
THE SAMPLING DISTRIBUTION OF PRIMES

By D. D. KOSAMBI

P. O. DECCAN GYMKHANA, POONA 4, INDIA

Communicated by H. S. Vandiver, November 2, 1962

The real half-line $x \geq x_0 \geq 2$, upon which the integers are marked off unit distance apart, is mapped onto $y \geq 0$ by the transformation $y = \int_{x_0}^{x} dt/\log t = \mathrm{li}(x) - \mathrm{li}(x_0)$. Cover the whole of $y \geq 0$ by a sequence of intervals, each of length $u > 0$, fixed. The nth such interval will be $(n - 1)u \leq y < nu$, and $\pi_n(u) = \pi(x_0, u; n)$ denotes the number of primes in its x-image. We show that the primes in an arbitrary connected stretch of the y-line have a Poisson distribution in the sense of probability theory, the sequences $\pi_n(u)$ constituting statistical *samples* thereof. Hereafter, take all positions of the initial point (on the y-line) as equally likely and x_0 neither restricted nor specified otherwise.

Textbook results in number theory and probability theory are taken for granted. In particular, LEMMA 1. *The number of primes $p \leq x$ is $\sim li(x) \sim y$ (for any x_0, as $x \rightarrow \infty$). If $\vartheta(x) = \sum \log p$, $p \leq x$, then $\vartheta(x) \sim x$. If p_k be the kth prime in order, starting from $p_1 = 2$, then $p_k \sim k \log k$.*

The first of these is the prime-number theorem,[1] and the other two are equivalent, as is well known.

LEMMA 2. *For $p \leq x$, $\Pi(1 - 1/p) \sim e^{-\gamma}/\log x$; γ, Euler's constant.*

This is a classical theorem of Mertens.[2]

LEMMA 3. *If for any set Z of primes, $\Pi p = x$, $p \subset Z$, then $\Pi(1 - 1/p)^{-1}$ is less than $C \log_2 x$, $p \subset Z$, x large.*

Proof: The product of $(1 - 1/p)^{-1}$ will be greatest for any given number of primes if the primes are 2, 3, . . . in sequence and all distinct. Then $\log x = \log \Pi p = \sum \log p$ by hypothesis, $p \subset Z$. Lemma 1 says that, packing the primes at the beginning of the sequence, $\max p \sim \sum \log p$, and here $\sum \log p = \log x$. By Lemma 2 (the product being not greater than in this case) $\Pi(1 - 1/p)^{-1} < C \log_2 x$, $p \subset Z$. *Q.E.D.*

LEMMA 4. *The proportion of u-intervals for which $\pi(x_0,u,n) \geq 2$ is less than cu^2 for small u, regardless of x_0, if x is large.*

Proof: The sieve of Viggo Brun leads to the theorem:[3] *The number of primes $p \leq x$ for which $p + b$ is also a prime is $< (cx/\log^2 x)\Pi(1 - 1/p)^{-1}$, $p|b$.* The u-intervals containing two or more primes must contain one such pair $p,p + b$ for some $b \leq u \log x$ approximately. Not all b, however, are admissible, as no odd b will do for $p > 2$. The number of admissible b's within the same u-interval is easily seen to be not greater than the number of integers in (the x-image of) the covering interval prime to $N = 2.3 \ldots p$, provided $N \lesssim u \log x$. Clearly, $p + b$ not a prime to N cannot be a prime except in the interval that begins from $x_0 = 2$, which may be ignored; moreover, such numbers are arranged cyclically modulo N, which, being about the length of the interval on the x-axis, cannot be materially changed in the vicinity of any given x. By Lemmas 2 and 3, the admissible set will contain less than $c'u \log x/\log_3 x$ members, for large x. The bound for $\Pi(1 - 1/p)^{-1}$ for primes dividing any b in the interval cannot ultimately be greater than $c'' \log_3 x$. Finally, the total number of covering intervals in the range is $\sim x/u \log x$. The estimate therefore is not in excess of $(cx/\log^2 x)(c'u \log x/\log_3 x)(c'' \log_3 x)(u \log x/x) = \bar{c}u^2$. *Q.E.D.*

LEMMA 5. *If $f_0,f_1,f_2, \ldots$ be the relative frequencies, $\sum f_i = 1$, with which small u-intervals containing 0,1,2, . . . primes occur in a large range of x, then $f_1 = u + o(u)$.*

Proof: Corresponding to the theorem cited in the proof of Lemma 4 is an extension by P. Erdös:[4] *The number of primes $p \leq x$ for which all the numbers $p + b_1, p + b_2, \ldots p + b_r$, $0 < b_1 < b_2 < \ldots < b_r$ are also primes is less than*

$$(cx/\log^{r+1} x) \prod_{p|E} (1 - 1/p)^{-(r+1-\omega(p))}, \qquad E = \prod_{i=1}^{r} b_i \prod_{1 \leq i < k \leq r} (b_k - b_i) \quad (1)$$

where $\omega(p)$ is the number of solutions mod p of $m(m + b_1) \ldots (m + b_r) \equiv 0 (\mathrm{mod}\ p)$. From this point, the reasoning of the previous lemma holds, except that the number of choices for the set of r b's will not exceed the binomial coefficient ${}^nC_r$, with $n = c \log x/\log_3 x$ and $\Pi p,p|E$ cannot exceed $(u \log x)^m$, with $m = r^2(r - 1)/2$ (an overestimate which we shall not stop to refine). The upper bound, for small u, is therefore $cu^{r+1}/r!$ for each r, and the same c may be taken throughout, quite obviously. For any u, the contribution of $f_2, f_3, \ldots$ to the expectation (mean value, average) of primes per covering interval may be assessed as not exceeding cu^2e^u. For, this mean value is $(0.f_0 + 1.f_1 + 2.f_2 + \ldots)$, so that f_0 contributes nothing. Any term from f_2 onwards, as assessed above, will contribute $0(u^2)$. The total contribution of those terms will be $0(u^2e^u)$, as may be seen from the upper bounds just given above. Now the mean value, by the prime number theorem, is exactly u, over the whole y-line, no matter what the x_0. It follows that for small u, $f_1 = u + 0(u^2)$. *Q.E.D.*

THEOREM. *With all x_0 equally likely, the probability that exactly r primes will lie in the x-image of $0 \leq y < t$ is $e^{-t}t^r/r!$ (the Poisson distribution, with parameter t).*

Proof: Given x_0, there is no question of any probability; the entire sample is completely defined for the whole y-line. But under the present conditions, the irregularity of primes permits the use of the concept "probability" the "event" being 0, 1, 2, ... primes lying in the interval $0 \leq y < t$. These events are exhaustive and mutually exclusive. The conditions for a Poisson process are given by the following postulates:[5] The probability for one prime in $t \leq y < t + h$ for small h is $h + o(h)$; the probability for more than one prime in the small interval is $o(h)$; the probability for the small interval being totally void of primes is $1 - h + o(h)$. Lastly, none of these are affected if it is known that k primes have actually occurred in $0 \leq y < t, k = 0, 1, 2 \ldots.$

These postulates are obviously satisfied in view of our lemmas above. Lemma 4 says that the probability (approximated arbitrarily closely by the corresponding frequency) for more than one prime in the small interval is $o(h)$. Lemma 5 gives the probability for a single prime as $h + o(h)$. Since these two cases and that of the h-interval being void of primes are mutually exclusive and exhaustive, the third postulate is satisfied. Finally, the lemmas hold regardless of x_0 and t, over the whole of the y-line, $y > t$. Moreover, the number of primes known to have occurred in $0 \leq y < t$ does not in any way affect the frequencies or probabilities or permit x_0 to be determined even approximately. (It is possible to go much further in this direction, for not even the precise knowledge of the points $t_1, t_2, \ldots$ at which these primes may actually have occurred changes the situation. If it could then be said that there *must* exist a prime in $t \leq y < t + h$, no matter how small the h, it would follow that the k + 1st prime could be located from the positions on the y-line of the first k, for all large primes and some k. This implies a recurrence relation between the primes; no such relation is known and an algebraic one of any finite degree is demonstrably impossible. There is no finite upper bound for the gap between consecutive primes on the y-line[6] and no known positive lower bound. On the other hand, it is known that subsequences of primes (of positive density) exist[7] for which the y-distance between consecutive primes is dense over a certain positive range, whose precise termini are not known. This shows the impossibility of using any but probability methods.) *Q.E.D.*

The Poisson distribution of our theorem may be quickly derived as follows. For the argument, allow x to be any point (with equal likelihood) of a range $R(x) \approx x^\alpha$, $38/61 < \alpha < 1$. It is known (Ingham, A. E., *Quart. J. Math.*, **8**, 255–266 (1937)) that the prime-number theorem holds asymptotically over $R(x)$ as $x \to \infty$. Further, let $I(x)$ be a randomly selected interval within $R(x)$ of y-length t, hence containing $\sim t \log x$ integers regardless of position (since the variation in $\log x$ is negligible over $R(x)$). No matter where $I(x)$ is located, alternate integers in it must be even, four out of every six (regularly arranged) divisible by 2 or 3, etc. This regularity of deletion by the sieve of Eratosthenes extends to all the smallest primes whose product $2.3.5 \ldots p = N \leq t \log x$. About $te^{-\gamma}\log x/\log_3 x = tg(x)$ integers in $I(x)$ will survive. Any p not a factor of N need not be the smallest prime factor of a surviving integer in $I(x)$ and a prime larger than $t \log x$ need not even have a multiple in $I(x)$, so that one of the "survivors" being deleted by any such prime is now a matter of chance with probability $1/p$. By the prime number theorem, the expectation of primes in $I(x)$ is exactly t (in the limit), hence the compound probability for primality of a "survivor" is asymptotic to $1/g(x)$. Moreover, if some k of these survivors be tested and found composite or prime (without revealing their numerical values), the knowledge does not modify the probability for primality for the rest. In all this, x is merely a background parameter, whose principal use is to furnish relative magnitudes of the various functions involved, as $x \rightarrow \infty$.

It follows that if P_r be the probability for precisely r primes in $I(x)$, then in the limit $P_0 = \lim(1 - 1/g)^{tg} = e^{-t}$. Using textbook definitions and procedures, the limit $P_1 = \lim(1 - 1/g)^{tg-1}(tg)(1/g) = te^{-t}$, and so on, with limit $P_r = e^{-t}t^r/r!$ But any limiting distribution over $R(x)$ as $x \rightarrow \infty$ will obviously be the distribution over the entire x-line, here the Poisson distribution with parameter t, as before.

[1] Prachar, K., *Primzahlverteilung* (Berlin, 1957), ch. 3.

[2] Hardy, G. H., and E. M. Wright, *An Introduction to the Theory of Numbers* (Oxford University Press, 1945), Theorem 430, pp. 349–354.

[3] Prachar, K., *op. cit.*, ch. 2, Theorem 4.4.

[4] *Ibid.*, Theorem 2.4.7.

[5] Feller, W., *An Introduction to Probability Theory and Its Applications* (New York: 1950), vol. 1, p. 366 *et passim.*

[6] For general known results on gaps in the sequence of primes, see Prachar, *op. cit.*, p. 154 ff.

[7] Ricci, G., "Sul pennello di quasi-asintoticità delle differenze di interi primi consecutivi," *Rend. Atti. Accad. Naz. Lincei*, **8**, 192–196 and 347–351 (1954–5).