THE SEQUENCE OF PRIMES

By S. Ducray

Received March 31, 1965

(Communicated by Sir C. V. Raman)

This note derives by analytic methods certain results in number theory developed elsewhere[1] from probability considerations. The presentation is not merely a translation into measure-theoretic terms. The approach is fundamentally different and does not rest upon sampling.

1. Lemmas in Number Theory

The following definitions and notation are used. The semi-infinite real line $x \geq x_0 \geq 2$ is transformed into $y \geq 0$ by

$$y = Li(x) - Li(x_0) = \int_{x_0}^{x} \frac{dt}{\log t}. \tag{1}$$

The positive integers are marked off on the x-line at unit intervals. Any set of the y-line is said to contain (or cover) a certain number of primes p if its x-image does so. The y-line is itself covered by the infinite sequence of intervals I_n: $(n-1)u \leq y < nu$; $n = 1, 2, 3, \cdots$ with $u > 0$ arbitrary but fixed for any given covering. The number of primes contained in I_k is denoted by X_k and is obviously a function of x_0, u, k, taking on the values $0, 1, 2, \cdots$. However, no finite number of such values determines the initial point x_0, so that the complete sequence for any given u cannot be determined by specifying some of its members. By S_n is meant the sum of the first n members of the sequence $\{X_n\}$.

Lemma 1. $S_n \sim nu$ *and* $S_n/n \to u$ *for all* x_0.

This is simply the prime-number theorem in our notation.

Lemma 2.—*For any two initial points* x_0, x_0', $S_n - S_n' = 0\,(\log n/\log\log n)$.

This follows from the known number-theoretic result that a length h of the x-line cannot contain more than $ch/\log h$ primes; here the terminal h is of order $\log n$ by hypothesis and the transformation (1).

LEMMA 3.—*The totality of distinct sequences* X_n *obtained by displacement of the initial point* x_0 *through a single* u*-interval contains a subset which maps in a* $1-1$ *manner onto the points of the interval* $0 \leq t < 1$.

This is theorem 1 of the paper cited[1]; it follows very simply from gap theorems due to P. Erdös and G. Ricci. Hereafter, we deal only with the subset of sequences that can be so mapped, without loss of generality, as will be seen.

2. LEMMAS ON MEASURE

The third lemma above gives a simple measure for the subset of sequences $\{X_n\}$, namely Lebesgue measure on $(0, 1)$. As the parameter t of the mapping shifts from 0 to 1, X_n will assume various values. These will be finite in number; zero or a positive integer. Moreover, each such value will have a measure attached to it because of the nature of the mapping employed. This last follows quite simply from the proof originally given for lemma 3, which shows that each value of X_n is assumed over a finite number of t-intervals, whence the measure would be the sum of the interval lengths. The existence of the measure and measurability are therefore not in doubt; the total measure for any X_n over all its possible values is obviously unity.

The *expectation* of any function $f(X_n)$ is defined as the Lebesgue integral of f, if it exists. It will be denoted by $E(f)$.

LEMMA 4.—*With the mapping and measure of lemma* 3, *we have,*

$$\Sigma E(X_r) \sim nu;\ \Sigma E(X_r^k) = 0(n) \tag{2}$$

for all k; *summation over* $1 \leq r \leq n$.

Of these, the first is derived from lemma *1* by interchanging the order of summation and integration. The second follows from certain estimates made by P. Erdös which extend Viggo Brun's work on prime pairs and were applied to find upper limits for the frequency with which $X_r = k$ could occur in any S_n, in a paper of Kosambi.[2]

The step-functions X_i, X_j are called independent (in measure) if for every pair of values r, k assumed, the measure of the t-set over which $X_i = r$ and $X_j = k$ simultaneously is the product of the two individual measures concerned. Similarly for independence of three or more of the X's. If the X's are independent no matter what finite number of them (however large) be taken, we shall say that they are completely independent, or just *independent*

without further qualification. There arise two cases, according to the complete independence or the X's, or otherwise.

Complete independence implies a *canonical mapping*, wherein each possible value of X_{n+1} is associated with the same proportional measure with each possible value of X_n. Without changing the values that occur or changing the measure with which each occurs in X_n, such a mapping would map all possible sequences $\{X_n\}$ with the given values and measures onto $0 \leq t < 1$. Our original sets are not by any means those of 'all possible sequences with the given values', as is again obvious. If, however, the set of sequences of primes $\{X_n\}$ happens to be of positive canonical measure the following would be true:

If the set of prime-sequences $\{X_n\}$ *be of positive measure in the canonical measure, then there exist two constants* $C > c > 0$ *such that*

$$-C\sqrt{n \log\log n} < s_n - nu < C\sqrt{n \log\log n}; \tag{3}$$

but each inequality would be false infinitely often as $n \to \infty$ *if* C *were replaced by* c.

The main idea is that the Khinchin law of the iterated logarithm applies in the case of independence, to almost all sequences, *in the canonical measure*. But inasmuch as our sets of sequences of primes in covering intervals is of positive measure here by hypothesis, it applies to almost all of the sequences actually obtained. There can be no exceptional sequences of primes in covering intervals, because of lemma *2*, which restricts the difference of sums to a lower order than required by (*3*), no matter how far apart the initial points may be. The mean values required by the law of large numbers are guaranteed by lemmas *1* and *4*, while lemma *4* also ensures that the dispersion (variance) is properly restricted, when we take $k = 2$ in the second of (*2*).

However, independence is difficult to prove even in this restricted sense, and not necessary for equation (*3*), though the other law with c in place of C does require it. We proceed to supply the equivalent needed for the validity of (*3*) as it stands.

3. Major Conclusions

Hereafter, we take some fixed $u > 0$, and an unspecified x_0 displaced through the x-image of a u-interval, with the set of sequences of primes (in covering intervals) $\{X_n\}$. Therewith, take the subset, mapping and measure of lemma *3*. The main question before us is the effect of any

possible lack of independence. The sequences are generated by the sieve of Eratosthenes, but x_0 runs through an unspecified set of values, hence the sole information available would be, at most, that such and such terms have occurred in a given sequence. What interests us, therefore, is the effect of S_n assuming a certain range of values, upon the sum of further terms. This, so far as relevant to the problem, may be stated as:

LEMMA 5.—*Given* r *and* m, *each of order* $\sqrt{n}$, *and the fact that* $S_{n+r} - S_n > ru$. *Then, in general, this may decrease the measure for* $S_{n+r+m} - S_{n+r} > mu$, *when compared to the case of complete independence, but cannot increase it. Similarly when both inequalities are reversed.*

That is, an excess above or deficiency below expectation in sums of this order of terms cannot be cumulative when deviations from expectation are considered; they could conceivably be compensatory in that sense. The proof needs only qualitative use of the sieve. For k consecutive covering intervals, the x-image contains roughly $\sim kulogn$ integers. The composite numbers from among these in the range taken would have a prime $p < \sqrt{nu \log nu}$ as deleting factor, and each such prime will have increasingly many multiples in the two ranges selected above. That there are unusually many prime numbers left in r consecutive covering intervals after the first n merely says that the deleting primes $p < \sqrt{nu \log nu}$ have multiplied each other oftener than would be expected on the average. If this implies anything about deletion in the next m covering intervals, can only be either: (*a*) there is no effect at all, or (*b*) the deleting primes, at the very worst, cannot multiply each other as often as on the average. An excess over expectation *mu* is then less likely than with independence in these m intervals. The same arguments may be repeated when inequalities are reversed.

This absence of cumulative effect, in itself quite obvious, leads to our next and main result.

THEOREM.—*Regardless of the complete independence of the* $\{X_n\}$ *in measure, or lack of such independence, there exists a constant* $C > 0$ *such that taking* $\pi(x)$ *as the number of primes* $p \leq x$, *we have*

$$-C\sqrt{x \log\log x/\log x} < \pi(x) - Li(x) < C\sqrt{x \log\log x/\log x} \quad (4)$$

Proof.—In the case of independence, or when the set is of positive canonical measure (3) and its complement for c in place of C hold. Putting in the proper values of x and y from (*1*), the formula (*4*) emerges, and much more besides,

In the opposite case, we have to note that (*3*) by itself forms only the *upper* law of the iterated logarithm. This rests upon the first Borel-Cantelli lemma, which is a proper measure-theoretic result though stated as for probability. It does not require complete independence in measure, nor does it rest upon the use of a canonical measure. Any totally finite outer measure would do, and lemma *3* furnishes one such. The proof of the upper law of the iterated logarithm turns upon two points: That the measure for deviations from the mean of sums (in either direction) should not exceed that with independence. Secondly, the measure for an extreme deviation from the expectation ku for the sum exceeding $C\sqrt{n \log\log n}$ for some intermediate sum of k terms should not itself exceed a fixed constant times the measure for n terms. Both of these follow in an obvious manner from lemma *5*, which says that the deviations cannot be cumulative. For k consecutive terms of order not exceeding $\sqrt{n}$, the reasoning adopted in proof of lemma *2* actually make the measure for large deviations zero, where it would be small but positive for total independence. Finally, lemma *2* prevents exceptional values of x_0 from arising.

References

1. Ducray, S. .. "Probability and prime numbers," *Proc. Ind. Acad, Sci.*, 1964, **60 A**, 159–64, with a correction to the first condition of lemma 1 there which requires gaps greater than any *ku*.

2. Kosambi, D. D. .. "The sampling distribution of primes," *Proc. Nat. Acad. Sci, U.S.A.*, 1963, **49**, 20–23.