemma 3 still remain valid. This is the key condition for the validity of the *upper* law of the iterated logarithm, which is based upon the first Borel-Cantelli[6] lemma and hence does not require independence (which is only a sufficient condition). Thus, regardless of the validity of 3.1 and 3.3, the result 3.2 still remains true, and the inequalities may at most be strengthened, never weakened.[7]

Theorem 3, however, may admit an exceptional set of measure zero, like any such unit-probability result. It remains to show that this must be

empty for the particular sample-sequences of primes in covering intervals. Consider two I_n whose coverings do not differ by more than u on the y-scale. The number of integers covered by any such I_n is not greater than $C \log n$ for large n. Therefore, the difference in the number of primes S_N for two different sample-sequences cannot be of greater order than $\log N$. This does not affect the order of magnitude as given in 3.2 which is therefore true for all covering sequences without exception. The result as translated here is thus proved.

The consequences of Theorem 4 are sufficiently well known to number-theorists and need not be detailed here.

REFERENCES

1. Prachar, K. .. *Primzahlverteilung*, Berlin, 1957, p. 157 *ff.*
2. Ricci, G. .. "Sul pennello di quasi-asintoticità della differenza di interi primi consecutivi," *Atti. Accad. Naz. Lincei* (*Rendiconti*), serie 8, 1954-55, **17**, 192–96 and 347–51.
3. In a private communication of Prof. P. Erdös to Prof. D. D. Kosambi.
4. Kosambi, D. D. .. "The sampling distribution of primes," *Proc. Nat. Acad. Sci.* (*USA*), 1963, **49**, 20–23.
5. Feller, W. .. *An Introduction to Probability Theory and its Applications* (New York), 1950, Vol. 1, for the laws of large numbers and of the iterated logarithm, pp. 157–61.
6. ———— .. *Ibid.*, p. 154.
7. Ducray, S. .. "Normal Sequences," to appear in the *J. Uni. Bombay.*