UCRL-16902

University of California

Ernest O. Lawrence Radiation Laboratory

STATISTICAL TREATMENT OF RPIME NUMBER DISTRIBUTIONS

Submitted to Mathematics of
Computations

UCRL-16902
Preprint

UNIVERSITY OF CALIFORNIA
Lawrence Radiation Laboratory
Berkeley, California

AEC Contract No. W-7405-eng-48

STATISTICAL TREATMENT OF PRIME NUMBER DISTRIBUTIONS

Edwin M. McMillan

May 23, 1966

STATISTICAL TREATMENT OF PRIME NUMBER DISTRIBUTIONS

Edwin M. McMillan

Lawrence Radiation Laboratory
University of California
Berkeley, California

May 23, 1966

1. Introduction

The distribution of prime numbers presents an apparent paradox: A completely determinate process, the sieve of Eratosthenes, leads to a sequence of primes which seems to be highly random. However, it is not hard to understand why this should be so. Suppose that we construct a "partial sieve" in which the primes $p = 2$ through $p = Q$ are eliminated, including the first occurrence of each prime. The resulting sequence of "potential primes" repeats with the period $\prod_{2}^{Q} p$, each period having a center of symmetry. The length of the period increases very rapidly with Q; for example, with $Q = 31$ it is about 2×10^{11}. On the other hand, the fraction of numbers remaining as potential primes is equal to $\prod_{2}^{Q} \frac{p-1}{p}$, which decreases rather slowly with Q, approximately like $1/2 \ln Q$; this fraction has the value 0.153 for $Q = 31$. The partial sieve is empty from 1 (which must be included as a potential prime to represent the repeating sequence properly) to the next prime Q_1 above Q. From Q to Q_1^2, (in our example, from 31 to 1369) the potential primes are real primes, but from there on most of the potential primes are composite, with factors above Q. Thus the partial sieve has regularity on a scale which rapidly outruns the region in which it determines the actual sequence of primes, and it seems reasonable to assume a random distribution of

primes in the remaining possible places. This is somewhat analogous to a computer routine for generating random numbers, in which a determinate process leads to a situation so involved that the determinateness becomes obscured. In other words, lack of manifest order is interpreted as randomness.

The partial sieve has absolute negative consequences. For example, when constructed through 2, it forbids the occurrence of consecutive primes, except for the pair 2, 3. Constructed through 3, it forbids sequences like p, p+2, p+4 or p, p+4, p+8 (except for ones including 3) but allows an endless sequence of primes differing by 6 or any multiple of 6. The elimination of 5 breaks this sequence, but allows an endless sequence differing by 30. Continuing the process, we get a general rule: the maximum number of primes that can be in a geometrical progression with a given interval is $p-1$, where p is the smallest prime that is not a factor of the interval, unless the sequence starts with p, when one more is allowed. Rules of this kind are, however, only permissive; the occurrence of primes in allowed configurations can be found only by trial, or their frequency can be estimated by statistical considerations.

2. Frequency Distribution of Numbers of Primes in Intervals

Consider now the problem: In a region where the density of primes is D, what is the frequency distribution $\nu_n(N, D)$ of the number n of primes occurring in an interval of fixed length N placed randomly in the region? Imagine that a partial sieve is constructed, with the fraction of numbers remaining as potential primes being equal to F.

An interval N then contains an average of FN potential primes and an average of ND real primes, so that the probability that any potential prime remains real is D/F. As the interval is moved along the sieve the number of potential primes within the interval will vary; let f_i be the fraction of intervals that contain i potential primes. For a group of intervals with fixed i, the frequency distribution of primes is the binomial distribution given by the expansion of

$$\left[\left(1 - \frac{D}{F}\right) + \frac{D}{F}\right]^i .$$

Adding the distributions for all values of i, we get:

$$\nu_n = \sum_{i \geq n} f_i \binom{i}{n} \left(\frac{D}{F}\right)^n \left(1 - \frac{D}{F}\right)^{i-n} \tag{1}$$

where $\binom{i}{n}$ is a binomial coefficient. Equation (1) can be rearranged to collect powers of D/F, with the result:

$$\nu_n = \sum_{m \geq n} (-1)^{m-n} \binom{m}{n} g_m (D/F)^m , \tag{2}$$

$$g_m = \sum_{i \geq m} \binom{i}{m} f_i . \tag{3}$$

Thus $g_0 = \sum f_i = 1$, $g_1 = \sum i f_i$ = mean value of i = FN,

$g_2 = \sum i (i-1)/2 f_i$, etc.

Also $$\bar{n} = \sum n \nu_n = g_1 (D/F) = DN,$$

$$\overline{n(n-1)} = \sum n(n-1) \nu_n = 2 g_2 (D/F)^2 .$$

From the last two relations we can obtain σ^2, the mean square deviation of n from its mean value $\bar{n}$:

$$\sigma^2 = \bar{n} \left[1 - \bar{n} \left(1 - \frac{2 g_2 D^2}{\bar{n}^2 F^2} \right) \right]. \qquad (4)$$

The next step is to evaluate the g's. The meaning of (3) can be expressed in the following way: f_i is the fraction of the intervals containing exactly i potential primes; g_m is the fraction containing all cases in which the presence of at least m potential primes is assured, regardless of how many others there are. To compute g_2, for example, we consider each combination of two places in the interval, allow the interval to assume all possible relations to the partial sieve, compute the fractions of these that leave potential primes in both places, and add these fractions for all combinations. It is easier to think of the interval remaining fixed, while the elements of the sieve are independently moved to all locations. In evaluating g_2, we start with the sieve constructed through 2, giving two locations with respect to the interval, which are to be handled separately and the results added. After this, each prime has p positions, of which $p-2$ avoid the chosen pair; the desired fraction for the chosen pair is then $\frac{1}{2} \prod_3^Q \frac{p-2}{p}$ unless the interval between the pair is a multiple of any p in the product. In that case, the corresponding factor should be $\frac{p-1}{p}$; this is taken care of by using the same product and multiplying by $\frac{p-1}{p-2}$. For p_3 and p_4, we start with the sieve

constructed through 3, and the products are $\frac{1}{6}\prod_{5}^{Q}\frac{p-3}{p}$ and $\frac{1}{6}\prod_{5}^{Q}\frac{p-4}{p}$; for p_5 and p_6, we start with the sieve constructed through 5, and the products are $\frac{1}{30}\prod_{7}^{Q}\frac{p-5}{p}$ and $\frac{1}{30}\prod_{7}^{Q}\frac{p-6}{p}$. The factors for commensurable intervals are $\frac{p-2}{p-3}$, $\frac{p-3}{p-4}$, $\frac{p-4}{p-5}$, $\frac{p-5}{p-6}$ for these cases. (If more than one interval in a given combination is commensurable with p, other factors will be needed; their derivation is obvious.)

Each g_m thus contains a divergent continued product. However, in (2) it is multiplied by $(1/F)^m$, with F given by $\prod_{2}^{Q}\frac{p-1}{p}$, which diverges in a compensating way; the product g_m/F^m is convergent. It is therefore possible to allow Q to become large, arriving at a unique result. The requirement that $F < D$ implies approximately that $Q < \sqrt{k}$, where k represents the position in the number scale where one is working. This limit is not important in the cases to be considered, and the continued products will be carried to infinity.

This treatment contains a tacit assumption that the partial sieve is "well mixed." The region over which the potential primes represent a complete set of all noncoinciding locations of the elements of the sieve is a repeat length, and the treatment is therefore strictly valid only for a repeat length. When it is applied to smaller regions, it is assumed that the arrangements of the elements are randomly selected from the complete set; this is related to the assumption of randomness for the filling of the remaining spaces.

Letting p_1 be the first prime above m, and p_0 the first prime below p_1, we can now write

$$g_m \left(\frac{D}{F}\right)^m = a_m A_m D^m , \tag{5}$$

$$A_m = \prod_{2}^{p_0} \frac{p^{m-1}}{(p-1)^m} \prod_{p}^{\infty} \frac{p^{m-1}(p-m)}{(p-1)^m} . \tag{6}$$

The coefficient a_m is the count of combinations of m potential primes found in a set of intervals containing all of the $\prod_{2}^{p_0} p$ possible locations of the partial sieve constructed through p_0, each entry into the count being multiplied by the appropriate commensurability factor where required.

The factors of A_m approach $[1 - \binom{m}{2}/p^2]$ as p gets large; therefore the products converge. A_2 has been evaluated by Shanks [1]. Ratios of powers of the A's can be made that converge rapidly, with the typical factor $(1 - \text{const}/p^3)$; this fact can be used to evaluate other A's in terms of A_2. Some values are

$$A_1 = 1$$

$$A_2 = 1.3203236\text{---}$$

$$A_3 = 2.859\text{----}$$

$$A_4 = 4.14\text{---}$$

$$A_5 = 10.1\text{---}$$

$$A_6 = 17.3\text{---}$$

Values of the a's to N = 20 and of a_2 to N = 30 are given in Table I. The value for an odd N is the mean of the values for the adjacent even N's. The first occurrence of a_7 is at N = 21.

[1] D. Shanks, Solved and Unsolved Problems In Number Theory, vol. 1, p. 30 (Washington 1962).

Table I

N	a_1	a_2	a_3	a_4	a_5	a_6	N	a_2
2	2	0	0	0	0	0	22	144-4/15
4	4	2	0	0	0	0	24	176-2/9
6	6	6	0	0	0	0		
8	8	14	4	0	0	0	26	$210-\frac{8}{45}$
10	10	24	12	2	0	0	28	$246-\frac{52}{165}$
12	12	36-2/3	26	8	0	0		
14	14	53-1/3	52	34	4	0	30	$284-\frac{422}{495}$
16	16	72-2/5	88	78	14	0		
18	18	93-7/15	132	146	36	2		
20	20	118-8/15	196	274	86	16		

3. Comparison with Empirical Data

Stein and Ulam [2] [3] have made a compilation of ν_m for intervals which increase with k in such a way that $\bar{n} \approx 1$, and present the cumulative values for several values of k. From their results one can pick ranges in which the interval N is constant, and find the values of ν_n for those ranges. Two of their ranges are given in Table II, in the columns marked "S & U."

Table II

Range of k	500,000 to 1,000,000			80,000,000 to 100,000,000		
N	14			19		
	S & U	calc. 1	calc. 2	S & U	calc. 1	calc. 2
ν_0	.2922	.2935	.294	.3068	.3063	.307
ν_1	.4358	.4293	.430	.4166	.4164	.415
ν_2	.2198	.2288	.228	.2158	.2178	.215
ν_3	.0475	.0444	.048	.0541	.0530	.054
ν_4	.0041	.0038	---	.0064	.0063	.006
ν_5	.0000	.00009	---	.0003	.00029	---
ν_6	.0000	0	---	.0000	.000004	---
$\bar{n}$	1.0349	1.0359		1.0376	1.0375	
σ^2	.7381	.7337		.8000	.7956	
x			3.553			4.451

Calculations made as described above, assuming that D=(1/ln k), and including small corrections for the variation of D over the intervals, are given in the columns marked "calc. 1". The agreement is very good, especially in the higher range, where one would expect a statistical theory to be better.

---

[2] S. M. Ulam, private communication.

[3] M. Stein and S. M. Ulam, An Observation on the Distribution of Primes, Am. Math. Monthly (in press).

4. A Simple Approximation

Since the most important parameters of a distribution are $\bar{n}$ and σ^2, and the distributions computed in Calc. 1 represent superpositions of binomial distributions, one may wonder how well the distributions will be represented by single binomial distributions with parameters chosen to give the same values of $\bar{n}$ and σ^2. In the binomial distribution given by the expansion of

$$\left[\left(1 - \frac{\bar{n}}{x}\right) + \frac{\bar{n}}{x}\right]^x ,$$

$$\sigma^2 = \bar{n}\left(1 - \frac{\bar{n}}{x}\right) . \qquad (7)$$

Setting this σ^2 equal to the value given by (4),

$$\frac{1}{x} = 1 - \frac{2\, a_2\, A_2}{N^2} . \qquad (8)$$

Since x is not in general an integer, the binomial expansion does not terminate, but the terms with $n < x$ seem to give a good representation of the major part of the distribution, as shown in the columns marked Calc. 2, computed in this way. The values of x are given in the table. The quantity x, which is a function of N only, can be considered the "capacity" of the interval, that is, the number of spaces in which the $\bar{n}$ primes must be distributed randomly in order to represent the major features of the distribution.

5. Distribution of Intervals Between Primes

The simplest problem is to find the density of intervals d between primes, regardless of the presence of intervening primes. The procedure for this is the same as that described above for evaluating

$g_2(D/F)^2$, except that only pairs differing by d are considered, N is allowed to become large, and the count of intervals is divided by N. The result is

$$\text{Density of intervals of length } d = A_2 D^2 \Pi_{(p/d)} \frac{p-1}{p-2}, \qquad (9)$$

where the product is taken over all odd primes p that are factors of d. This formula covers two previous results: that of Shanks [1], who computed the density of twin primes (d = 2), and that of Pólya [4], who gave the ratio of the interval density for any d to that for d = 2. Shanks compared his computation with the number of twin primes in the range 1 to 37,000,000, with agreement to one part in a thousand. Pólya gave a compilation of interval ratios to d = 70 in the range 1 to 30,000,000; these fit the formula very well, including d = 62, with the prime factor 31.

It is a more difficult problem to find the frequency of empty intervals, or gaps of length G, between primes. Equation (9), with d = G, must be multiplied by a factor that gives the probability that no primes exist within the interval. The computation of this probability is related to that of ν_0, but is modified by the constraints on the location of the elements of the sieve imposed by the presence of the two known primes at the ends of the gap. This computation has been carried out to G = 10, the largest gap that contains at most two potential primes, and the results are compared with some gap counts made by P. and M. Stein and sent to me by Ulam [2]. The counts for the range 80,000,000 to 100,000,000 are entered in Table III in the column marked "S & U." The computed values are given in the column marked "calc."

(4) G. Pólya, Heuristic Reasoning in the Theory of Numbers, Am. Math. Monthly 66, 375 (1959).

These are obtained by multiplying 20,000,000 A_2D^2 by the factors given in the last column, assuming D = (1/ln k), and making a correction for the variation of D over the range.

Table III

G	S & U	calc.	factor
2	78,862	78,740	1
4	78,911	78,740	1
6	138,855	138,860	$2\,[1 - (A_3/A_2)\,D]$
8	60,796	60,860	$1 - 2\,(A_3/A_2)\,D + (A_4/A_2)\,D^2$
10	78,522	78,552	$4/3\,[1 - 9/4\,(A_3/A_2)\,D + 3/2\,(A_4/A_2)D^2]$

Cases in which intervening primes are required can also be considered. Examples are "double twins," like 5, 7, 11, 13, and "bracketed twins," like 7, 11, 13, 17. The computed densities for these are A_4D^4 and $2A_4D^4$ respectively. Another variation concerns configurations in which certain intervening primes are required, while others are disregarded. The example chosen to illustrate this is a geometrical progression of six primes with a common difference equal to 30, like 7, 37, 67, 97, 127, 157. The computed density for this case is $8A_6D^6$. These formulas predict that in the range 80,000,000 to 100,000,000, there will be 740 double twins, 1480 bracketed twins, and 75 geometrical progressions of 6 primes differing by 30. As far as I know, counts of these have not been made.

6. Discussion

The method developed in Section 2 appears to be a valid way of applying statistical methods to prime number distributions, but the work becomes very involved for large values of N; the number of cases of commensurability which enter into the computation of a_m increases rapidly, and the alternating signs of Eq. (2) place a high requirement on the accuracy of these coefficients. However, in Section 4, it is shown that only a_2 is needed to determine the major part of the distribution for large N. This offers a simple method for handling these cases. The evaluation of a_2 is easily done by a systematic procedure:

If N is even, the number of pairs of positions in the interval contributing to a_2 is $\frac{N}{2}\left(\frac{N}{2} - 1\right)$. If N is even and divisible by a prime p, the amount to be added because of the commensurability factor is $\frac{N}{2}\left(\frac{N}{2p} - 1\right)\frac{1}{p-2}$. If N is even and divisible by any product $p_1 p_2 ---$, the further amount $\frac{N}{2}\left(\frac{N}{2p_1 p_2 ---} - 1\right)\frac{1}{(p_1-2)(P_2-2)---}$ must be added. For a general value of N, the number to be used in each case is a linear interpolation between those given by the above formulas for particular values of N.

A remarkable feature of the distribution of prime numbers is the behavior of σ^2 as N is varied in a region of constant D. It is found that, because x increases less rapidly than N, σ^2 also increases less rapidly than N. If an interval is made up of adjacent smaller intervals, σ^2 for the whole interval is less than the sum of the values of σ^2 for its parts. This behavior is caused by the fact that adjacent intervals are not independently random in their locations with respect to the partial

sieve. The value of σ^2 is always less than $\bar{n}$, and may be much less in some cases. For $N = 60$ ($x = 10.8018$) and $D = 1/12$ ($k = 160{,}000$), the value 0.54 is computed for $\sigma^2/\bar{n}$. One cannot go to much larger values of D without getting into a region of k too small to give confidence in statistical reasoning, but much smaller values of $\sigma^2/\bar{n}$ will be found for very large values of N. Prime distributions are not as random as one would guess from conventional statistical arguments, and the agreement between computed frequencies and the empirical values will usually be found to be better than such conventional arguments would suggest.

It will be interesting to collect further empirical data for various values of N and D. In the collection of such data, the intervals should either be distributed randomly throughout the range of k, or they should be taken at every value of k, as was done in Ref. 3. This avoids any correlation with the periodicities of the number scale.

Acknowledgments: I should like to thank Martin Gardner, who published a problem in The Scientific American that started me thinking about the statistics of prime numbers; Stanislaw Ulam, who provided data that gave me the impetus to continue; and Derrick Lehmer, who informed me of the paper by Pólya.