Primes: What is Known and Unknown

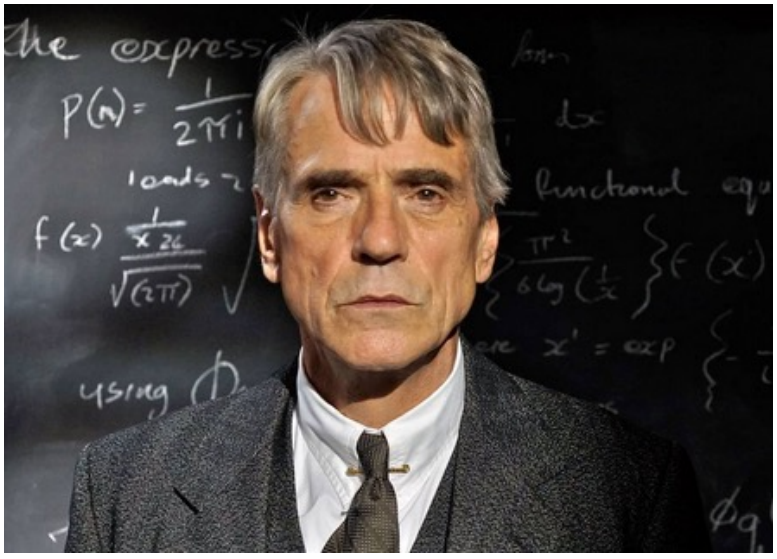
Keith Conrad
Theory Day

March 30, 2017

Outline

- How to use primes?
- How to find primes?
- What do we know about primes?
- What don't we know about primes?

Using primes in the real world



The theory of numbers has always been regarded as one of the most obviously useless branches of pure mathematics. The accusation [is] never more just than when directed against the parts [...] *concerned with primes.* G. H. Hardy (1915)

Very large primes are essential for cryptography, used indirectly every day by people and banks (really computers, phones, etc.).

1. The RSA cryptosystem
2. Digital signature algorithms

Using primes in the real world

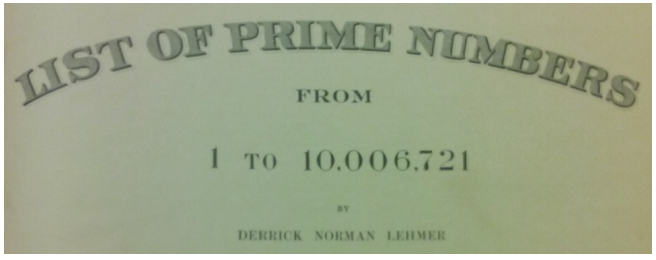
The theory of numbers has always been regarded as one of the most obviously useless branches of pure mathematics. The accusation [is] never more just than when directed against the parts [...] *concerned with primes.* G. H. Hardy (1915)

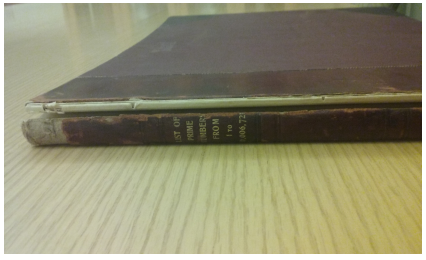
Very large primes are essential for cryptography, used indirectly every day by people and banks (really computers, phones, etc.).

1. The RSA cryptosystem
2. Digital signature algorithms

How to find primes 100 years ago: look in a book

Below is an excerpt from D. N. Lehmer's 1914 book.





1	541	1223	1987
2	47	29	93
3	57	31	97
5	63	37	99
7	69	49	2003
11	71	59	11
13	77	77	17
17	87	79	27
19	93	83	29
23	99	89	39

Why don't we regard 1 as prime anymore?

- Unique prime factorization would fail: $12 = 2^2 \cdot 3 = 1^5 \cdot 2^2 \cdot 3$
- In abstract algebra, prime numbers generalize to prime ideals while 1 generalizes to units: different concepts.

How to find primes of a special form: Mersenne primes

A *Mersenne prime* is a prime of the form $2^p - 1$, where p is prime. (If $2^n - 1$ is prime then n must be prime.)

Examples: $2^2 - 1 = 3$, $2^3 - 1 = 7$, $2^5 - 1 = 31$, $2^7 - 1 = 127$. (Note $2^{11} - 1 = 2047 = 23 \cdot 89$.)

There is a special-purpose test to check primality of Mersenne numbers $2^p - 1$, called the *Lucas–Lehmer test*.



There are 49 known Mersenne primes, latest being $2^{74,207,281} - 1$ (found in Jan. 2016) with over 22,000,000 digits. Largest known prime has nearly always been a Mersenne prime.

How to find general primes the elementary way: look for factors

Call a factor of n strictly between 1 and n a *division witness* for n: it is a "witness" to the compositeness of n. A prime does not have any division witness, so finding one implies n is composite.

Example. The number 3423701 has two division witnesses, 1801 and 1901.

Searching for division witnesses does not work quickly if factors are hard to find, and they don't even exist if the number is prime.

Example. The number 3423713 is prime. Trial division will verify primality only after testing all numbers up to $\sqrt{3423713} \approx 1850.3$. Finding no division witness up to 1000 *essentially tells us nothing*.

To check if $n > 1$ is composite we need to factor it, right? **Wrong!** Often it is easier to prove n factors than to find factors, by showing n lacks a property that all primes have. What is an example of such a property?

How to find general primes for applications: by probability

Fermat's little theorem: *For prime p and* **every** *a from 1 to $p-1$,*

$$a^{p-1} \equiv 1 \bmod p.$$

Here it is for $p = 5$:

$$1^4 \equiv 1 \bmod 5,\ 2^4 \equiv 1 \bmod 5,\ 3^4 \equiv 1 \bmod 5,\ 4^4 \equiv 1 \bmod 5.$$

This result can be *written* with general n in place of primes:

$$\boxed{1 \le a \le n-1 \stackrel{?}{\Longrightarrow} a^{n-1} \equiv 1 \bmod n}$$

Writing this doesn't make it true. Fails at some $a \Rightarrow n$ isn't prime.

Example. Let $n = 2047$. Then $2^{2046} \equiv 1 \bmod 2047$: so what? But $3^{2046} \equiv 1013 \bmod 2047$, which proves 2047 is composite without showing how to factor it. We call 3 a *Fermat witness* for 2047.

Theorem. *If $a^{n-1} \not\equiv 1 \bmod n$ for some a in $\{1, \ldots, n-1\}$ then n is composite and proportion of Fermat witnesses is <u>often</u> $> 50\%$.*

Example. If $n = 2047$ then about 76% of all a from 1 to 2046 are Fermat witnesses: $a^{2046} \not\equiv 1 \bmod 2047$.

How to find general primes for applications: by probability

Example. If $n = 11004252611041$ then 100 random choices of a from 1 to $n-1$ all satisfy $a^{n-1} \equiv 1 \bmod n$. Does that mean n is prime? **NO**: $n = 12241 \cdot 24481 \cdot 36721$.

The *Miller–Rabin test* is a refinement of the Fermat test. Steps in the test not described here, but they are all elementary. Introduced by Miller as a deterministic primality test needing a big hypothesis.

Theorem (Miller, 1976). *If Generalized Riemann hypothesis is true then there is a constant C such that primality of n is same as no $a \leq C(\log n)^2$ being a Miller–Rabin witness for compositeness of n.*

What's C? If GRH is true then we can use $C = 2$ (Bach, 1984).

Example. If $n = 3423713$ then $2(\log n)^2 \approx 452.77$.

Fact: odd composite $n < 10^{10}$ have 2, 3, 5, 7, or 11 as MR witnesses.

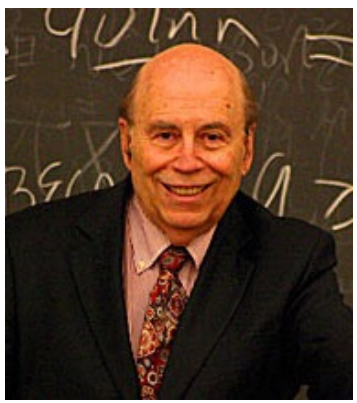
Theorem (Rabin, 1980). *For odd composite n, at least* **75%** *of a from 1 to $n-1$ are Miller–Rabin witnesses for compositeness of n.*

Rabin's theorem avoids GRH but has randomization.

Efficient proof of primality

In practice, Miller–Rabin test used as a probabilistic primality test. It would be a deterministic primality test running in polynomial time (bounded by a fixed power of $\log n$) if Generalized Riemann hypothesis is true.

First unconditional (not depending on any unproved conjectures) deterministic polynomial time primality test due to Agrawal, Kayal, and Saxena in 2002.


Other (unconditional) deterministic primality tests run faster than the AKS test in practice, but there is *no proof* yet that they do so on all inputs.

Counting primes

Euclid (*Elements, Book IX, Prop.* 20) proved there are infinitely many primes: $p_1 p_2 \cdots p_k + 1$ has prime factor not in $p_1, \ldots, p_k$.

Euler (1737) found another proof using analysis: divergence of harmonic series $\sum_{n \geq 1} \frac{1}{n}$ implies divergence of $\sum_p \frac{1}{p}$.

Dirichlet (1837) showed arithmetic progression $b, a+b, 2a+b, \ldots$ when $\gcd(a, b) = 1$ has infinitely many primes: $\sum_{p=an+b} \frac{1}{p} = \infty$.

Ex. $\{10n + 7\}$ has primes $7, 17, \cancel{27}, 37, 47, \cancel{57}, 67, \cancel{77}, \cancel{87}, 97, \ldots$

Dirichlet's proof used complex numbers and novel analytic ideas.

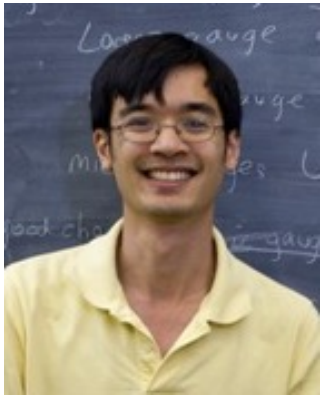
Dirichlet created a new branch of mathematics, which uses the infinite series introduced by Fourier in the theory of heat, to explore properties of prime numbers. Jacobi (1846)

ALL primes in arithmetic progression

Erdős conjectured for every set S of positive integers satisfying

$$\sum_{n \in S} \frac{1}{n} = \infty$$

contains arbitrarily long arithmetic progressions. This is still open in general. The set of primes fits the hypothesis, by Euler. That they fit the conclusion as well is the Green–Tao theorem (2004), whose proof uses ideas from ergodic theory.



Length of longest known arithmetic progression of primes: **26**.

Unsolved problems about counting primes

1. Are there infinitely many Mersenne primes $2^p - 1$?
2. Are there infinitely many twin primes?

$$(3,5),\quad (5,7),\quad (11,13),\quad (17,19),\quad (29,31),\quad (41,43),\quad \ldots$$

Could there be infinitely many "triple primes" $p, p+2, p+4$?
No besides 3, 5, 7: one of $n, n+2, n+4$ is a multiple of 3.

3. (Hardy–Littlewood, 1923) Are there infinitely many prime k-tuples $n+h_1, \ldots, n+h_k$ "unless there obviously aren't"?
Don't want all $\{n+h_1, \ldots, n+h_k\}$ to contain a multiple of a common prime. (Bad: $\{h_1, \ldots, h_k\} = \{0, 1, \ldots, p-1\}$ mod p for some prime $p \leq k$.)
Ex. $(k=3)$ $n, n+2, n+4$ bad: at least one is a multiple of 3.
Ex. $(k=3)$ $n, n+2, n+6$ good: doesn't always include an even number, doesn't always include a multiple of 3. Expect infinitely many such triples are all prime.

Progress on prime gaps

Theorem. (Zhang, 2013) *There are infinitely many prime pairs differing by at most* 70,000,000*.*

Work of Maynard, Tao, and others reduced prime pair gap bound infinitely often to at most 246 and got more:

Theorem. *For each* $k \geq 2$ *there is a* k*-tuple* $(h_1, \ldots, h_k)$ *such that* $n + h_1, \ldots, n + h_k$ *are all prime infinitely often.*

This is a pure existence result: for no $k \geq 2$ is a specific k-tuple $(h_1, \ldots, h_k)$ known that provably fits the conclusion.

Reciprocals of primes

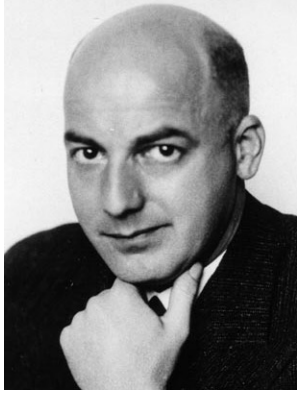
The number $\frac{1}{7} = .142857142857\ldots$ has decimal period 6. Let

$$E(x) = \frac{|\{p \leq x : 1/p \text{ has even decimal period}\}|}{|\{p \leq x\}|}.$$

How do you think $E(x)$ behaves as x grows?

x	10	10^2	10^3	10^4	10^5	10^6
$E(x)$	.25	.52	.6488	.6664	.6666	.6666

Theorem. (Hasse, 1965) *The "probability" $1/p$ has even decimal period is $2/3$: as $x \to \infty$, $E(x) \to 2/3$.*



Reciprocals of primes

Decimal period for $\frac{1}{d}$ is $\leq d-1$. Equality requires prime d.

Examples: $\frac{1}{7} = .\underbrace{\overline{142857}}_{6}$, $\frac{1}{17} = .\underbrace{\overline{0588235294117647}}_{16}$

Works for $d = 7, 17, 19, 23, 29, 47, 59, 61, \ldots, 2017, \ldots$

For $b \geq 2$, base-b period of $\frac{1}{d}$ is $\leq d-1$. Equality needs prime d.

Examples when $b = 2$: $\frac{1}{3} = .\overline{01}$, $\frac{1}{5} = .\underbrace{\overline{0011}}_{4}$, $\frac{1}{11} = .\underbrace{\overline{00101101}}_{10}$

Works if $b = 2$ for $d = 3, 5, 11, 13, 19, 29, 37, 53, \ldots, \cancel{2017}, \ldots$

Artin (1927) conjectured base-b period for $1/p$ is $p-1$ infinitely often whenever b is not a perfect square. What is known?

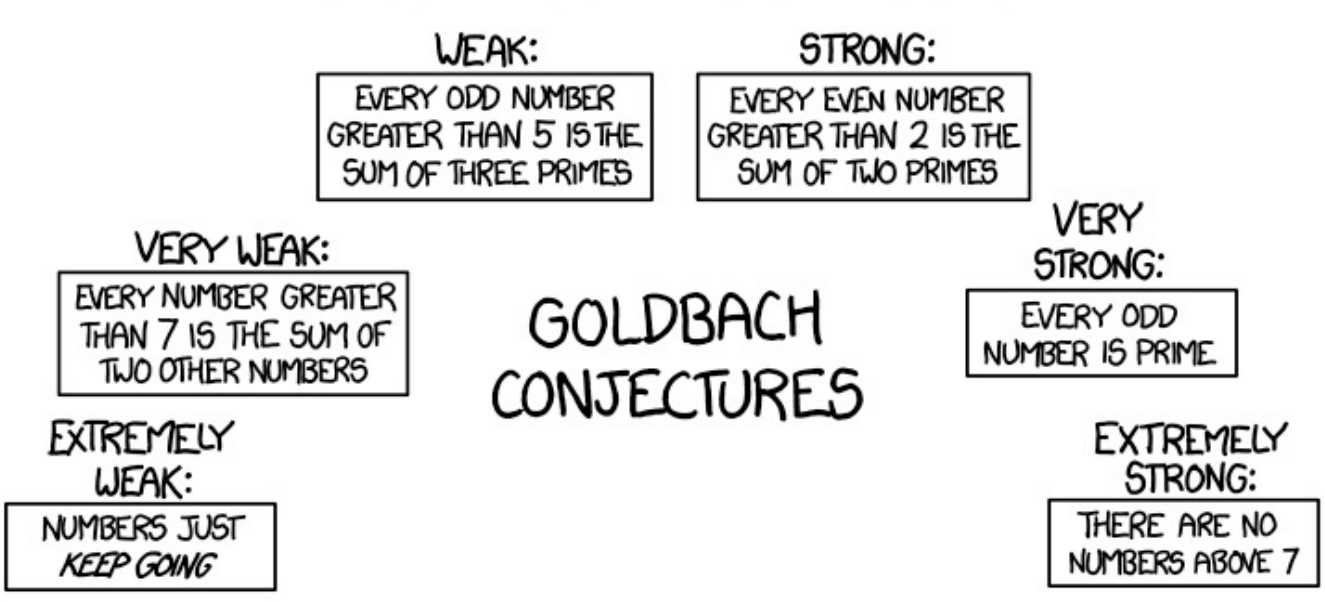
- (Hooley, 1967) Follows from Generalized Riemann hypothesis.
- (Heath-Brown, 1986) True for at least one of $b = 2, 3$, or 5.

Adding primes

Strong Goldbach conjecture: all even $n > 2$ a sum of two primes.
Weak Goldbach conjecture: all odd $n > 5$ a sum of three primes.
Since $n > 5 \implies n - 3 > 2$, Strong GC implies Weak GC.

1. 1923: Hardy and Littlewood showed Generalized Riemann hypothesis (GRH) implies Weak GC true for odd $n \gg 0$.
2. 1937: Vinogradov showed Weak GC true for odd $n \gg 0$.
3. 2013: Helfgott proved Weak GC unconditionally.

Obligatory xkcd:



Counting prime pairs

$$\begin{aligned} \pi(x) &= |\{p \le x : p \text{ prime}\}| \\ \pi_{\text{twin}}(x) &= |\{p \le x : p, p+2 \text{ prime}\}| \\ \pi_{T,T+6}(x) &= |\{p \le x : p, p+6 \text{ prime}\}|. \end{aligned}$$

Ex. $\pi(10) = |\{2,3,5,7\}| = 4$, $\pi_{\text{twin}}(10) = 2$, $\pi_{T,T+6}(10) = 2$.

x	10^4	10^5	10^6	10^7	10^8	10^9
$\pi_{\text{twin}}(x)$	205	1224	8169	58980	440312	3424506
$\pi_{T,T+6}(x)$	411	2447	16386	117207	879980	6849047

Expect these counts $\to \infty$, but no proof! Observations? Looks like $\boxed{\pi_{T,T+6}(x) \sim 2\pi_{\text{twin}}(x)}$.

Comparing $n, n+2$ with $n, n+6$ for $n \ge 1$, they share the same statistics for divisibility by 2, but **not** for divisibility by 3:

- $n, n+2$ not divisible by 3 exactly when $n = 3m+2$.
- $n, n+6$ not divisible by 3 exactly when $n = 3m+1$ or $3m+2$.

These pairs have same statistics for divisibility by each prime $\neq 3$.

Density of primes around large numbers

Mathematicians have tried in vain [...] *to discover some order in the sequence of primes, and we have reason to believe that it is a mystery the human mind will never penetrate.* Euler (1751)

One of my first projects was [...] *the decreasing frequency of primes, to which end I counted primes in several chiliads. I soon recognized that behind all of its fluctuations, this frequency is on average inversely proportional to the logarithm.* Gauss (1849)

Density of primes around large numbers

n	$\lvert\{\text{primes in } [n, n+999]\}\rvert/1000$	$1/\log n$
10K	.106	.1085
50K	.089	.0924
100K	.081	.0868
500K	.079	.0762
1 M	.075	.0723
1.5 M	.083	.0703
2 M	.069	.0689
2.5 M	.064	.0678
3 M	.062	.0670

This suggests the useful heuristic

$$\text{Prob}(n \text{ prime}) = \frac{1}{\log n}.$$

Strictly speaking, this is problematic:

1. $\frac{1}{\log 1} = \infty$, $\frac{1}{\log 2} > 1$, and $\sum_{n \geq 2} \frac{1}{\log n} = \infty$.
2. Being prime is *not* a probabilistic concept.

The basic probabilistic heuristic

Here is a more elementary heuristic:

$$\text{Prob}(n \text{ even}) = \frac{1}{2} \text{ and } \sum_{n \le x} \text{Prob}(n \text{ even}) \sim \frac{x}{2} \sim |\{\text{even } n \le x\}|. \checkmark$$

The **expected number** of primes up to x ("successes") should be a sum of probabilities over n up to x:

$$\pi(x) \stackrel{?}{\sim} \sum_{n \le x} \text{Prob}(n \text{ prime}) = \sum_{2 \le n \le x} \frac{1}{\log n} \stackrel{!}{\sim} \frac{x}{\log x},$$

and that **is** true: it's the Prime Number Theorem.

Theorem (Hadamard, de la Vallée-Poussin, 1896) *As* $x \to \infty$,

$$\boxed{\pi(x) \sim \frac{x}{\log x}.}$$

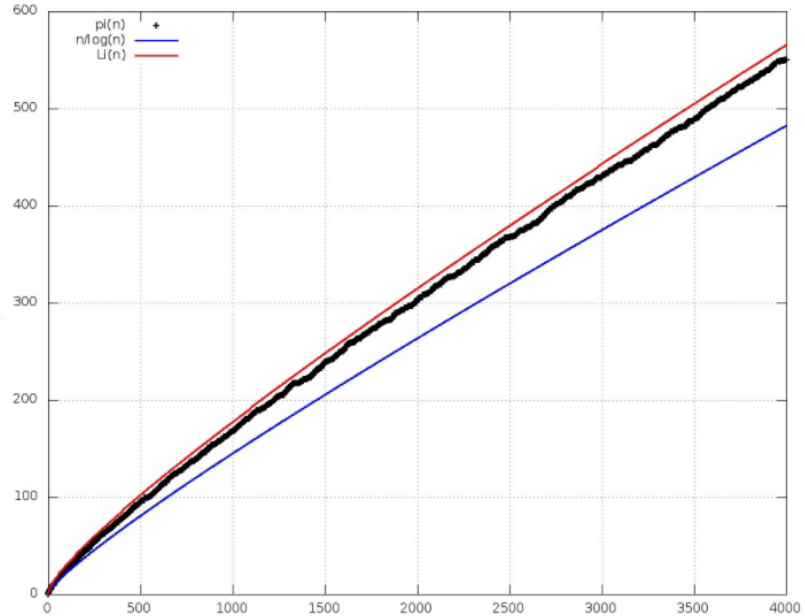
Data for Prime Number Theorem compared to sum of probabilities

While $\pi(x) \sim \dfrac{x}{\log x} \sim \displaystyle\sum_{2 \le n \le x} \dfrac{1}{\log n}$, the last formula is a **much better** asymptotic estimate for $\pi(x)$.

x	10^4	10^5	10^6	10^7	10^8
$\pi(x)$	1229	9592	78498	664579	5761455
$x/\log x$	1085	8685	72382	620420	5428681
Ratio	1.1319	1.1043	1.0844	1.0711	1.0612
$\pi(x)$	1229	9592	78498	664579	5761455
$\displaystyle\sum_{2 \le n \le x} \dfrac{1}{\log n}$	1245	9629	78627	664918	5762209
Ratio	.9863	.9960	.9983	.9994	.9998

For verification within the limits of calculation, [*the formula used*] *is by no means indifferent and it will be found that it makes a vital difference in the plausibility of the results.* Hardy & Littlewood

Comparing growth of $\pi(x)$ and approximations



Plot shows $\pi(x)$, $\dfrac{x}{\log x}$, and smoothed version of $\displaystyle\sum_{2\le n\le x} \frac{1}{\log n}$. Blue below black for $x \ge 17$. No known $x > 2$ where red below black.

- Littlewood (1914) proved red below black infinitely often as $x \to \infty$.
- Skewes (1933, 1955) showed red below black before a huge but explicit bound. Now known to occur before $\sim 10^{316}$ and not before 10^{19}.

Error terms in the Prime Number Theorem

Bounding the error $\left|\pi(x) - \sum_{2\le n\le x} \frac{1}{\log n}\right|$ is one way the famous Riemann hypothesis (1859) can be expressed. It says this error grows no faster than a constant multiple of $\sqrt{x}\log x$ as x grows:

$$\left|\pi(x) - \sum_{2\le n\le x} \frac{1}{\log n}\right| \overset{\text{RH}}{\le} C\sqrt{x}\log x.$$

The key issue on the right is the exponent on x: $\sqrt{x} = x^{1/2}$. Not true with smaller exponent, would be a breakthrough to get $x^{1-\varepsilon}$.

This is **not** how people usually think about RH: there are several equivalent formulations and more technical ones are more useful.

Work of Littlewood and Skewes each had two parts: first prove result if Riemann hypothesis is true and then prove the result if Riemann hypothesis is false. Thus theorem proved either way! (If RH is false, it is provably false.)

Refining the probabilistic heuristic

When counting primes of a special form, we should **take into account** divisibility properties of such numbers. This is like *conditional probability*.

Example. Prob($n \in \{1, \ldots, 100\}$ is a multiple of 4) $= 1/4$, but Prob($n \in \{1, \ldots, 100\}$ is a multiple of 4 **if** n is even) $= \frac{1/4}{1/2} = 1/2$.

Example. Pairs $n, n+2$ for $n \geq 1$ are

$$(1,3),\ (2,4),\ (3,5),\ (4,6),\ (5,7),\ (6,8),\ (7,9),\ (8,10),\ldots$$

1. A random pair (n, m) has both terms odd 25% of the time, but n and $n+2$ both odd 50% of the time (see above). This makes it *more likely* that $n, n+2$ both prime than n, m.
2. For prime $p > 2$, a random pair (n, m) has both terms not a multiple of p with probability $(1-1/p)^2$, but n and $n+2$ both not multiple of p with probability $1 - 2/p < (1-1/p)^2$. This makes it *less likely* that $n, n+2$ are both prime than n, m.

Twin prime heuristic

To quantify how often n and $n+2$ are both prime, our basic heuristic is that for each prime p,

$$\text{Prob}(n \text{ and } m \text{ not multiples of } p) = \left(1 - \frac{1}{p}\right)^2$$

because events "n, m not a multiple of p" are "independent," while

$$\text{Prob}(n \text{ and } n+2 \text{ not multiples of } p) = \begin{cases} 1/2 & \text{if } p = 2, \\ 1 - 2/p & \text{if } p > 2. \end{cases}$$

New heuristic: $\text{Prob}(n \text{ and } n+2 \text{ prime}) = \dfrac{C}{(\log n)(\log(n+2))}$

where

$$C = \frac{1/2}{(1-1/2)^2} \prod_{p>2} \frac{1-2/p}{(1-1/p)^2} = 2 \prod_{p>2} \frac{1-2/p}{(1-1/p)^2} \approx 1.320323.$$

Twin prime conjecture

Conjecture (Hardy–Littlewood, 1923). *As* $x \to \infty$,

$$\pi_{\text{twin}}(x) \stackrel{?}{\sim} \sum_{2 \le n \le x} \frac{C}{(\log n)(\log(n+2))} \sim C \frac{x}{(\log x)^2},$$

where

$$C = 2 \prod_{p>2} \frac{1 - 2/p}{(1 - 1/p)^2} \approx 1.320323.$$

In table below, "Approx." comes from the summation up to x.

x	10^4	10^5	10^6	10^7	10^8
$\pi_{\text{twin}}(x)$	205	1224	8169	58980	440312
Approx.	213	1248	8247	58753	440367
Ratio	.9599	.9807	.9904	1.0038	.9998

This conjecture for twin prime growth goes back to Hardy and Littlewood, who did not use probability. They wrote "*Probability* is not a notion of mathematics, but of philosophy or physics."

Twin primes vs. other prime pair

The count up to x of prime pairs $p, p+2$ or $p, p+6$ should both grow like a constant multiple of $x/(\log x)^2$.

For twin primes the constant is

$$C = \frac{1/2}{(1-1/2)^2} \prod_{p>2} \frac{1-2/p}{(1-1/p)^2} = 2 \prod_{p>2} \frac{1-2/p}{(1-1/p)^2}$$

while for $p, p+6$ the constant is

$$\frac{1/2}{(1-1/2)^2} \frac{2/3}{(1-1/3)^2} \prod_{p>3} \frac{1-2/p}{(1-1/p)^2} = 2C.$$

That's consistent with earlier guess that $\boxed{\pi_{T,T+6}(x) \sim 2\pi_{\text{twin}}(x)}$ coming from numerical data.

Bias in prime counts: Chebyshev's bias

Every prime other than 2 or 5 ends in digit 1, 3, 7, or 9. Dirichlet's theorem in quantitative form says as $x \to \infty$ that for $d = 1, 3, 7, 9$,

$$\frac{|\{p \leq x : p \text{ ends in } d\}|}{|\{p \leq x\}|} \to \frac{1}{4} = 25\%.$$

For $d = 1, 3, 7, 9$, set $\pi_d(x) = |\{p \leq x : p \text{ ends in } d\}|$.

x	$\pi_1(x)$	$\pi_3(x)$	$\pi_7(x)$	$\pi_9(x)$
100	5	7	6	5
1,000	40	42	46	38
10,000	306	310	308	303
100,000	2387	2402	2411	2390
1,000,000	19617	19665	19621	19593

Observations? Assuming Generalized Riemann hypothesis and a bit more, Rubinstein and Sarnak (1994) showed

1. $\pi_1(x), \pi_9(x) < \pi_3(x), \pi_7(x)$ "most of the time"
2. $\pi_3(x)$ vs. $\pi_7(x)$ not biased
3. $\pi_1(x)$ vs. $\pi_9(x)$? Intermediate data suggest not biased

Bias in prime counts: beyond Chebyshev

Picking two digits among 1, 3, 7, 9, count primes $p \leq x$ such that p ends in a digit d and the **next prime** after p ends in a digit d'.

Example. Four primes up to 100 end in 3 and next prime ends in 9: 23 (then 29), 53 (then 59), 73 (then 79), and 83 (then 89).

Since $\pi_1(x), \pi_9(x) < \pi_3(x), \pi_7(x)$ "most of the time," what is an example of a pair of final digits that should make such a count bigger than usual? Here are some results over first 10^8 primes.

d	d'	Count	d	d'	Count
3	3	4,442,562	7	7	4,439,355
3	7	7,043,695	7	3	6,755,195

Assuming a quantitative form of the Hardy–Littlewood conjecture, Lemke Oliver and Soundararajan (2016) showed such counts have a formula with same dominant term but a second term whose sign (positive or negative) depends on digits being different or equal.

Final comments

Although the prime numbers are rigidly determined, they somehow feel like experimental data. Gowers (2002)

It is easier to test primality with a probabilistic algorithm than a deterministic one.

There are far more unsolved problems than solved problems about sets of primes being infinite.

Many statistical conjectures about primes can be motivated by heuristics based on probabilistic ideas such as independence and conditional probability. Use with caution: Artin's first quantitative form of his conjecture on periods of $1/p$ did not fit the data!

Proofs of theorems about primes use techniques from complex analysis, Fourier analysis, ergodic theory, algebraic geometry, etc.

The Generalized Riemann hypothesis is far more important in applications than the Riemann hypothesis alone.