

Statistical properties of the primes and the Riemann zeros

P.J. Forrester

My objective in this article is to introduce the reader to some fundamental probabilistic and statistical properties of the primes and the Riemann zeros. This topic is at once profound — containing the celebrated Riemann hypothesis — and accessible, relating to familiar quantities and allowing for a degree of experimentation. The experimentation may take the form of numerical computations, which in relation to the primes can be a simple exercise once one is equipped with appropriate software, or heuristic reasoning, whereby partial insights can be used to formulate precise conjectures.

Private hand produced tables of primes in the late 1700's and early 1800's by Gauss and Legendre led to the first conjectured statistical property of the primes. This concerned an asymptotic formula for $\pi(N)$, the number of primes out of the first N natural numbers. The data suggests that up to 10^n , for increasing n , the proportion of primes is about $1 : 2.3n$. This then led to Legendre's logarithmic law, which states that to leading order, out of the first N natural numbers 1 in $\log N$ are prime, and thus

$$\pi(N) \sim \frac{N}{\log N}. \quad (1)$$

Gauss refined the estimate (1) by observing that for large N the mean spacing between primes is to leading order $1/\log N$. In general integration of the density gives the expected number, so Gauss obtained

$$\pi(N) \sim \int_2^N \frac{dt}{\log t} =: \text{Li}(N). \quad (2)$$

The quantity $\text{Li}(N)$ is called the logarithmic integral. Its leading asymptotic form reproduces (1), but in its entirety Gauss' estimate is typically much more accurate than Legendre's (see e.g. [3, Table 7-3]). The empirical observation (1) is indeed a theorem, referred to as the prime number theorem, as proved in separate works near the end of the 1800's by Hadamard and de la Vallée Poussin.

We now turn our attention to an empirical observation associated with (2) which turns out to be false in general. This relates to the sign of $\text{Li}(N) - \pi(N)$. All numerical data gives that $\text{Li}(N) - \pi(N) > 0$ and further indicates this to be an increasing function of N . However in 1914 Littlewood proved that these numerical observations do not persist for general N , and in fact $\text{Li}(N) - \pi(N)$ changes sign infinitely often. In recent times it has been proved that a sign change must occur in the vicinity of 1.39822×10^{316} [3, pg. 236], which is the best upper bound on the first sign change known to date. The enormity of this number indicates that at the least some degree of caution must be exercised when relying solely on numerical data to formulate conjectures of this type. As an aside we remark that recent analytic studies [12] have quantified the proportion of x values for which

$$\text{Li}(x) - \pi(x) < 0. \quad (3)$$

In particular, it has been shown that the logarithmic density of the set (3), which is the value of $\frac{1}{\log X} \int dt/t$ integrated over all values of t for which (3) holds in the range $[2, X]$, is approximately equal to 0.26×10^{-6} as $X \rightarrow \infty$.

A natural question relating to $\text{Li}(N) - \pi(N)$ is a bound on its size. Littlewood showed that this difference oscillates in both directions by at least $\text{Li}(\sqrt{x}) \log \log \log x$. Van Koch noted that the Riemann hypothesis (RH) implies

$$\pi(x) = \text{Li}(x) + O(\sqrt{x} \log x), \tag{4}$$

and furthermore it is known that the bound on the error term in (4) implies RH [2]. Thus (4) is in fact equivalent to RH. It is thus appropriate to now recall RH. Encoding Euclid’s theorem on unique factorization of a natural number into primes (the fundamental theorem of arithmetic) is the formula

$$\sum_{n=1}^{\infty} \frac{1}{n^s} = \prod_{\text{primes } p} (1 - p^{-s})^{-1}, \quad \text{Re}(s) > 1$$

due to Euler. Riemann studied this function of s , now referred to as the Riemann zeta function $\zeta(s)$, for complex values of s in relation to his studies on the Gauss formula (2). He was led to conjecture that the analytic continuation of $\zeta(s)$, for $0 \leq \text{Re}(s) \leq 1$, has zeros along the critical line $\text{Re}(s) = 1/2$ only (of which there must be an infinite number), and his conjecture is now referred to as RH.

The occurrence of the factor \sqrt{x} in the bound (4) is most interesting, for it ties in with probability theory. As an illustration, suppose we choose a sequence of numbers x_1, x_2, x_3, \dots uniformly at random from the unit interval $[0, 1]$. What is the distribution of the sum $\sum_{j=1}^N x_j$ for large N ? This question is answered by the central limit theorem, which tells us that

$$\sum_{j=1}^N x_j \sim \frac{1}{2}N + O(\sqrt{N}).$$

More generally the central limit theorem tells us that fluctuations about the mean of order \sqrt{N} are typical of processes involving independent random numbers.

This observation gives motivation for Cramér’s probabilistic model of the primes. Here it is hypothesized that statistical properties of the primes are well described by a model in which each natural number N is a prime with probability $1/\log N$. Note in particular that there are no built in correlations between primes. By construction, this model is consistent with the prime number theorem. Moreover, it can be used to predict the asymptotic form of the quantity $\pi_m(x)$ — the number of prime pairs $(p, p + m)$. The probabilistic model suggests that the number of such prime pairs is proportional to

$$\int_2^x \frac{dt}{\log t \log(t + m)} \underset{x \rightarrow \infty}{\sim} \frac{x}{(\log x)^2}.$$

This can’t possibly be correct for general natural numbers m as for a start $\pi_m(x) = 0$ for m odd. Nonetheless, at a heuristic level (see [11]) these arithmetic considerations can be overcome, and for m even one is led to a well known conjecture of Hardy and Littlewood, namely that

$$\pi_m(x) \sim 2C_2 \frac{x}{(\log x)^2} \prod_{\substack{p > 2 \\ p|m}} \frac{p-1}{p-2}, \tag{5}$$

where $C_2 = \prod_{p > 2} (1 - 1/(p-1)^2)$ (all products are over primes only).

Associated with (5) is a density function

$$d_m(x) = \frac{d}{dx} \pi_m(x). \quad (6)$$

(For fixed x , (6) only makes sense when considering a smoothed version of $\pi_m(x)$, but this is not a concern in relation to (5).) If $d_0(x)$ denotes the asymptotic density of primes at x , so that $d_0(x) = 1/\log x$, we see from (5) and (6) that

$$\lim_{x \rightarrow \infty} \frac{d_m(x)}{(d_0(x))^2} = 2C_2 \prod_{\substack{p > 2 \\ p|m}} \frac{p-1}{p-2}. \quad (7)$$

If the primes were truly asymptotically independent the right hand side of (7) would equal unity. Its nontrivial dependence on m is due to the arithmetic effects incorporated in (5). For large m the right hand side of (7) has the smoothed form [7]

$$1 - \frac{1}{2m}. \quad (8)$$

This is a feature of great importance in the study of the statistics of the Riemann zeros.

The Riemann zeros refer to the zeros of $\zeta(s)$ on the critical line $\text{Re}(s) = \frac{1}{2}$. Riemann showed that the number of zeros in the critical strip $0 \leq \text{Re}(s) \leq 1$ with imaginary part between 0 and T grows asymptotically as $(T/2\pi) \log(T/2\pi)$. Assuming RH, this implies that the density of the Riemann zeros has the leading asymptotic form $(1/2\pi) \log T$. Unlike the situation with primes, there is no truth in modelling the zeros as occurring independently at random with this density along the critical line. Rather, analytic evidence of Montgomery [9] and numerical evidence of Odlyzko [10] indicates a markedly different scenario. This is that the statistical properties of the large Riemann zeros agree with the statistical properties of the eigenvalues of large random Hermitian matrices. Here it is assumed that both the sequences of zeros and eigenvalues have been rescaled to have their mean spacing unity.

To test this hypothesis (referred to as the Montgomery-Odlyzko law [6]) numerically, in addition to having available large sequences of high zeros [10], one requires knowledge of the statistical properties of large Hermitian matrices. This is a topic initiated in theoretical physics over 40 years ago, and is still an active research area today (see e.g. [5]). It is also my own main research area. One reason for the continued activity is the essential role played by Painlevé transcendents. To illustrate a result of this type of relevance to the Riemann zeros, let us consider the problem of computing the distribution of closest neighbour spacings $p_{\text{cn}}(s)$ of the bulk eigenvalues for large Hermitian matrices. For a sequence $\dots < x_{n-1} < x_n < x_{n+1} < \dots$ the closest neighbour spacing at x_n is defined as $\min(x_n - x_{n-1}, x_{n+1} - x_n)$. If $E(-s, s)$ is the probability that about an eigenvalue at the origin, there is no other eigenvalue in the interval $(-s, s)$, it is easy to see that

$$p_{\text{cn}}(s) = -\frac{d}{ds} E(-s, s).$$

In the case of random Hermitian matrices from the so called Gaussian unitary ensemble, orthogonal polynomial methods allow $E(-s, s)$ to be expressed as a Fredholm determinant, and this Fredholm determinant can in turn be characterized as the solution of a nonlinear equation [4]. The final result is that

$$p_{\text{cn}}(s) = -\frac{\sigma(2\pi s)}{2\pi s} \exp\left(\int_0^{2\pi s} \frac{\sigma(t)}{t} dt\right) \quad (9)$$

where σ satisfies the nonlinear equation

$$(s\sigma'')^2 + 4(-1 + s\sigma' - \sigma)\left((\sigma')^2 - \{1 - (1 - s\sigma' + \sigma)^{1/2}\}^2\right) = 0 \tag{10}$$

subject to the boundary condition

$$\sigma(s) \underset{s \rightarrow 0^+}{\sim} -\frac{s^3}{12\pi}.$$

In [14] σ is expressed in terms of a Painlevé V transcendent, while in [13] the equation (10) itself is identified with the equation satisfied by an auxiliary Hamiltonian in the Painlevé III theory.

The result (9) was used in [4] as a test of the Montgomery–Odlyzko law. Thus one compares the empirical determination of $p_{\text{cn}}(s)$ for large sequences of Riemann zeros, starting at different positions along the critical line, with the random matrix distribution (9). The results, which are consistent with the Montgomery–Odlyzko law, are reproduced in Figure 1.

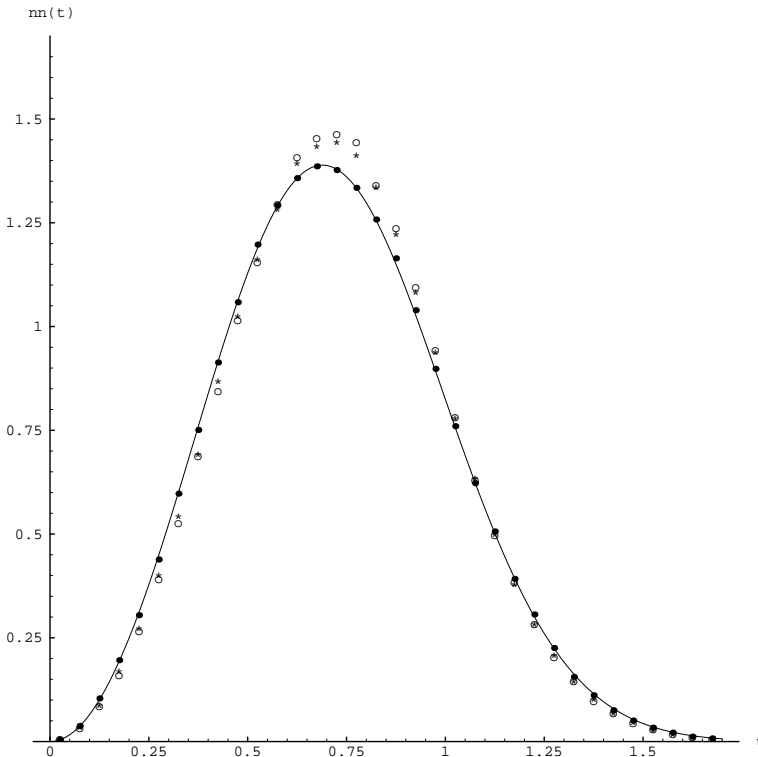


Figure 1. Comparison of $nn(t) := p_{\text{cn}}(t)$ for the GUE (continuous curve) and for 10^6 consecutive Riemann zeros, starting near zero number 1 (open circles), 10^6 (asterisks) and 10^{20} (filled circles).

A statistical quantity for the eigenvalues of large Hermitian random matrices which has a particularly simple functional form is the two-point correlation function $\rho_2(s)$. This is defined as the density of eigenvalues a distance s from a particular eigenvalue. One has that

$$\rho_2(s) = 1 - \frac{\sin^2 \pi s}{(\pi s)^2}.$$

From this one computes the so called structure function

$$S(k) := 1 + \int_{-\infty}^{\infty} (\rho_2(x) - 1)e^{ikx} dx$$

$$= \begin{cases} \frac{|k|}{2\pi}, & |k| < 2\pi \\ 1, & |k| \geq 2\pi. \end{cases} \quad (11)$$

Montgomery proved that $S(k)$ for the Riemann zeros has exactly this functional form for $|k| < 2\pi$. Heuristic reasoning in [7] to give (11) for $|k| \geq 2\pi$ makes essential use of the smoothed form (8) for the asymptotic density of prime pairs (7).

Methods developed in the semi-classical analysis of chaotic billiards [1] gives much insight into the interplay between the Riemann zeros and the primes. Before finishing, I would like to mention too that random matrix theory has shed light not only on the distribution of the Riemann zeros, but also the value distribution of $\zeta(s)$ itself on the critical line (see e.g. [8]).

Acknowledgement

My research is supported by the Australian Research Council.

References

- [1] M.V. Berry, J.P. Keating, *The Riemann zeros and eigenvalue asymptotics*, SIAM Rev. **41** (1979), 236–266.
- [2] E. Bombieri, *Problems of the millennium: the Riemann hypothesis*, <http://www.claymath.org/millennium/>.
- [3] J. Derbyshire, *Prime Obsession* (John Henry Press Washington D.C. 2003).
- [4] P.J. Forrester, A.M. Odlyzko, *Gaussian unitary ensemble eigenvalues and Riemann ζ function zeros: a non-linear equation for a new statistic*, Phys. Rev. E **54** (1996), 4493–4495.
- [5] P.J. Forrester, N.S. Snaith, J.J.M. Verbaarschot, *Developments in random matrix theory*, J. Phys. A **36** (2003), 1–10.
- [6] N. Katz, P. Sarnak, *Zeros of zeta functions and symmetry*, Bull. Amer. Math. Soc. **36** (1999), 1–26.
- [7] J.P. Keating, *The Riemann zeta function and quantum chaos*, in: *Quantum Chaos* (ed. I. Guarneri, U. Smilansky) (North Holland Amsterdam 1993), 145–185.
- [8] J.P. Keating, N.C. Snaith, *Random matrices and L-functions*, J. Phys. A **36** (2003), 2859–2881.
- [9] H.L. Montgomery, *The pair correlation of zeros of the zeta function*, Proc. Sympos. Pure Math. **24** (Amer. Math. Soc., Providence, R.I., 1973), 181–193.
- [10] A.M. Odlyzko, *The 10²⁰th zero of the Riemann zeta function and 70 million of its neighbours*, Preprint, 1989.
- [11] M. Rubinstein, *A simple heuristic proof of Hardy and Littlewoods conjecture B*, Amer. Math. Monthly **100** (1993), 456–460.
- [12] M. Rubinstein, P. Sarnak, *Chebyshev’s bias*, Experimental Math. **3** (1994), 173–197.
- [13] N.S. Witte, *Gap probabilities for double intervals in Hermitian random matrix ensembles as τ -functions — spectrum singularity case*, <http://arXiv.org/math-ph/0307063> (2003).
- [14] N.S. Witte, P.J. Forrester, *Gap probabilities in the finite and scaled Cauchy random matrix ensembles*, Nonlinearity **13** (2000), 1965–1986.

Department of Mathematics and Statistics, The University of Melbourne, VIC 3010

E-mail: P.Forrester@ms.unimelb.edu.au