The structure of error terms in number theory and an introduction to the Sato-Tate Conjecture.

Barry Mazur

December 27, 2006

Contents

Abstract

It is wonderful to see the individual strengths of otherwise separate mathematical subdisciplines coming together and connecting with each other (in as startling a way as the theory of continental drift connects the shape of disparate continents) and then providing for us the resolution of a long-sought conjecture. This is indeed what happened last Spring, when a conjecture about certain important probability distributions in number theory, posed forty years ago by Mikio Sato and John Tate, was finally verified for a large number of cases as the culmination of three major works:

- *in the study of modular liftings and automorphic representation theory* (work of Laurent Clozel, Michael Harris, and Richard Taylor [**1**])
- *in algebraic geometry and automorphic representations* (work of Michael Harris, Nicholas Shepherd-Barron, and Richard Taylor [**3**])
- *in Galois deformation theory* (work of Richard Taylor [**12**]).

the last-mentioned breakthrough establishing the result.

My aim is just to discuss, in concrete terms, two "sample problems" —one still open, and one settled by the recent work—that are addressed by the Sato-Tate Conjecture.

1 Error Terms and the Sato-Tate Conjecture

1.1 Why are there still unsolved problems in Number Theory?

Eratosthenes, to take an example—or other ancient Greek mathematicians—might have imagined that all they needed were a few powerful insights and then everything about numbers would be as plain, say, as facts about triangles in the setting of Euclid's *Elements of Geometry*. If Eratosthenes had felt this, and if he now—transported by some time machine—dropped in to visit us, I'm sure he would be quite surprised to see what has developed.

Of course, geometry has evolved splendidly but has expanded to higher realms and more profound structures. Nevertheless, there is hardly a question that Euclid could pose with his vocabulary about triangles that we don't know the answer to today. And, in stark contrast, many of the basic naive queries that Euclid or his contemporaries might have had about primes, perfect numbers, and the like, would still be open.

Sometimes, but not that often, in number theory, we get a complete answer to a question we have posed, an answer that finishes the problem off. Often something else happens: we—perhaps after some major effort—-manage to find a fine, simple, *good approximation* to the data, or phenomena, that interests us, and then we discover that yet deeper questions lie hidden in the error term—in the measure of how badly our approximation misses its mark.

A telling example of this, and of how in the error term lies richness, is the manner in which we study of $\pi(X) :=$ the number of prime numbers less than X. The function $\pi(X)$ is shown below, in various ranges as step functions giving the "staircase" of numbers of primes.

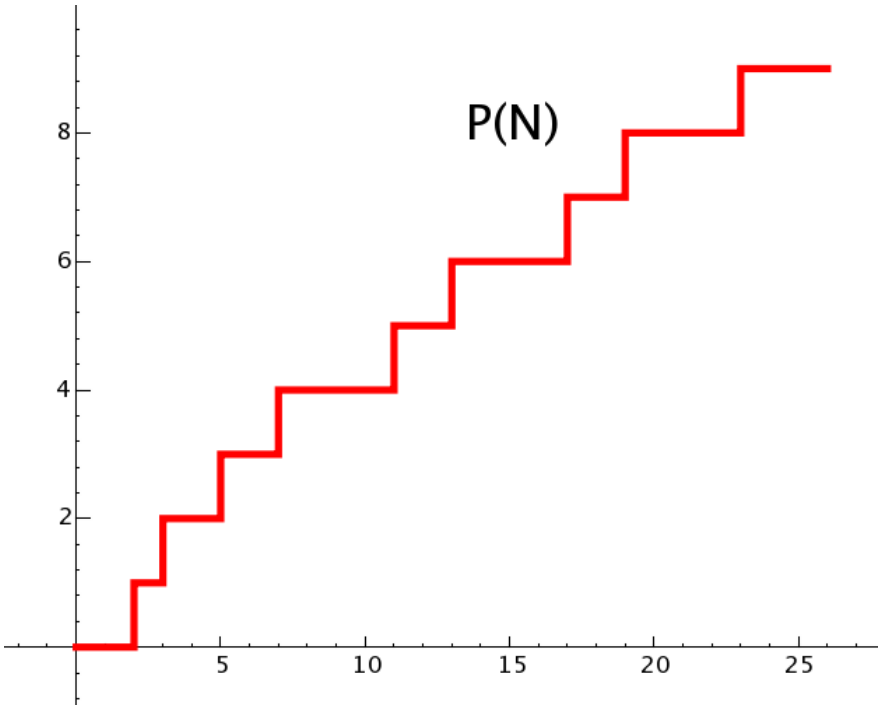

Figure 1.1: The step function $\pi(N)$ counts the number of primes up to N

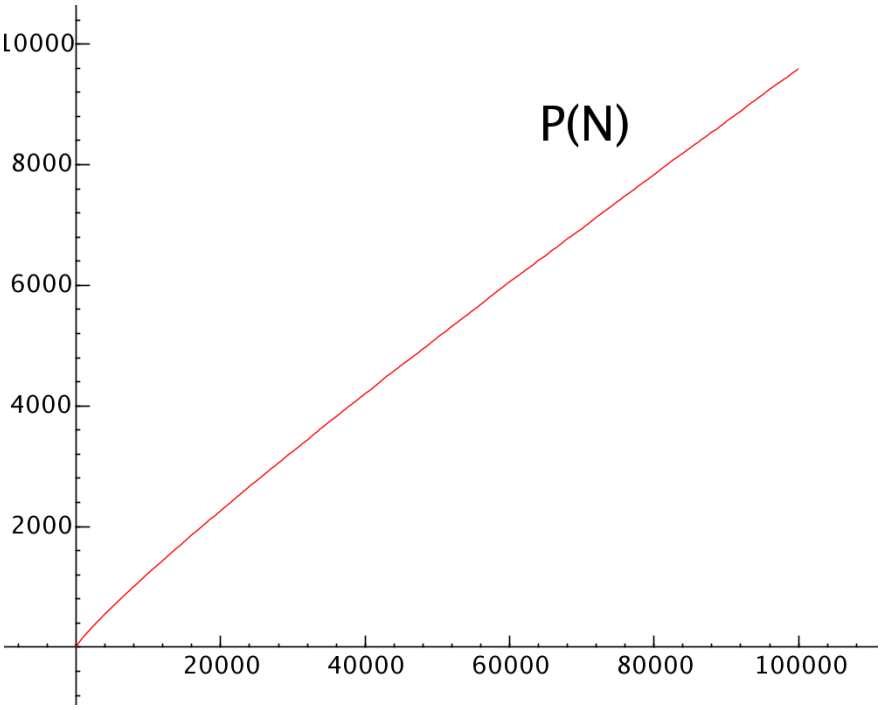

Figure 1.2: The step function $\pi(N)$ counts the number of primes up to N

As is well known, Carl Friedrich Gauss, two centuries ago, computed tables of $\pi(X)$ by hand, for X up to the millions, and offered us a probabilistic "first" guess for a nice smooth approximating curve for this data; a certain beautiful curve that, experimentally, seems to be an exceptionally good fit for the staircase of primes.

The data, as we clearly see, certainly cries out to us to guess a *good approximation*. If you make believe that the chances that a number N is a prime is inversely proportional to the number of digits of N you might well hit upon Gauss's guess, which produces indeed a very good fit. In a letter written in 1849 Gauss claimed that as early as 1792 or 1793 he had already observed that the density of prime numbers over intervals of numbers of a given rough magnitude X seemed to average $1/\log X$.

The Riemann Hypothesis is equivalent to saying that the integral $\int_2^X dx/\log x$ (i.e., the area under the graph of the function $1/\log x$ from 2 to X) is *essentially square root close* to $\pi(X)$. *Essentially square root close* by the way just means that for any given exponent greater than $1/2$ (you choose it: 0.501, 0.5001, 0.50001 for example) and for large enough X—the size, here, depending on your choice of exponent—the difference between $\int_2^X dx/\log x$ and $\pi(X)$ in absolute value is less than X raised to that exponent (e.g. $X^{0.501}$ etc.).

1.2 Much of the depth of the problem is hidden in the structure of the error term.

In a general context, once we make what we hope to be a good approximation to some numerical data, we can focus our attention to the *error term* that has thereby been created, namely:

$$\text{Error term} \;=\; \text{Exact Value} \;-\; \text{Our "good approximation."}$$

In our attempt to understand $\pi(x)$, i.e., the placement of primes in the sequence of natural numbers, we might choose, with Gauss, our *good approximation* to be $\int_2^X dx/\log x$. If so, then we will have focused our mind on the *error term* which so that in this instance we have

$$\text{Error}(x) \;=\; \pi(x) \;-\; \int_2^X dx/\log x.$$

It is Riemann's analysis of—what is in effect— this error term that first showed us the immense world of structure packaged in it. For Riemann did what is, in effect, a Fourier analysis of $\pi(e^t)$ expressing Error(x) (or, more precisely, a closely related function that has the same information as Error(x)) as an *exact* infinite sum of corrective terms, each of these corrective terms easily described in terms of the value of a *zero of the Remann zeta function*; all of these corrective terms are square root small if and only if "his" hypothesis holds[1].

[1] All the data in figures appearing in this article have been tabulated by William Stein.

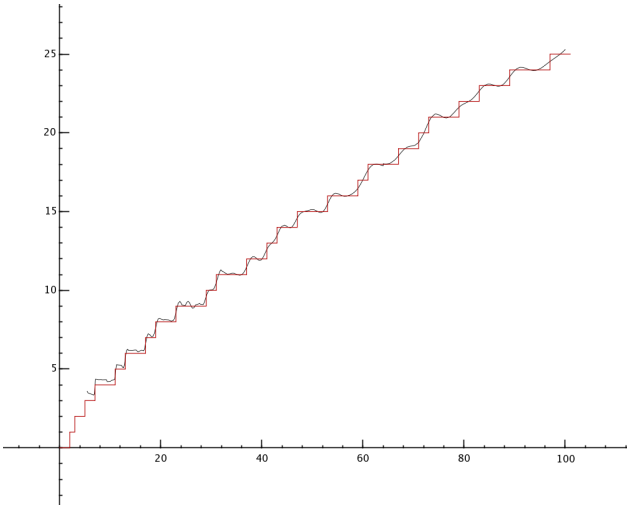

Figure 1.3: The smooth function slithering up the staircase of primes up to 100 is Riemann's approximation that uses the "first" 29 zeroes of the Riemann zeta function

1.3 Strict square-root accuracy

We will be considering a somewhat different class of number theoretic problem than the example that we have been discussing, and for those an even stronger notion of *square-root approximation* is relevant. We will be interested in situations where the *error term* is less than a *fixed constant* times the square root of the quantity being approximated; let us say that an approximation to numerical data has **strict square-root accuracy** if its error term has this property.

We have witnessed great successes in the last century in obtaining good approximations to important problems in Number theory, with error terms demonstrated to be strictly square-root accurate. Specifically, through the work of Helmut Hasse in the 1930s, André Weil in the 1940s and Pierre Deligne in the 1970s, a large class of major approximations were proved to have this kind of accuracy.

1.4 Some Sample Arithmetic Problems

It has known since the time of Fermat, and proved by Euler, that a prime p can be written as a sum of two square numbers if and only if $p \not\equiv 3$ modulo 4 and if it can be written as a sum of two squares, it can be done so in only one way (not counting the order of the two squares). For example:

$$401 = 1^2 + 20^2$$

is the only way (up to changing the order of the two summands) to express the prime number 401 as a sum of two squares. This result is, for many reasons, a much more central and important classical result than it may first appear to be. The problem, which seems to mix *prime numbers* with *geometry* (squares of distances to the origin of integral lattice points in the plane) has the

virtue that its answer is equivalent to knowledge of the splitting properties of primes, and the validity of the unique factorization theorem, in the ring of gaussian integers.

In how many ways can the prime p be expressed as *a sum of the squares of three integers?* The answer for $p \geq 5$ —due to Gauss—can be given in terms of the function $h(-d)$ the class number of the quadratic imaginary field of discriminant $-d$. The number of ways that $p \geq 5$ be expressed as *a sum of the squares of three integers* is:

- $12h(-4p)$ if $p \equiv 1, 5$ modulo 8;
- $24h(-p)$ if $p \equiv 3$ modulo 8;
- 0 if $p \equiv 7$ modulo 8.

The rules of the game here is that the ordering of the summands, and the signs of the integers chosen, count in the tally so for $p = 2$ we have $2 = 0^2 + (\pm 1)^2 + (\pm 1)^2 = (\pm 1)^2 + 0^2 + (\pm 1)^2 = (\pm 1)^2 + (\pm 1)^2 + 0^2$ and therefore we have that 2 can be written "as a sum of three squares" in $3 \cdot 2^2 = 12$ ways.

These two problems are simply the first two of a series of companion questions that have a long history,

In many ways can the prime p be expressed as a sum of the squares of r integers?

To get some sample problems that drive home a point I want to make in this exposition—and for no other reason—of I'll restrict consideration to certain select values of r.

For $r = 4$ we have a simply stated, exact, solution: the prime p can be expressed as a sum of four squares in $8p + 8$ ways.

For $r = 8$, any odd prime number p can be expressed as a sum of eight squares in $16p^3 + 16$ ways.

In both of these cases the answer to our problem (at least for $p > 2$) is a polynomial in p of degree $r/2 - 1$ (i.e., of degree 1 and 3, respectively). Things, however, don't remain as simple, for larger values of r—probably for most larger values of r. To illustrate how things can change, let us focus on $r = 24$.

1.5 Our first "sample problem."

Define, then, $N(p)$ to be the number of ways in which p can be written as a sum of 24 squares of whole numbers.

Recall that squares of positive numbers, negative numbers and zero are all allowed, and the ordering of the squares of the numbers that occur in this summation also counts. Thus, the first prime

number, 2, can already be written as a sum of 24 squares of whole numbers in $1,104$ ways. So:

$$N(2) = 1,104.$$

What about $N(p)$ for the other prime numbers $p = 3, 5, 7, 11, \ldots$? Here is some data.

2	1104
3	16192
5	1362336
7	44981376
11	6631997376
13	41469483552
17	793229226336
19	2697825744960
23	22063059606912
29	282507110257440
31	588326886375936
37	4119646755044256
41	12742799887509216
43	21517654506205632
47	57242599902057216
53	214623041906680992
59	698254765677746880
61	1007558483942335776
67	2827903926520931136
71	5351602023957373056
73	7264293802635839712
79	17319684851070915840
83	29819539398107307072
89	64258709626203556320
97	165626956557080594016

Eyeballing the data, it is already convincingly clear that $N(p)$ is growing less than exponentially, for otherwise the shadow of figures on the page would probably look triangular. Following the pattern we've seen for the smaller values of r we have considered we might expect that $N(p)$ be a polynomial in p of degree $r/2 - 1 = 11$. If we had enough data I imagine we might "curve-fit" a polynomial approximation. But happily, without having to lean on numerical experimentation, certain theoretical issues—which I don't want to get into—allow us to guess the following *good approximation* for the values $N(p)$; namely the polynomial in p of degree 11:

$$N_{\text{approx}}(p) := \frac{16}{691}(p^{11} + 1).$$

The difference, then, between the data and our good approximation is:

$$\text{Error}(p) := \quad N(p) \quad - \quad N_{\text{approx}}(p) \quad = \quad N(p) \quad - \quad \frac{16}{691}(p^{11} + 1).$$

This error term been proven to be square-root small. And perhaps one should emphasize that this square-root smallness statement is hardly an elementary result: it is a consequence of deep work of Deligne. In fact, using the work of Deligne I am alluding to, you can show that:

$$|\text{Error}(p)| \leq \frac{66,304}{691}\sqrt{p^{11}}.$$

What with that hefty constant, $\frac{66,304}{691}$, the "smallness" of our error term here may not impress us for quite a while as we systematically tabulate the values of $N(p)$, but—of course— this result tells us that as we get into the high prime numbers our data will hug startlingly close to the simple smooth curve

$$f(x) = \frac{16}{691}(x^{11} + 1).$$

1.6 The "next question"

Whenever some element of some theory is settled, or is considered settled, many of us mathematicians propose a subsequent plan of inquiry with that phrase: "So, the next question to ask is . . . "

Here too. Given the precise inequality

$$|\text{Error}(p)| \leq \frac{66,304}{691}\sqrt{p^{11}}$$

described in the previous section, and given the fact that this represents one consequence of what has been a great project that has spanned half a century of progress in number theory, some natural (and related) "next" questions arise. We might—for example—ask

- Is the bound on this error term (e.g., the constant $\frac{66,304}{691}$) is the best possible?
- Is $f(x) = \frac{16}{691}(x^{11} + 1)$ the *best* polynomial approximation to our data?
- Might we, more specifically, find another polynomial $g(x)$ which *beats* $f(x)$ in the sense that the absolute values of the corresponding error terms $|N(p) - g(p)|$ are $\leq C\sqrt{p^{11}}$ with a constant C that is strictly less than $\frac{66,304}{691}$?
- For any given constant $C < \frac{66,304}{691}$ is there a positive proportion of prime numbers p for which

$$|N(p) - f(p)| \leq C\sqrt{p^{11}}.$$

- We might ask what that proportion is, as a function of C.
- We might ask for the proportion of primes p for which the error term is positive, i.e., where our good approximation is an undercount.

To be sure, we would want to phrase such questions not only about our specific "sample problem" but about the full range of problems for which we have—thanks to Deligne et al— such good square-root close approximations.

It is the *Sato-Tate Conjecture* that addresses this "next," more delicate, tier of questions[2].

1.7 The distribution of scaled error terms

Given that in our sample problem we know the bound

$$|\text{Error}(p)| \leq \frac{66,304}{691}\sqrt{p^{11}},$$

let us focus our microscope on the fluctuations here. Namely, consider the *scaled* error term

$$\text{Scaled Error}(p) := \quad \frac{\text{Error}(p)}{\frac{66,304}{691}\sqrt{p^{11}}} \quad = \quad \frac{N(p) \quad - \quad \frac{16}{691}(p^{11}+1)}{\frac{66,304}{691}\sqrt{p^{11}}}$$

so that we have:

$$-1 \quad \leq \quad \text{Scaled Error}(p) \quad \leq \quad +1.$$

About this type of *scaled error value distribution,* let me recall the words of Susan Holmes, a mathematician and statistician at Stanford, who—when I sent her some numerical computations related to a similar number theoretic problem for which I had some statistical questions—exclaimed: "what beautiful data!"

But what can we say further about this data? How do these scaled error values distribute themselves on the interval $[-1,+1]$? That is, what is the function $I \mapsto \mathcal{P}(I)$ that associates to any subinterval I contained in $[-1,+1]$ the *probability* $\mathcal{P}(I)$ that for a randomly chosen prime number p its scaled error term $\text{Error}(p)$ lies in I?

In 1960, Mikio Sato (by studying numerical data) and John Tate (following a theoretical investigation) predicted—for a large class of number theoretic questions including many problems of current interest, of which our example is one—that the values of the scaled error terms for data in these problems conforms to a specific probability distribution, Usually the Sato-Tate conjecture predicts that this distribution is no more complicated than the elementary function $t \mapsto \frac{2}{\pi}\sqrt{1-t^2}$, i.e., the thing whose graph is a semi-circle of radius 1 centered at the origin, but squished vertically to have its integral equal to one. This makes it far from the Gaussian normal distribution! Indeed, Sato

[2] As is only to be expected, there are whole books of questions about this sample problem that one could ask, and mathematicians have asked—some of these questions being structurally important, and some at least traditionally of great interest. Eg., how often is our approximate value $N_{\text{approx}}(p)$ above *exactly* equal to the actual value $N(p)$? A conjecture of Lehmer would say that this never happens.

and Tate predict this type of behavior in our example problem, so that their conjecture would have it that

$$\mathcal{P}(I) = \frac{2}{\pi} \int_I \sqrt{1-t^2} dt.$$

This is still an open question, for our sample problem! Nevertheless, we have an impressive amount of data in support of it (see below).

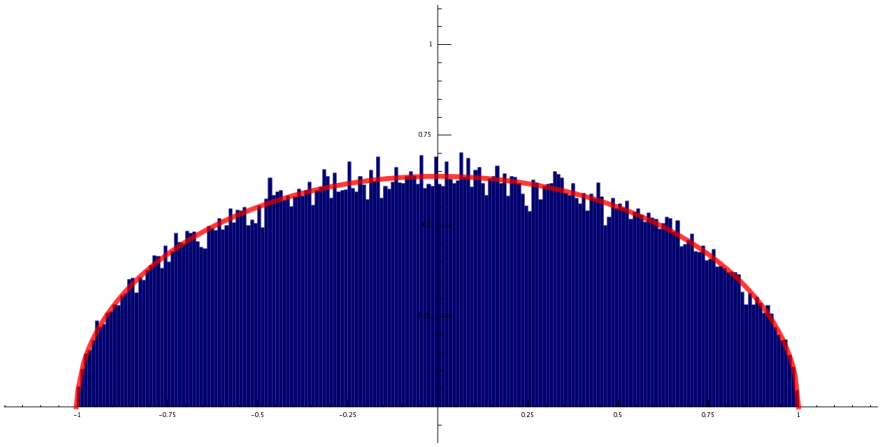

Figure 1.4: **Probability distribution of error terms.** The Sato-Tate distribution $\frac{2}{\pi}\sqrt{1-t^2}$, the smooth profile curve in this figure, can be compared with the probability distribution of *scaled error terms* for the number of ways $N(p)$ in which a prime number p can be written as a sum of 24 squares ($p < 10^6$). These computations were made by William Stein.

To compute this data for primes up to 10^6 in reasonable time required much ingenuity on the part of William Stein. For a more complete description of what these computations entail, together with other background material, consult William Stein's: http://sage.math.washington.edu:8100/193

The great breakthrough last Spring was the resolution of the Sato-Tate conjecture for a large class of elliptic curves.

2 An elliptic curve, and a new "sample problem"

2.1 The number of points of an elliptic curve mod p; for varying p

The example we will use is one of the favorites of many number theorists, namely the curve in the plane, call it E, cut out by the equation

$$y^2 + y = x^3 - x^2.$$

This is an elliptic curve that is something of a showcase for number theory, in that it has been extensively studied—much is known about it—and yet it continues to repay study, for—as with all other elliptic curves—its deeper features have yet to be understood.

This curve $E : y^2 + y = x^3 - x^2$ when extended to the projective plane has exactly one rational point on the line at infinity, and if you stipulate that that unique point "at infinity" be the *origin,* there is a unique algebraic group law on E, allowing us—for any field k of characteristic different from 11 (i.e., any field where $11 \neq 0$)—to endow the set consisting of ∞ and the points of E with values $(x, y) = (a, b) \in k$ with the structure of an abelian group. Let k be of characteristic different from 11 and let us denote by $E(k)$ this group of k-rational points of E. The reason why we have to exclude 11 is that the polynomial equation above modulo 11 has a singular point.

Every one of these groups $E(k)$ contains the five rational points

$$\{\infty,\ (0,0),\ (0,-1),\ (1,0),\ (1,-1),\}$$

and it isn't difficulty to check that these five points comprise a cyclic subgroup of $E(k)$ of order five. The *data* we shall be focussing on, in this problem is *the number of rational points that E has over the prime field containing p elements* (excluding, again, $p = 11$). So, let p be a prime number (different from 11) and let $\mathbf{F}_p$ denote the field of integers modulo p, and define

$$N(p) := \text{the number of elements in the finite group } E(\mathbf{F}_p).$$

There is much that is surprising in this "data.' That is the numerical function

$$p \longmapsto N(p)$$

or (essentially equivalently) the error term we are now concentrating on:

$$p \longmapsto Error(p) = N(p) - (p + 1)$$

and it can be expressed in a few quite different-sounding ways. Here is one: Expand the infinite product

$$q \prod_{n=1}^{\infty} (1 - q^n)^2 (1 - q^{11n})^2 = \sum a_n q^n$$

and we have that:

$$Error(p) = -a_p.$$

Here is what $N(p)$ looks like for small primes p:

p	2	3	5	7	13	17	19	23	29	31	37	41	43	47	53	59	61	67	71
$N(p)$	5	5	5	10	10	20	20	25	30	25	35	50	50	40	60	55	50	75	75

Since, from the first of the two definitions, $N(p)$ is the order of a finite group that contains a cyclic group of order five, we know, from Lagrange's theorem of elementary group theory that $N(p)$ is divisible by 5, but what more can we say about the data

$$p \longmapsto N(p)?$$

This, now, will constitute our *sample problem* on which be focussing for the rest of this article.

For starters, following the format of the the previous sections of this article, we should look for a "good approximation" to $N(p)$. An old result due to Helmut Hasse tells us that a square-root accurate approximation to $N(p)$ is given by the simple expression: $p+1$, which is, by the way, just the number of points on a line in the projective plane over $\mathbf{F}_p$.

It s a deep theorem (proved in the PhD thesis of Noam Elkies) that for this elliptic curve as well as any other elliptic curved defined over Q there are an infinite number of primes p such $N(p)$ is equal to *precisely* this simple expression $p+1$. But, it is generally true that the error term for this approximation is quite small. Explicitly, writing

$$\text{Error}(p) := \ N(p) \ - \ (p+1)$$

Hasse proved the inequality

$$|\text{Error}(p)| = |N(p) \ - \ (p+1)| \ \leq \ 2\sqrt{p}.$$

Another way of saying this is that there is a conjugate pair of complex numbers $e^{i\theta_p}$ and $e^{-i\theta_p}$ for which the error term can be written as

$$\text{Error}(p) := \ N(p) \ - \ (p+1) \ = \ \sqrt{p}(e^{i\theta_p} + e^{-i\theta_p}) = \ 2\sqrt{p}\cos(\theta_p).$$

Following, again, the format of our example-problem of the previous sections, we might ask for the distribution of error values, and here we can do this just by asking for the statistics of the rule that assigns to prime numbers p the conjugate-pair of complex numbers on the unit circle in the complex plane

$$p \longmapsto e^{\pm i\theta_p}.$$

Here is some data:

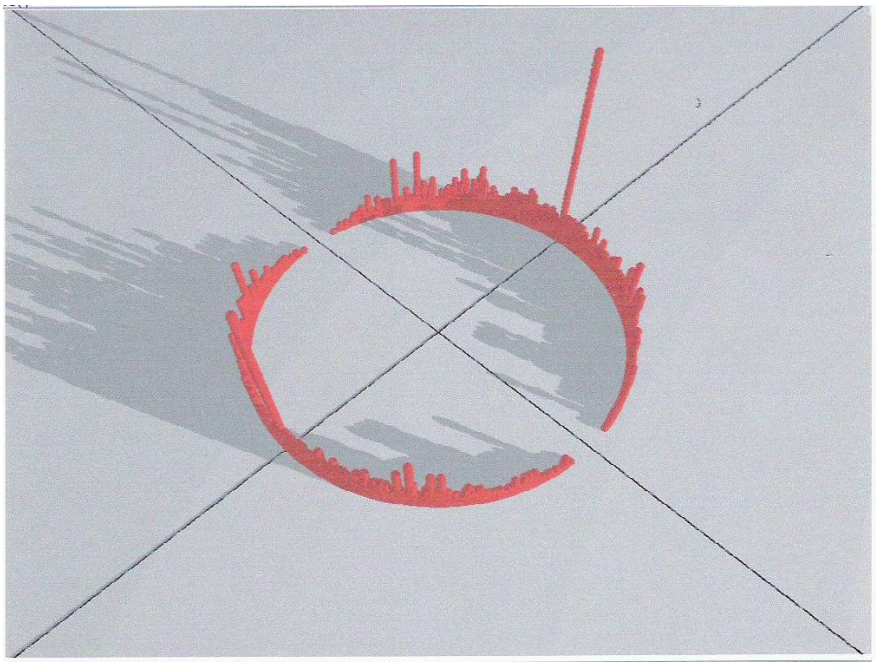

Figure 2.1: The accumulation of red dots on a position in this unit circle of the complex plane corresponds to the frequency of occurrences of θ_p in a small arc around that position for a good number of primes p. The two axis lines are the x-axis and y-axis; from the data—which conforms to the Sato-Tate statistics—you can guess which is which.

2.2 The Sato-Tate distribution

Thanks to the recent advance due to Taylor et al, the data

$$p \longmapsto \cos(\theta_p) = 1/2(e^{i\theta_p} + e^{-i\theta_p})$$

of the previous section conforms to the Sato-Tate distribution $\frac{2}{\pi}\sqrt{1-t^2}$. That is,

Theorem 2.1. (The Sato-Tate Conjecture for our sample case) *For any continuous function $F(t)$ on the interval $[-1,+1]$ we have that the limit*

$$\lim_{X\to\infty} \sum_{p\leq X} F(\cos\theta_p)/\pi(X)$$

exists and is equal to the integral

$$\frac{2}{\pi}\int_{-1}^{+1} F(t)\sqrt{1-t^2}dt.$$

To express our expected distribution in terms of the θ_p's, one could make the change of variables $(t \mapsto \cos\theta)$

$$\frac{2}{\pi}\int_{-1}^{+1} F(t)\sqrt{1-t^2}dt \;=\; \frac{1}{\pi}\int_{-\pi}^{+\pi} F(\cos\theta)\sin^2\theta d\theta \;=\; \frac{2}{\pi}\int_{0}^{+\pi} F(\cos\theta)\sin^2\theta d\theta,$$

i.e., expressing things in terms of θ we get a "sine-squared" distribution. Here is what the data looks like in these terms:

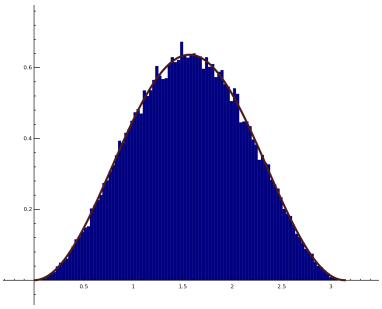

Figure 2.2:

In a sequel to these notes (a sequel yet to be written) I would like to say a few things for nonexperts about the actual proof of this theorem. But to conclude here let us see how the problem reduces to a study of L-functions.

To prove the above theorem, it would be enough to show that

$$\lim_{X\to\infty} \sum_{p\leq X} F(\cos\theta_p)/\pi(X) = \frac{2}{\pi}\int_{-1}^{+1} F(t)\sqrt{1-t^2}dt.$$

for all real-valued polynomial functions $F(t)$ by the Weierstrass approximation theorem, and therefore, since our task is linear, we could concentrate on proving this for $F(t) =$ all the powers of the variable t, i.e.,

$$1, t, t^2, t^3, \ldots$$

or, for that matter it would suffice to prove it for $F(t) =$ any other $\mathbf{R}$-basis of the ring of real-valued polynomials.

2.3 Bases for the ring of polynomials

Write the variable x as a sum $\alpha + \alpha^{-1}$ so that any polynomial in x (with, e.g., real coefficients) is a polynomial in α and α^{-1} invariant under the interchange $\alpha \leftrightarrow \alpha^{-1}$, and conversely: any polynomial in α and α^{-1} invariant under the above interchange is a polynomial in x. Consider then, these polynomials (let's call them *symmetric power polynomials*)

$$\begin{aligned} s_0 &= 1 \\ s_1 &= \alpha + \alpha^{-1} \\ s_2 &= \alpha^2 + 1 + \alpha^{-2} \\ s_3 &= \alpha^3 + \alpha^1 + \alpha^{-1} + \alpha^{-3} \\ s_4 &= \alpha^4 + \alpha^2 + 1 + \alpha^{-2} + \alpha^{-4} \\ s_5 &= \alpha^5 + \alpha^3 + \alpha^1 + \alpha^{-1} + \alpha^{-3} + \alpha^{-5} \\ &\ldots \end{aligned} \tag{2.1}$$

which, when expressed as polynomials in x, look like

$$\begin{aligned} s_0(x) &= 1 \\ s_1(x) &= x \\ s_2(x) &= x^2 - 1 \\ s_3(x) &= x^3 - 2x \\ s_4(x) &= x^4 - 3x^2 + 1 \\ s_5(x) &= x^5 - 4x^3 + 3x \\ &\ldots \end{aligned} \tag{2.2}$$

where $s_n(x)$ is a monic polynomial in x of degree n (these are also called the *Chebychev polynomials of the second kind*). They form a basis of the vector space of polynomials in the variable x. Any collection of products

$$\{s_m(2t)s_n(2t)\}_{(m,n)\in\mathcal{I}}$$

forms a basis of the vector space of polynomials in the variable t where $\mathcal{I}$ is a collection of a pairs of nonnegative integers such that the sums $m+n$ run through all nonegative numbers with no repeats.

Here is an elementary calculus exercise:

Proposition 2.2. *If $F(t) = s_m(2t)s_n(2t)$ with $m \neq n$ then*

$$\frac{2}{\pi}\int_{-1}^{+1} F(t)\sqrt{1-t^2}dt = 0.$$

Corollary 2.3. *Theorem 2.1 would follow if for every positive integer k there is a pair of distinct nonnegative integers (n,m) with $n+m=k$ and such that*

$$\lim_{X\to\infty} \sum_{p\leq X} s_m(2\cos\theta_p)s_n(2\cos\theta_p)/\pi(X) = 0.$$

But how can we get that such limits vanish? A standard strategy—in fact, it seems, the only known strategy—is to invoke L functions. So we turn to:

2.4 L-functions

To study the data $\{p \longmapsto \pm\theta_p\}$ effectively it is a good idea to "package it" into complex analytic functions (Dirichlet series) whose behavior will tell us about the limits described in Corollary 2.3.

Let us do this. For any choice of prime number p different from 11 and for any pair of nonnegative numbers $0 \leq m \leq n$, define *the local factor at p of the L-function* $L_{m,n}(s)$ as follows:

$$L_{m,n}^{\{p\}}(s) := \prod_{j=0}^{m}\prod_{k=0}^{n}\left(1-e^{i(m+n-2j-2k)\theta_p}p^{-s}\right)^{-1}.$$

If m (or n) is zero, the factors in "$\prod_{j=0}^{m}$" (or "$\prod_{k=0}^{n}$") don't occur, so, for example:

$$L_{0,n}^{\{p\}}(s) := \prod_{k=0}^{n}\left(1-e^{i(n-2k)\theta_p}p^{-s}\right)^{-1}.$$

Now form the infinite product over all prime numbers p different from 11:

$$L_{m,n}(s) := \prod_{p} L_{m,n}^{\{p\}}(s)$$

and expand this to get a Dirichlet series

$$L_{m,n}(s) = \sum_{r=0}^{\infty} a_{m,n}(r) r^{-s}.$$

The terms $a_{m,n}(r)$ are easily computed: we have, for example, that $a_{m,n}(p) = s_m(2\cos\theta_p)s_n(2\cos\theta_p)$ for p a prime number different from 11, and for any positive integer r the term $a_{m,n}(r)$ is bounded from above in absolute value by a fixed polynomial (depending only on m and n) in $\log(r)$. This guarantees that the Dirichlet series $L_{m,n}(s)$ converges in the half-plane $Re(s) > 1$.

Here we rely on analytic number theory (in the form of a classical theorem of Wiener and Ikehara) which gives us that if we know enough further analytic facts about these Dirichlet series $\sum a_{m,n}(r)r^{-s}$ we can control limits of the form

$$\lim_{X\to\infty} \frac{\sum_{p<X} a_{m,n}(p)}{\pi(X)},$$

i.e., since $a_{m,n}(p) = s_m(2\cos\theta_p)s_n(2\cos\theta_p)$ $(p \neq 11)$ these are exactly the limits we are interested in.

Proposition 2.4. *Let $m < n$. If $L_{m,n}(s)$ extends to a meromorphic function on the entire complex plane, holomorphic on* $\mathrm{Re}(s) \geq 1$ *and nonzero on all points* $\mathrm{Re}(s) = 1$ *other than $s = 1$ then*

$$\lim_{X\to\infty} \sum_{p\leq X} s_m(2\cos\theta_p)s_n(2\cos\theta_p)/\pi(X) = 0.$$

If, by the way, $L_{m,n}(s)$ extended to a meromorphic function on the entire complex plane, holomorphic on $\mathrm{Re}(s) \geq 1$ except for having a pole of order k at $s = 1$ the analytic proposition above would tell us that the limit is k, rather than 0.

This analytic theorem follows from classical results due to Weiner and Ikehara. A beautiful discussion of these ideas and proofs can be found in the Appendix to Chapter 1 of Serre's monograph [**7**]. See also Tate's article [**11**], Shahidi's article [**9**] and Serre's letter to Shahidi [**8**] that discusses in some detail the implications in the direction of the Sato-Tate conjecture that would follow if one assumes that the $L_{0,\nu}$'s satisfy the hypotheses of Proposition 2.4 for $\nu \leq d$. This knowledge is known for $\nu = 1$ (our "sample problem" comes from a modular form; indeed by the celebrated results regarding modularity of elliptic curves, it would be known for any elliptic curve defined over Q that we care to choose). It is also known for $\nu = 2$ using an integral representation due to Shimura [**10**]; see also Gelbart's and Jacquet's article [**2**]. It is known for $\nu = 3, 4$ by work of Shahidi; see the enlightening discussion in [**9**] about this)[3].

2.5 The coming together of different mathematical viewpoints

But how can we get that Dirichlet series such as $L_{m,n}(s)$ extend meromorphically to the entire complex plane for *enough* values of (m, n) to guarantee that we have computed all the moments of

[3]The corresponding symmetric cube and fourth power of the modular form of weight two (corresponding to our sample problem) are cuspidal automorphic forms; cf. the articles [**4**], [**5**] by Kim and Shahidi.

the distribution determined by our data? And how can we determine whether these meromorphic extensions have (or better: don't have) zeroes or poles on the line $Re(s) = 1$? A standard strategy—in fact, it seems, the only known strategy to get L-functions to have all the analytic properties that they need to have is to connect these L-functions with automorphic forms over Q or with pairs of automorphic forms on GL_m and GL_n over Q relying on ideas of Rankin-Selberg. For the problem we are interested in, it turns out that one gets sufficiently valuable information if one can make the analogous connection with automorphic forms over *some* number field F—not necessarily Q—so long as F is Galois over Q, and totally real.

Part of the beauty of the new theorem we are discussing—which applies, in fact, to all elliptic curves over Q that have at least one prime of multiplicative reduction—is how it pulls together work from significantly different viewpoints. There are three major pieces that go into it: work of Laurent Clozel, Michael Harris, and Richard Taylor) on modular lifting and *automorphic representation theory*; work of Michael Harris, Nicholas Shepherd-Barron, and Richard Taylor bringing in an extraordinary piece of *algebraic geometry:* the pencil of Calabi-Yau varieties

$$X_0^{n+1} + X_1^{n+1} + \cdots + X_n^{n+1} = (n+1)tX_0X_1 \ldots X_n$$

for even values of n, parametrized by the variable t; and the last: Richard Taylor's major discovery in *Galois deformation theory* which, using ideas of Mark Kisin, improved dramatically the mechanism of modular lifting, allowing Richard Taylor to prove this extraordinary result.

2.6 Expository accounts of this recent work

Different audiences benefit from different shapes of exposition. I wrote a brief "news" article in the journal Nature **[6]** meant to give a hint of the nature of the Sato-Tate Conjecture and some related mathematical problems to scientists who are not necessarily familiar with much modern mathematics. For professional mathematicians, a number of excellent articles and videos—requiring different levels of prerequisites of their audiences—are devoted to exposing this material:

1. Available through the MSRI website (http://www.msri.org/):
   (a) An introductory one hour lecture by Nicholas Katz emphasizing the background and the historical perspective of the work.
   (b) A series of lectures for a number theory workshop, by Richard Taylor; by Michael, Harris; and by Nicholas Shepherd-Barron, where an exposition of the proof itself is given.
2. Two hours of expository lectures by Laurent Clozel on this topic which goes in considerable detail through the ideas of the proof, aimed at a general mathematical audience, delivered in the conference on Current Developments in Mathematics, at Harvard University. The notes for these should soon be available as well.
3. An expository article by Michael Harris: "The Sato-Tate Conjecture: introduction to the proof." This will be submitted to the proceedings of the École d'été Franco-Asiatique, held at the IHES during the summer of 2006.

References

[1] Clozel,L., Harris, M., Taylor, R.: Automorphy for some ℓ-adic lifts of automorphic mod l representations (preprint) http://www.math.harvard.edu/ rtaylor/

[2] Gelbart, S., Jacquet, H.: A relation between automorphic representations of $G(2)$ and $GL(3)$, Ann. Sci. École Norm. Sup. (4) **11** (1978) 471-552

[3] Harris, M., Shepherd-Barron, N., Taylor, R.: Ihara's lemma and potential automorphy (preprint) http://www.math.harvard.edu/ rtaylor/

[4] Kim, H., Shahidi, F.: Cuspidality of symmetric powers with applications, Duke Math. J. **112**, no. 1 (2002), 177197

[5] Kim, H., Shahidi, F.: Functorial products for GL(2) $\times$ GL(3) and the symmetric cube for GL(2), Annals of Math. **155** (2002), 837-893

[6] Mazur, B.: Controlling our Errors, Nature Vol **443**, **7** (2006) 38-40

[7] Serre, J.-P.: *Abelian ℓ-adic Representations* Benjamin (1968)

[8] Serre, J.-P.: Letter to F. Shahidi, January 24, 1992; Appendix (pp. 175-180) in reference 9 below,

[9] Shahidi, F.: Symmetric power L-functions for GL(2), pp. 159-182, in *Elliptic Curves and Related Topics,* Volume 4 of CRM Proceedings and Lecture Notes (Eds.: H. Kisilevsky, M.R. Murty), American Mathematical Society, (1994)

[10] Shimura, G.: On the holomorphy of certain Dirichlet series, Proc. London Math. Soc (3) **31** (1975) 79-98

[11] Tate, J.: Algebraic Cycles and Poles of Zeta Functions, pp.93-110 in *Arithmetic Algebraic Geometry*, Proceedings of a conference held in Purdue, Dec. 5-7 1963, Harpers (1965)

[12] Taylor, R.: Automorphy for some ℓ-adic lifts of automorphic mod ℓ representations. II (preprint) http://www.math.harvard.edu/ rtaylor/