The Distribution of Primes

The Sieve of Eratosthenes

The following table shows the integers from 1 to 30, with all the numbers divisible by 2, 3 and 5 sieved out.

1	2	3	4	5	6
7	8	9	10	11	12
13	14	15	16	17	18
19	20	21	22	23	24
25	26	27	28	29	30

It will be noted that this process singles out 8 numbers which include all the primes in the region with the exception of the fundamental primes 2, 3 & 5. Also, for these purposes, 1 is counted as a pseudo-prime because it is not divisible by 2, 3 or 5.

The table for the next 30 integers looks like this:

31	32	33	34	35	36
37	38	39	40	41	42
43	44	45	46	47	48
49	50	51	52	53	54
55	56	57	58	59	60

This time, the process picks out all the primes correctly (with the exception of 49 because this is divisible by the next prime up). In fact it is clear that in any block of 30 integers, there will be exactly 8 numbers which are not divisible by either 2, 3 or 5. Why is this?

Well, sieving out the numbers divisible by 2 halves the numbers in the table leaving 15. Sieving out the 3's reduces the number by a third leaving 10 and sieving out the 5's reduces it again by a fifth leaving 8.

We can summarise this process as follows:

$$N = 30 \times \frac{1}{2} \times \frac{2}{3} \times \frac{4}{5} = 8$$

Alternatively we can say that the *density* of integers which are not divisible by either 2, 3 or 5 is equal to:

$$\sigma_{2,3,5} = \frac{1}{2} \times \frac{2}{3} \times \frac{4}{5} = \frac{8}{30}$$

In general we can say that, in any block of $p_1 . p_2 \ldots p_n$ integers where $p_1 , p_2 \ldots p_n$ are primes there will be exactly

$$p_1 p_2 \ldots p_n \times \frac{(p_1-1)}{p_1} \times \frac{(p_2-1)}{p_2} \times \ldots \times \frac{(p_n-1)}{p_n} = (p_1-1) \times (p_2-1) \times \ldots \times (p_n-1)$$

numbers which are not divisible by any of the listed primes. (Note that the block does not have to

start and finish in any particular place because the pattern of the seive in every block is identical.)

Lets see what happens if the list includes the next prime up – 7. The size of the block needed is 210 and our theorem predicts that in any block of 210 numbers there will be

$$1 \times 2 \times 4 \times 6 = 48$$

numbers which are not divisible by 2, 3, 5 or 7.

Now let us compare this figure with the number of real primes in the respective blocks.

Between 1 and 210 there are 46 real primes. The discrepancy is due to three factors:

> Firstly, the numbers 2, 3 5 and 7 are not counted as prime because they are (of course) divisible by themselves. This brings our predited count of real primes up to 52
>
> Secondly, the numbers 121, 143, 187, 209 and 169 are counted as prime when in fact they are not (being 11^2, 11×13, 11×17, 11×19 and 13^2) (bringing the total of real primes to 47).
>
> Thirdly, the number 1 is counted as prime (so the predicted count is reduced to 46).

Between 211 and 420 there are 35 primes.

The discrepancy this time is due to the fact that there are 13 composite numbers between 211 and 420 but which are not divisible by 2, 3, 5 or 7 namely:

$11 \times 23 = 253$	$13 \times 17 = 221$	$17 \times 17 = 289$	$19 \times 19 = 361$
$11 \times 29 = 319$	$13 \times 19 = 247$	$17 \times 19 = 323$	
$11 \times 31 = 341$	$13 \times 23 = 299$	$17 \times 23 = 391$	
$11 \times 37 = 407$	$13 \times 29 = 377$		
	$13 \times 31 = 403$		

This time, since the numbers 1, 2, 3 5 and 7 are not included in the block, all we have to do is subtract 13 leaving 48 – 13 = 35

Of course, if we want to estimate the number of primes in a certain block of integers, we really should include all the primes up to the square root of the largest integer. For example, if we wish to estimate the number of primes in the first 400 integers, then we should take into account all the primes less than 20.

This gives us the following estimate:

$$N \approx 400 \times \frac{1}{2} \times \frac{2}{3} \times \frac{4}{5} \times \frac{6}{7} \times \frac{10}{11} \times \frac{12}{13} \times \frac{16}{17} = 72$$

to which we must add the number of fundamental primes used (7) and subtract 1 giving uas a best estimate of the number of primes between 1 and 400 of 78.

This can only be an estimate because our formula is only exact for block sizes equal to the product of the fundamental primes used which in this case is 510,510.

So how dooes our estimate measure up?

The number of real primes between 1 and 400 is exactly 78!

What about the block 401 to 800? This time we must throw in the primes 19 and 23 as well.

$$N \approx 400 \times \frac{1}{2} \times \frac{2}{3} \times \frac{4}{5} \times \frac{6}{7} \times \frac{10}{11} \times \frac{12}{13} \times \frac{16}{17} \times \frac{18}{19} \times \frac{22}{23} = 65$$

(there are no corrections to be made this time because the block starts above the largest prime used.)

Now the number of primes between 401 and 800 is actually only 61 – a significant discrepancy

from our estimate. The figures for subsequent blocks are as follows:

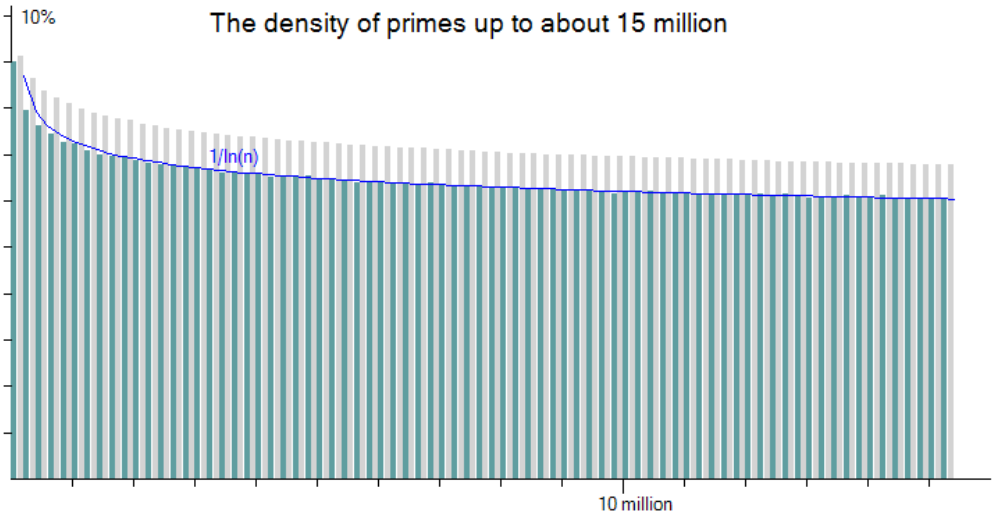
	Highest prime	Estimate	Actual
1 - 400	17	78	78
401 - 800	23	65	61
801 - 1200	31	61	57
1201 - 1600	39	59	55
1601 - 2000	43	57	52
2001 - 2400	47	55	54
2401 - 2800	47	55	50
2801 - 3200	53	54	45
3201 - 3600	59	54	51
3601 - 4000	61	53	47

The first thing to notice is that the numbers in the last column fluctuate randomly. This is because we are only looking at a small part of the complete cyclic block so the number of primes in a block of a certain size will partrly depend on exactly where the block starts and finishes. The estimate will always decrease monotonically. as more primes are factored in.

More importantly, the actual number of primes in a block always appears to be *less* than the estimated number. This is more puzzling because our previous analysis seemed to suggest that in any cyclic block, there would always be exactly the same number of primes. This analysis is false for the following reason. Our theorem states that, in any block of size $p_1 . p_2 ... p_n$ there will always be the same number of numbers which are not divisible by $p_1 , p_2 ... p_n$. But this is not the same thing as saying that the blocks will contain the same number of primes because blocks which contain larger numbers obviously have more potential primes by which they can be divided. In fact the only block size which could be said to work is the one we started with, namely $2 \times 3 \times 5 = 30$ because the highest prime which is less than the square root of 30 is 5. As we have see, the $2 \times 3 \times 5 \times 7 = 210$ block has to take into account the primes 11, 13, 17 and 19 as well.

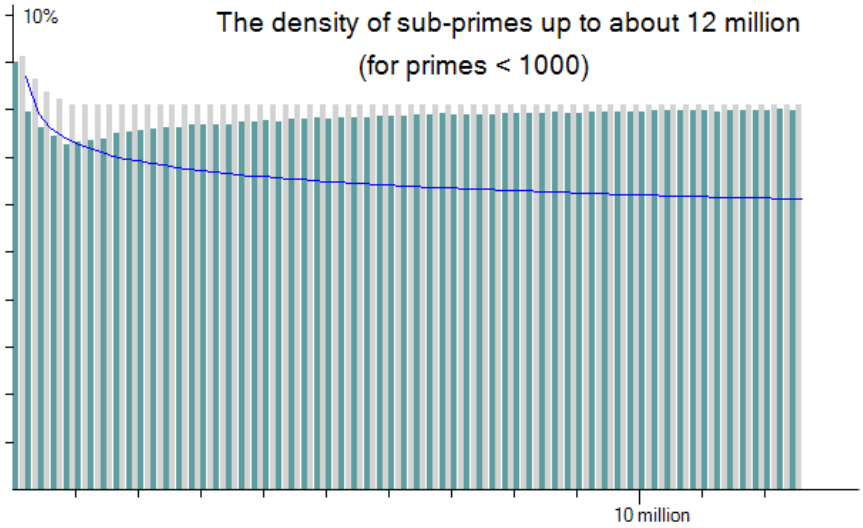
But why is the number of actual primes always *less* than the estimated number? Surely there ought to be an equivalent number of blocks which contain more than the estimated number of primes?

The following graph shows the estimated density of primes (in grey) and the actual density of primes (in blue) for 75 blocks of 200,000 integers from 1 to about 15 million.



It is, I think, extremely significant that the shape of the estimated density closely follows the actual density but in every case the actual density is significantly less.

Let us try a little experiment. Instead of using all the primes up to the square root of 15 million, let us use just the primes up to the square root of 1 million (ie primes < 1000) and just count as 'primes,' only those numbers which are not divisible by this collection of primes.



As you would expect, for all blocks > 1 million, the expected density of sub-primes is constant (because we are only dealing with all the primes up to 1000)

As I think you would expect too, the actual density of subprimes in the largest blocks closely approaches the expected density. Indeed, when we reach numbers of the order of the product of all the primes from 1 to 1000, I would expect the two graphs to be identical.

The question then remains – why is it that, when the numbers involved are of the order of the square of the highest prime (1,000,000 in this case), the number of subprimes is consistently *less* than the estimate?

I can only conjecture the following explanation. The formula which I have been using for the density of primes in the region of *n* is

$$P_n = \prod (p_i - 1)/p_i$$

where the product is taken over all the primes up to √n.

Essentially what we are assuming here is that the *probability* of zapping a given number with a prime such as 7 is 1:7 and this is totally independent of the probability of the number being zapped with any other prime. It is clear that when the numbers are small (ie < the product of all the primes being considered) this assumption must be false. And the reason for this connection is that *all the primes zap the number zero*. If we constructed the Sieve of Eratosthenes by striking out, not all those numbers whose modulus with respect to a certain prime *p* was 0 but instead some random number less than *p*, then I suspect that we would find that the number of 'primes' closely followed the prediction because there would be no correlation with the different primes at any number.

If this is correct, we can now see why the acrual number of primes in any block is always less than the expected number. The reason is that any number you choose (>48) is always far smaller than than the product of all its potential primes factors. So every block is always in the region where there are significant correlations between the prime factors.

Crook March 2013