An Analytic Approximation to the Density of Twin Primes

[1]Rodel B. Azura and [2]Dionisel Y. Regalado
[1]Agusan del Sur State College of Agriculture and Technology
[2]University of Science and Technology of the Philippines
rodelaz.geralden@gmail.com

Abstract

The highly irregular and rough fluctuations of the twin primes below or equal to a positive integer x ($x \leq 10^7$) are considered in this study. The occurrence of a twin prime on an interval [0,x] is assumed to be random. In particular, we considered the waiting time between arrivals of twin primes as approximated by a geometric distribution which possesses the discrete memory-less property. For large n, the geometric distribution is well-approximated by the exponential distribution. The number of twin primes less or equal to x will then follow the Poisson distribution with the same rate parameter as the exponential distribution. The results are compared with the Hardy-Littlewood conjecture on the frequency of twin primes. We successfully demonstrated that for large n, the proposed model is superior to the H-L conjecture in predicting the frequency of twin primes.

Keywords: twin primes, inter-arrival time

1.0 Introduction

Primes are building blocks of the numeration system. They are defined as positive integers $p \in \mathbb{Z}^+$ whose only factors are one and themselves. Every integer N can be expressed as products of primes:

$$N = P_1 P_2 \cdots P_n \quad (1)$$

Euclid proved that there are infinitely many primes by *reductio ad absurdum*. However, there are no known formulas to date that generate the prime numbers. Goldbach (1742), on the other hand, considered pairs $(P, P+2)$ of primes called twin primes and conjectured that there are also infinitely many twin primes. Attempts to prove the Goldbach conjecture by the same technique as Euclid have failed, and today, the Goldbach conjecture remains an open problem.

If $\pi(x)$ is the number of primes less or equal to x, the Prime Number Theorem, proved by de la Vallee Poussin and Riemann (1869), states that:

$$\pi(x) \sim \frac{x}{\log x} \quad (2)$$

In the case of twin primes, no such theorem exists but Hardy & Littlewood (1984) conjectured that:

$$\pi_{2(x)} \sim \frac{2cx}{(\log(x))^2} \quad (3)$$

where $c = 0.660161815\ldots$ is an irrational constant. Padua & Libao (2017) and Padua & Frias (2017) developed (3) by dynamical Systems and Martingale models, respectively. In this paper we derive an approximate density of the twin primes by using symbolic regression with primes below $N = 10^7$as data. We derived three (3) models for the density and compare their accuracy using the actual number of twin primes less or equal to x.

2.0 Basic Concepts

The concept of a limiting distribution in Statistics is an important feature that describes the distribution of random variables for large n.

Definition 1. Let $\varphi_k(x), k = 1,2,3,\ldots,n$ be a sequence of probability distribution functions for random samples of size k. If

$$\lim_{k\to\infty} \varphi_k(x) = \varphi(x) < \infty,$$

then the distribution $\varphi(x)$ is called the *limiting distribution of the sequence of distribution functions.*

If we require that $\varphi_k \to \varphi$ uniformly as $k \to \infty$, then $\varphi(x)$is the *ergodic or stationary*

distribution of the random variables. We note that ergodicity is a stronger concept than limiting distribution. In the context of twin primes, the distribution $\pi_{2k}(x)$ of twin primes possesses a limiting distribution but not an ergodic distribution because of non-stationarity i.e. both the mean number of twin primes and the variance are functions of k.*In this paper, we considered* $k = 10^7$ *as large enough to properly approximate the limiting distribution* $\varphi(x)$.

Another concept of importance in this paper involves stochastic processes, namely, a Renewal Process. The following definitions and results are found in standard textbooks in Stochastic Processes.

Definition 2. (Ross, 1984) An arrival process is an increasing sequence $0 < S_1 < S_2 < \cdots < S_n$ of positive random variables where $x_i = S_i - S_{i-1}$ are called *inter-arrival times* and $\{S_n : n \geq 1\}$ are called *arrival epochs.*

Definition 3. (Ross, 1984) The number of arrivals prior to and up to t,$\{ N(t): t \geq 0\}$ is called a counting process.

Theorem 1. The fundamental relationship between $N(t)$ and S_n is given by Ross in 1984:

$$\{S_n \leq t\} = \{N(t) \geq n\}$$

Definition 4. (Ross, 1984) A *renewal process*is an arrival process for which the inter-arrival time $X_1, X_2, \ldots$ are independent and identically distributed *(iid).*

Definition 5. (Ross, 1984) A *Poisson process* is a renewal process for which the inter-arrival times are exponentially distributed(λ):

$$f(x_i) = \lambda e^{-\lambda x_i}, x_i \geq 0.$$

Definition 6. (Ross, 1984) A random variable X is memoryless iff $x \geq 0$ and $P(x \geq t + x) = P(X \geq t) \cdot P(X \geq x)orP(X \geq t | X \geq x) = P(X \geq t)$.

Theorem 2. (Ross, 1984) A Poisson process is a memoryless process.

Theorem 3. (Ross, 1984) A random variable X is memoryless if and only if X is exponential(λ).

Theorem 4. (Ross, 1984) For a Poisson process with rate λ and for $t \geq 0$, the interval Z until the next arrival after $t$has an exponential distribution $F_z^{\,c}(z) = e^{-\lambda z}, \forall z > 0$.The random variable Z is independent of $N(t) = n$ and of all arrival epochs $S_1, S_2, \ldots, S_n$ and all arrivals $N(\tau) | 0 < \tau < t$.

Definition 7. (Ross, 1984) The Poisson counting process $N(t, t') = N(t') - N(t)$ for $t' > t$ has the property of stationary and independent increments:

$$N(t') - N(t) \sim N(t' - t).$$

Finally, we provide a brief description of *Symbolic Regression.* In traditional regression analysis, given ordered pairs $(x_i, y_i), i = 1,2,3, \ldots, n$, we fit a linear model of the form:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

where the $\beta's$ are unknown parameters to be estimated from the data. The term linear model is used to denote the fact that the model is linear in the unknown parameters. While the linear model is useful in most instances, its utility is limited when the actual functional relationship is non-linear in the unknown parameters. Thus, if the underlying relationship happens to be:

$$y_i = \cos(\beta_0 \beta_1) \exp(-x_i)$$

then the linear model will not be sufficient to discover such a relationship. Symbolic regression addresses this problem by assuming that:

$$y_i = f(x_i)$$

and leaving the functional form of $f(.)$ unspecified. A set of building blocks of functions is built into the computational algorithm e.g. logarithmic, exponential, trigonometric functions and others, and a genetic search is initiated. The genetic algorithm tries all possible combinations of building blocks that best fit the observed data. If the actual $f(x)$ is one of the combinations of these building blocks, symbolic regression through genetic search process will be able to discover such a functional relationship.

3.0 Results

Table 1. Shows the frequency distribution for the inter arrival time (t) and the frequency for each time t.

Table 1. *Frequency Distribution of the Inter-arrival time*

Time	Frequency	Frequency (%)	Time	Frequency	Frequency (%)	Time	Frequency	Frequency (%)
1	1	0.000122	22	78	0.009549	43	6	0.000735
2	863	0.105656	23	75	0.009182	44	6	0.000735
3	840	0.10284	24	61	0.007468	45	2	0.000245
4	783	0.095862	25	58	0.007101	46	4	0.00049
5	622	0.076151	26	58	0.007101	47	5	0.000612
6	550	0.067336	27	40	0.004897	48	1	0.000122
7	515	0.063051	28	44	0.005387	49	2	0.000245
8	494	0.06048	29	34	0.004163	50	3	0.000367
9	401	0.049094	30	29	0.00355	53	2	0.000245
10	354	0.04334	31	33	0.00404	54	1	0.000122
11	334	0.040891	32	29	0.00355	55	1	0.000122
12	306	0.037463	33	21	0.002571	58	1	0.000122
13	239	0.029261	34	14	0.001714	60	1	0.000122
14	203	0.024853	35	9	0.001102	61	1	0.000122
15	201	0.024608	36	15	0.001836	64	1	0.000122
16	207	0.025343	37	10	0.001224	65	1	0.000122
17	164	0.020078	38	5	0.000612	69	1	0.000122
18	131	0.016038	39	6	0.000735	72	1	0.000122
19	109	0.013345	40	3	0.000367	75	1	0.000122
20	101	0.012365	41	8	0.000979	103	1	0.000122
21	74	0.00906	42	4	0.00049			

Figure 1 shows the histogram of the actual frequency distribution for primes below.

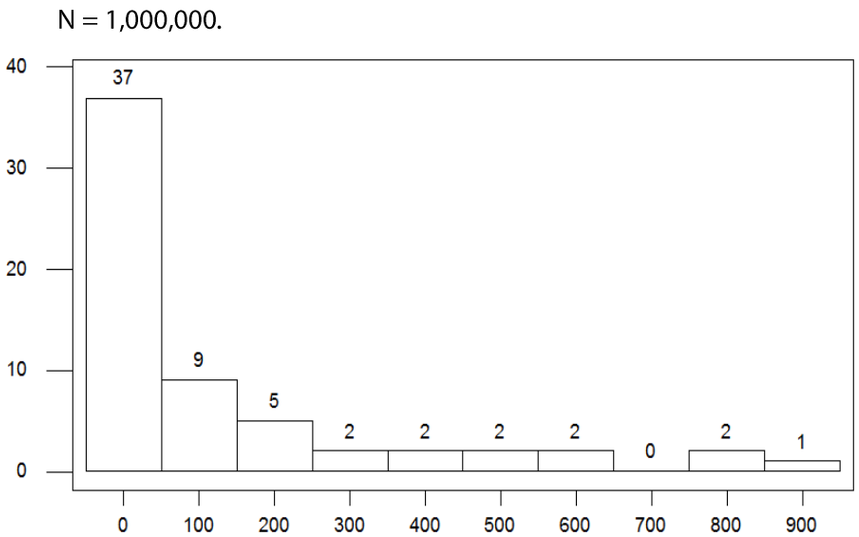

Figure 1. *Histogram of the actual frequency distribution*

Table 2 shows the approximate analytic densities derived by symbolic regression.

Table 2. *Approximate Analytic densities*

Model	R^2	Mean Absolute Error
$y = 0.155exp\left(\frac{-0.0874}{x} - 0.127x\right)$	0.995	.00087
$y = 0.156exp\left(\frac{-0.136}{x - 0.998} - 0.127x\right)$	0.997	.00076
$y = 0.155exp\left(\frac{-0.133}{x^2 - 1.98x + 0.998} - 0.126x\right)$	0.998	.000704

Figure 2 shows the plot s of the three (3) models with the actual frequency together with theHardy -Littlewood conjecture. The density of the inter-arrival times for the Hardy - Littlewood Conjecture is obtained as follows: Consider the appearance of a twin prime as a success. Then, the probability of observing a twin prime on the interval [0,x] is given by:

$$P\{twin\ prime\ on\ [0,x]\} = \frac{2c}{(\log(x))^2} = p.$$

The probability of not observing a twin prime is therefore, q = 1-p. Using the geometric distribution as the memory-less inter-arrival process, we obtain:

$$P\{waiting\ until\ Time\ T\ for\ the\ first\ success\} = q^{T-1}p.$$

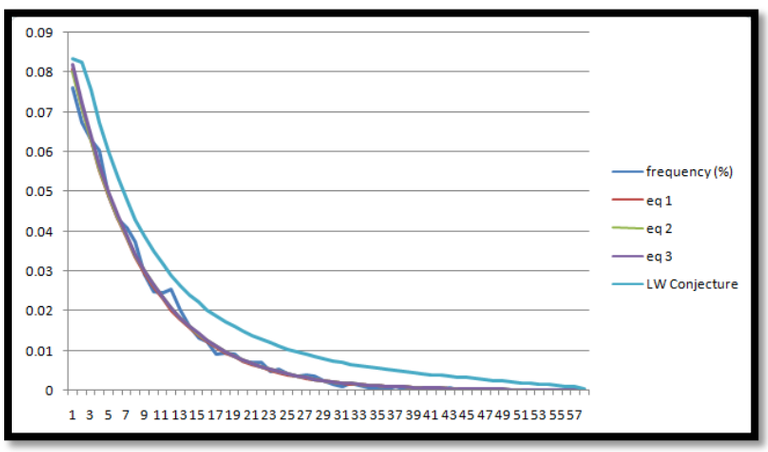

Figure 2. *Approximate Analytic Expression*

Table 3 shows a comparison of the errors of the three models based on the actual frequencies.

Table 3. *Mean-Absolute Errors of the Models Based on Actual Frequencies*

Model	Mean-Absolute Error
I	.00087
II	.00076
III	.000704
Littlewood Conjecture	.005161

Theorem: Let x be the random variable representing the inter-arrival times of twin primes on the interval $[0, N]$, $N \to \infty$. An exponential inter-arrival distribution with rate parameter λ: is almost surely equivalent (a.s) $f(x) = \lambda e^{-\lambda x}$, $x > 0, \lambda > 0$is equivalent to a geometric inter-arrival waiting time of Littlewood's conjecture if $\lambda = \frac{(\log N)^2 - 2c}{2c}$.

Proof: The maximum-likelihood estimator of λ is:

$$\hat{\lambda} = \frac{1}{\bar{x}}$$

and the expected value of x for an exponential waiting time is :

$$E(x) = \frac{1}{\lambda} \cong \frac{1}{\frac{1}{\bar{x}}} = \bar{x}$$

On the other hand, the expected value of x using Littlewood's conjecture is:

$$E(x) = \frac{p}{q} \text{ where } p = \frac{2c}{(\log N)^2}, q = 1 - p.$$

It follows that:

$$\begin{aligned} \bar{x} &= \frac{\left(\frac{2c}{(\log N)^2}\right)}{\left(1 - \frac{2c}{(\log N)^2}\right)} \\ &= \frac{2c \cdot (\log N)^2}{(\log N)^2 \cdot [(\log N)^2 - 2c]} \\ &= \frac{2c}{(\log N)^2 - 2c} \end{aligned}$$

Since

$$\bar{x} \cong \frac{1}{\lambda},$$

we obtain

$$\frac{1}{\lambda} \cong \frac{2c}{(\log N)^2 - 2c}$$

and

$$\lambda = \frac{(\log N)^2 - 2c}{2c} \text{ for } N \geq 5 \blacksquare$$

4.0 Conclusion

The number of twin primes less or equal to x will then follow the Poisson distribution with the same rate parameter as the exponential distribution. The results are compared with the Hardy-Littlewood conjecture on the frequency of twin primes. The paper demonstrated that for large n, the proposed model is superior to the Hardy – Littlewood conjecture in predicting the frequency of twin primes.

References

Azura, R. B., Tarepe, D. A., Borres, M. S., & Panduyos, J. T. (2017). The density of primes less or equal to a positive integer up to 20,000: Fractal approximation. *Recoletos Multidisciplinary Research Journal*, *1*(2).

Banks,W.D., Freiberg,T.,& Maynard, J. (2014). *On the limit points of the sequence of normalized prime gaps.* Retrieved from arXiv preprint arXiv:1404.5094

Hardy, G. H., & Littlewood, J. E. (1984). *Some problems of'Partitio Numerorum'; iii: on the expression of a number as a sum of primes.* In Goldbach Conjecture (pp. 21-60). Retrieved from: https://doi.org/10.1142/9789814542487_0002

Fliegel, Henry F.; Robertson, Douglas S. (1989). "Goldbach's comet: The numbers related to Goldbach's conjecture". *Journal of Recreational Mathematics*. 21 (1): 1–7.

Ford,K., Green,B., Konyagin,S., & Tao, T.(2014*). Large gaps between consecutive prime numbers.* Retrieved from arXiv preprint arXiv: 1408.4505

Maynard, J. (2013). Bounded length intervals containing two primes and an almost-prime. *Bulletin of the London Mathematical Society*, *45*(4), 753-764.

Maynard,J.(2013). Small gaps between primes. Retrieved from arXiv preprint arXiv: 1311.4600

Padua, R. N., & Libao, M. F. (2017). On stochastic approximations to the distribution of primes and prime metrics. *Journal of Higher Education Research Disciplines*, *1*(1), 33-38.

Regalado, D. & Azura, R.On a Mixed Regression Estimator for the Density of Prime Gaps. *Journal of Higher Education Research Disciplines.*1 (2), 48-57.

Sloane, N.J.A. (ed.). "Sequence A001692 (Number of irreducible polynomials of degree n over GF(5); dimensions of free Lie algebras)". *The On-Line Encyclopedia of Integer Sequences*. OEIS Foundation. *Retrieved 2011-02-02. -- A page of number theoretical constants.*

Young, J., &Potler, A. (1989). First occurrence prime gaps. *Mathematics of Computation*, 221-224.