5. DISTRIBUTION OF PRIMES

§5.1. Infinitely Many Primes

So far we've been dealing with primes individually. In this chapter we turn our attention to the set of all primes and the way they're distributed. Primes seem to become scarcer as we go out among larger and larger numbers. It's not inconceivable that they could run out altogether. True, there are infinitely many numbers altogether, and every number is a product of primes, but even with just one prime, such as 2, we can produce infinitely many powers of 2.

The fact that there's no last integer is easily proved. If N was the largest integer then we'd get a contradiction when we considered N + 1. However, such a simple argument won't work with primes. If N is the largest prime then N + 1 won't be a prime. Even N + 2

may not be a prime. There's no easy way of getting the next prime and so create a contradiction.

The following proof that there are infinitely many primes is due to Euclid.

Theorem 1 (EUCLID): There are infinitely many primes.

Proof: Suppose there is a largest prime N. Now take N! = N(N − 1)(N − 2) … 3.2.1. Every prime will therefore divide N! because every prime will appear as one of its factors. Now take N! + 1. No prime number will divide it because they all divide N! and no number bigger than 1 can divide two successive numbers. But every number bigger than 1 is divisible by a prime number, so we get a contradiction. Hence there are infinitely many prime numbers.

§5.2. A Formula For Primes

Since there are infinitely many primes we're justified in denoting the *n*'th prime by p_n, confident that p_n exists for all *n*. The next thing we might want to do is to find a formula for p_n. Such a formula has remained elusive despite centuries spent in trying to find one.

There are artificial formulae for the *n*'th prime that are totally useless in practice. A very simple minded formula is $p_n = \sqrt{q_n}$ where q_n is the *n*'th prime squared. There are formulae that are much more sophisticated than this but they are no more useful.

A less ambitious goal would be to find a formula for the next prime after a given one. This would be a recurrence equation of the form $p_{n+1} = F(p_n)$. Again, no satisfactory recurrence equation has been found. There appears to be a randomness about the primes.

A less ambitious goal is to find a simple function, such as a polynomial, all of whose values are primes. A remarkable possibility is $n^2 + n + 41$. This is a good example to show how mathematics differs from science.

Imagine a fantasy where a team of scientists in America test the hypothesis that $n^2 + n + 41$ only gives prime values by trying $n = 0, 1, 2, 3, ..., 9$. We'll pretend that testing for each n takes a lot of work and costs a lot of money, so stopping after just 10 values can be justified. The values they would obtain are:

41, 43, 47, 53, 61,71, 83, 97, 113, 131.

Some primes, such as 59, get missed but all these values are prime.

So they publish their results and claim that the hypothesis is true. "The values of $n^2 + n + 41$ are always prime." But, as generally happens in the scientific community, other scientists would repeat the experiment. Suppose a Russian team tried $n = 10, 11, 12, ..., 19$. Again the values are prime: 151, 173, 197, 223, 251, 281, 313, 347, 383, 421, 461.

A third group of scientists from Finland decides to test $n = 20, 21, 22, ..., 29$. They publish their results. The values are: 461, 503, 547, 593, 641, 691, 743, 797, 853, 911, 971. Yes, all are prime.

By this stage the scientific community has agreed that $n^2 + n + 41$ will always be prime and the 'fact' begins to be published in school textbooks.

Continuing this fantasy we'll suppose that, as an exercise in experimental number theory, various research students test for $n = 30, 31, 32, ..., 39$. Nobody doubts the outcome but it's good training in scientific research. Of course the values will all be prime:

971, 1033, 1097, 1163, 1231, 1301, 1373, 1447, 1523, 1601.

Then a young high school student points out that when $n = 41$, $n^2 + n + 41$ will, of course, be divisible by 41. In fact, when $n = 40$, $n^2 + n + 41 = 40^2 + 40 + 41 = 40.41 + 41$ and it is also divisible by 41.

Such a situation has probably never happened in real science but several astronomical discoveries have been made by young amateurs that were not noticed by professional astronomers.

Now, in fairness, the only method open to scientists is the experimental method. They make a hypothesis and they carry out experiments to test it. They can't *prove* things in the way mathematicians can. (It's true that theoretical physicists can use mathematics to draw consequences from certain laws of physics, but ultimately those laws rest upon experiment.)

Mathematical induction might have proved that the hypothesis is true always. If so, there would be no risk of it ever being overturned. But clearly no such proof could ever exist, because our hypothetical schoolgirl has found a counter-example.

Are 40 and 41 just isolated exceptions? Perhaps $n^2 + n + 41$ is always prime except when n is a multiple of 41 or one less than a multiple of 41. Let's continue.

For $n = 42, 43, ..., 80$ the values of $n^2 + n + 41$ are prime except for:

n	**$n^2 + n + 41$**	**factors**
44	2021	43×47
49	2491	47×53
56	3233	53×61
65	4331	61×71
76	5893	71×83

So even above 41 there are not that many composite values of $n^2 + n + 41$. What causes this remarkable phenomenon is that the discriminant of $n^2 + n + 41$ is -163 and -163 is not a square modulo any prime up to 37. It follows that all the prime divisors of $n^2 + n + 41$ will be at least 41.

§5.3. Gaps Between Primes

One of the most elusive aspects of number theory is the way the primes are distributed. If you examine a list of primes you'll observe that they become rarer as

the numbers get larger, but apart from that there doesn't seem to be any obvious pattern. The square numbers also become rarer as the numbers get larger yet in this case the pattern is quite predictable. To get the next square after the *n*'th you add $2n + 1$. If you work out the size of the gaps between successive squares you get a clear pattern.

n	1	2	3	4	5	6	7	8
n'th square	1	4	9	16	25	36	49	64
gap		3	5	7	9	11	13	15

n	9	10	11	12	13	14	15
n'th square	81	100	121	144	169	196	225
gap	17	19	21	23	25	27	29

If you do the same for primes you get considerable irregularity.

n	1	2	3	4	5	6	7	8
n'th prime	2	3	5	7	11	13	17	19
gap		1	2	2	4	2	4	2

n	9	10	11	12	13	14	15
n'th prime	23	29	31	37	41	43	47
gap	4	6	2	6	4	2	4

You'll observe that the gaps are all even, except between 2 and 3. This, of course, is because all primes after 2 are odd. You might be tempted to believe that the gaps, apart from the first, are only ever 2, 4 or 6 but when we get to the prime 89 the next prime is 97 and so the gap is 8. In fact the gaps can be arbitrarily large because for all n, none of the numbers

$$n! + 2, n! + 3, n! + 4, ..., n! + n$$

can be prime. This is because, if $m \leq n$, the number $n! + m$ will be divisible by m. So if p_k is the prime just before $n! + 2$ the next prime, $p_{k+1} \geq n! + n + 1$ and so the gap will be at least n.

You'll notice in the above table that a gap of 2 is quite common. Two primes that differ by 2 are called a **prime pair**. These become rarer as the numbers get larger, but there's no reason to believe that they stop altogether. A very famous hypothesis in number theory is the **Twin Prime Hypothesis**, stating that there are infinitely many twin primes. We don't know if this is indeed true – it is one of the many unsolved problems in number theory.

Example 1: The twin primes up to 133 are:

3, 5	5, 7	11, 13	17, 19	29, 31	41, 43	59, 61	71, 73	97, 101	101, 103	107, 109	131,133

Number theory is quite unique, among other branches of mathematics, in the number of unsolved

conjectures where the statement is quite elementary. Another such conjecture is the **Goldbach Conjecture** which states that every even number greater than 2 is the sum of two primes.

Example 2: The Goldbach conjecture verified for numbers up to 90.

	$4 = 2 + 2$	$6 = 3 + 3$	$8 = 3 + 5$	$10 = 3 + 7$
$12 = 5 + 7$	$14 = 3 + 11$	$16 = 3 + 13$	$18 = 5 + 13$	$20 = 3 + 17$
$22 = 3 + 19$	$24 = 5 + 19$	$26 = 3 + 23$	$28 = 5 + 23$	$30 = 7 + 23$
$32 = 3 + 29$	$34 = 3 + 31$	$36 = 5 + 31$	$38 = 7 + 31$	$40 = 3 + 37$
$42 = 5 + 37$	$44 = 3 + 41$	$46 = 3 + 43$	$48 = 5 + 43$	$50 = 3 + 47$
$52 = 5 + 47$	$54 = 7 + 47$	$56 = 3 + 53$	$58 = 5 + 53$	$60 = 7 + 53$
$62 = 3 + 59$	$64 = 3 + 61$	$66 = 5 + 61$	$68 = 7 + 61$	$70 = 3 + 67$
$72 = 5 + 67$	$74 = 3 + 71$	$76 = 3 + 73$	$78 = 5 + 73$	$80 = 7 + 73$
$82 = 3 + 79$	$84 = 5 + 79$	$86 = 3 + 83$	$88 = 5 + 83$	$90 = 7 + 83$

Indeed it is true up to 4.000.000,000,000,000,000. The scientific method would say that it’s obvious that it’s a

‘law of nature’, but in mathematics we can’t just leave it at that.

Many even numbers have several representations as the sum of two primes. For example 22 is also equal to 5 + 17 and, of course, it is 11 + 11. It might seem that we never need to go beyond 7 for the smaller prime in such a representation but watch when we continue up to 100.

92 = 3 + 89	94 = 5 + 89	96 = 7 + 89	98 = 19 + 79	100 = 3 + 97

But even so, we can also write 98 as 31 + 67 and 37 + 61. It doesn’t seem to be such a special thing to be able to write an even number as a sum of two primes, and it is probably always possible. But it’s amazing that its proof has eluded number theorists for so long.

§5.4. Primes in Arithmetic Sequences

All primes, after 2, are odd. Therefore all odd primes have the form $4n + 1$ or $4n + 3$. As there are infinitely many odd primes altogether it might be surprising if there were only finitely many of one type and infinitely many of the other, but it isn’t obvious that this is so. The following proof is a variation on Euclid’s one.

Theorem 2: There are infinitely many primes of the form $4n + 3$.

Proof: Suppose there are only finitely many primes of the form $4n + 3$ and let P be the largest of them. Let N = 2.3.5. ... .P, be the product of all the primes up to an including P and let $M = 2N - 1$. Clearly $M \equiv 3 \pmod 4$.

Now for all integers a, b, if a and b are congruent to 1 modulo 4 then $ab \equiv 1 \pmod 4$.

It follows that not all the prime divisors of M can be congruent to 1 modulo 4, otherwise M would be also.

So M must have a prime divisor, p, of the form $4n + 3$. But clearly we can't have $p \leq P$ because then $p \mid N$ and so $M \equiv -1 \pmod p$.

Hence $p > P$, contradicting the assumption that P is the largest prime of the form $4n + 3$.

Theorem 3: There are infinitely many primes of the form $8n + 5$.

Proof: Suppose there are only finitely many primes of the form $8n + 5$ and let P be the largest.

Let N = 3.5. ... .P, be the product of all the odd primes up to an including P and let $M = N^2 + 4$. Clearly $N \equiv 1 \pmod 2$ and so $N^2 \equiv 1 \pmod 8$ and hence $M \equiv 5 \pmod 4$.

Let p be an odd prime divisor of M. Now it can be shown that an odd prime divisor of a sum of two coprime squares, such as M, must have the form $4n + 1$, that is of the form $8m + 1$ or $8m + 5$.

If all the odd prime divisors of M were congruent to 1 mod 8 then so would $M \equiv 1 (\text{mod } 8)$.
It follows that at least one of the odd prime divisors, p, of M must be of the form $8n + 5$.
But $p > P$, a contradiction.

These results are special cases of a very deep theorem due to Dirichlet.

Theorem 4 (DITICHLET): If a, b are coprime integers, with $a > 0$, then there are infinitely many primes of the form $an + b$.
Proof: This is a very deep theorem and the proof is omitted.

§5.5. Number of Primes up to *n*

In this, and the following section, $\sum_{p \le n} F(p)$ and $\sum_{a \le p \le n} F(p)$ denote the sums over all *primes* satisfying the inequality and similarly for products.

Let $\boldsymbol{\pi(n)}$ be the number of primes that are less than or equal to n. The fact that there are infinitely many primes means that $\pi(n) \to \infty$ as $n \to \infty$. Can we find a simple upper bound for $\pi(n)$ in terms of n? Well, of

course $\pi(n) < n$, but can we do a lot better than that? I'll answer this later. A related functions is $\boldsymbol{\rho(n)} = \sum_{p \le n} \log p$.

Example 3: $\pi(16) = 6$, $\rho(16) = \log 2 + \log 3 + \log 5 + \log 7 + \log 11 + \log 13 \approx 10.31$.

Theorem 5: For all $n \ge 7$, $\pi(n) < \rho(n)$.
Proof: $\pi(7) = 4$ and $\rho(7) = \log 2 + \log 3 + \log 5 + \log 7 \approx 5.35$.
For larger n, each extra prime, p, contributes 1 to $\pi(n)$ and log p to $\rho(n)$ and since $\log p > 1$ for all $p > 7$ the inequality continues.

We can find an upper bound for $\rho(n)$ as a means of finding an upper bound for $\pi(n)$.

Theorem 6: For all m, $\rho(2m + 1) < \rho(m) + 2m \log 2$.
Proof: Let $M = \binom{2m+1}{m} = \frac{(2m+1)(2m) \ldots (m+1)}{m!}$.
Since M occurs twice in the binomial expansion of $(1 + 1)^{2m+1}$ we conclude that $2M < 2^{2m+1}$.
Hence $M < 2^{2m}$ and so $\log M < 2m \log 2$.
Let $P = \prod_{m+1 < p \le 2m+1} p$ be the product of all primes p with $m + 1 < p \le 2m + 1$.
Every one of these primes divides M because it divides the numerator but not the denominator.
Hence P divides M and so $P \le M$.

Now $\rho(2m + 1) - \rho(m) = \log\left(\prod_{m+1< p\leq 2m+1} p\right) = \log P \leq \log M < 2m \log 2$.

Note that we needed to use 2m + 1 so that we'd have two largest binomial coefficients in the expansion of $(1 + 1)^{2m+1}$, whereas the expansion of $(1 + 1)^{2m}$ only has one.

Theorem 7: $\rho(n) < 2n\log 2$ for all n.

Proof: Let Suppose N is a minimal counter-example.

If N is even $\rho(N) = \rho(N - 1)$
$< 2(N - 1)\log 2$
$< 2N \log 2$.

If N is odd, say $N = 2m + 1$ then

$\rho(N) = \rho(2m + 1) < \rho(m + 1) + 2m \log 2$
$< 2(m + 1)\log 2 + 2m \log 2$
$= 2(2m + 1) \log 2$
$= 2N \log 2$.

Hence the theorem holds for N, and so is not a counter-example.

So the theorem holds for all n.

Corollary: $\pi(n) < 2n\log 2$ for all n.

Proof: For $n \geq 7$, $\pi(n) < \rho(n)$. A simple check shows that it is true for $n = 1, 2, 3, 4, 5, 6$.

This upper bound is not particularly tight. For example, if $n = 10^9$, $\pi(n)$ is about 51 million while $2n\log 2$ is about 1386 million.

§5.6. Bertand's Postulate

The infinitude of primes guarantees that, given a prime there exists a prime beyond it. But we don't seem to be able to predict how much further it is to the next prime. Can we, at least, place an upper bound on the gap $p_{n+1} - p_n$? Bertrand's Postulate gives a positive answer to this question.

It is known as Bertrand's Postulate rather than Bertrand's Theorem because Joseph Bertrand [1822-1900] postulated it in 1845 but it was proved in 1850 by Chebyshev [1821-1894]. The following is based on the proof by Erdös [1913-1996].

It's presented on pp, 343-344 of the third edition of *An Introduction To The Theory of Numbers* by G.H. Hardy and E.M. Wright.

[**NOTE:** There is an error at the top of page 344 where it is asserted, in effect, that
$(4/3)n \log 2 \le 2(1 + \sqrt{2n})\log(2n)$.]

Theorem 8 (BERTRAND-CHEBYSHEV): For all $n \ge 2$ there exists a prime p such that

$$n < p < 2n.$$

Proof: We prove the theorem first for $n > 512$.
Suppose that there is no prime p with $512 < n < p < 2n$.
Let $N = \binom{2n}{n} = \frac{(2n)!}{(n!)^2}$.

(1) Let p be a prime divisor of N. By our assumption $p \le n$.

Suppose that $p > \frac{2}{3}n$. Then $2p \le 2n < 3p$.

Hence $p^2 > \frac{4}{9}n^2 > 2n$ since $2n > 9$.

Now the number of multiples of p in $\{1, 2, ..., n\}$ is the same as the number of multiples of p in
$\{n + 1, n + 2, ..., 2n\}$ and so the factors of p in the factorials in N cancel completely, contradicting the fact that $p \mid$ N.

Hence the prime dividers of N are all at most $\frac{2}{3}n$.

(2) Now $\sum_{p|N} \log p \le \sum_{p \le (2/3)n} \log p = \rho\left(\frac{2}{3}n\right) < \frac{4}{3}n.\log 2.$

(3) Let k_p be the number of factors of p that divide N.
If $k_p \ge 2$ then $p^2 \le 2n$ and so $p \le \sqrt{2n}$.

(4) N is the largest term of $(1 + 1)^{2n}$ so $2^{2n} \le 2nN$.
[If we group the first and last terms together there are $2n$ terms in the expansion of $(1 + 1)^{2n}$ and each is less than or equal to N.]

(5) Now $N = \prod_{p|N} p^{k_p}$ so

$$\log N = \sum_{p|N} k_p \log p$$

$$= \sum_{k_p=1} \log p + \sum_{k_p \geq 2} \log p^{k_p}$$

$$\leq \sum_{p|N} \log p + \sqrt{2n} \log(2n)$$ since there are at most $\sqrt{2n}$ primes with $k_p \geq 2$ and each $p^{k_p} \leq 2n$.

$$\leq \frac{4n}{3} \log 2 + \sqrt{2n} \log(2n).$$

But from (4), $2n \log 2 \leq \log(2n) + \log N$

$$\leq \log(2n) + \frac{4n}{3} \log 2 + \sqrt{2n} \log(2n)$$

$$= \frac{4n}{3} \log 2 + (1 + \sqrt{2n}) \log(2n).$$

Hence $(2n)\log 2 \leq 3(1 + \sqrt{2n})\log(2n)$.

(6) Let $\varepsilon = \frac{\log 2n}{\log 4} \geq \frac{\log 1024}{\log 4} = 5$.

Hence $2\varepsilon \log 2 = \log 2n$ and so $2n = 2^{2\varepsilon}$.

(7) Substituting in the inequality in (5) we get:
$2^{2\varepsilon} \log 2 \leq 3(1 + 2^{\varepsilon})\log(2^{2\varepsilon}) = 6\varepsilon(1 + 2^{\varepsilon})\log 2$.
Therefore $2^{2\varepsilon} \leq 6\varepsilon(1 + 2^{\varepsilon})$ and hence $2^{\varepsilon} \leq 6\varepsilon(1 + 2^{-\varepsilon})$.
Since $\varepsilon \geq 5$, $6(1 + 2^{-\varepsilon}) \leq 6(1 + 2^{-5}) = \frac{193}{32}$.

So $2^{\varepsilon} \leq \frac{193}{32} \varepsilon$.

The graphs of $y = 2^x$ and $y = \frac{193}{32}x$ intersect just once, and this is at a value between $x = 4.8790$ and $x = 4.8791$. Since $\varepsilon \geq 5$ we get $2^\varepsilon > \frac{193}{32}\varepsilon$ and hence a contradiction.

It remains to check that the theorem hold for $n < 512$. This is easily done by reference to a table of prime numbers.

A consequence of Bertand's Postulate is that there are infinitely many primes, though if that's all we wanted then Euclid's proof would be simpler! But another consequence is that it provides a lower bound on $\pi(n)$.

Theorem 9: For all $m \geq 3$, $\pi(2^m) \geq m + 1$.
Proof: Suppose M is a minimal counter-example.
Since $\pi(8) = 4$, $M > 3$.
Then $\pi(2^M) \geq \pi(2^{M-1}) + 1$ by Bertrand's Postulate.
$\geq M + 1$, contradicting the definition of M.
Hence $\pi(2^m) \geq m + 1$ for all $m \geq 3$.

Theorem 10: $\frac{\log n}{\log 2} \leq \pi(n) < 2n\log 2$ for all n.

Proof: Suppose that $2^m \leq n < 2^{m+1}$.
Then $\log n < (m + 1)\log 2$.
Hence $\pi(n) \geq \pi(2^m) \geq m + 1 > \frac{\log n}{\log 2}$.

When $n = 10^9$, $\frac{\log n}{\log 2} \approx 2624.918$, while $\pi(n)$ is about 51 million so this is a pretty weak lower bound!

There have been many improvements on Bertrand's Postulate. In 2006 it was shown that for $n \geq 2$ there's always a prime between $2n$ and $3n$ and in 2011 it was shown that for $n \geq 2$ there's always a prime between $3n$ and $4n$.

It has been postulated that there is always a prime between n^2 and $(n + 1)^2$. This seems likely, and indeed if it was a scientific hypothesis it would be considered fact. But it has not yet been proved, or disproved.