Prime Gaps in the First One Million Primes: A Computational and Statistical Study

Gideon Afriyie

Abstract

This research presents a computational and statistical analysis of prime gaps in the first 1,000,000 prime numbers. Topics explored include twin primes, prime gap distributions, mode/mean/median comparisons, histogram analysis, Cramér's Conjecture, modulo patterns, quantile-quantile plots, empirical PDF and CDF, and logarithmic growth. The research uses Python, Excel, and abstract mathematics/number theory insights to uncover the structure of prime distribution. The project was inspired by coursework in abstract mathematics, which introduced the author to primes, Mersenne numbers, and Fermat primes, prompting a deeper summer investigation into patterns that define the fundamental atoms of arithmetic.

1 Introduction

In an abstract mathematics course, we studied the properties of primes, Mersenne numbers, and Fermat primes. While the theoretical discussions sparked my interest, the course didn't delve deeply into prime numbers or their statistical behaviors. Over the summer, I set out to investigate questions like: *Is there a predictable formula for primes? Are Fermat primes truly finite? How do gaps between primes behave?*

Through programming and statistical tools, I analyzed 1,000,000 primes and documented results from both classical conjectures and empirical observations.

Note: This research was conducted independently by the author. The pronoun *"we"* is used throughout in accordance with academic writing conventions.

2 Methods

- Generated the first 1,000,000 primes using Python.
- Calculated gaps between consecutive primes in Excel and Python.
- Conducted frequency analysis and computed:
  - Mean: 15.49
  - Median: 12

  - Mode: 6 (146,518 occurrences)
  - Max Gap: 154
  - Min Gap (excluding first): 2

- Plotted:
  - Histogram of prime gaps
  - Gaps vs. prime index
  - Logarithmic plots
  - Quantile-quantile (Q-Q) plots
  - Empirical PDF and CDF of gaps
  - Cramér's Conjecture

- Verified:
  - Cramér's Conjecture using $\log^2(p)$
  - Twin primes: 86,027 pairs
  - Gap modulo 6 patterns

Excel Formulas Used for Gap Analysis

The statistical summaries of the prime gaps were calculated using the following built-in Excel functions:

- **Prime Gaps:**

  Assuming column A contains the ordered list of primes starting from cell A2, and column B is used for gaps, enter in cell B3:

  ```
  =A3 - A2
  ```

  Drag down to compute all consecutive prime gaps.

- **Mean (Average Gap):**

  ```
  =AVERAGE(B3:B1000000)
  ```

- **Median Gap:**

  ```
  =MEDIAN(B3:B1000000)
  ```

- **Mode Gap (Most Frequent Gap):**

  ```
  =MODE.SNGL(B3:B100000)
  ```

- **Frequency Table:**

  Use `=COUNTIF(B:B, x)` where `x` is the specific gap value. Example: `=COUNTIF(B:B, 6)` counts how many times a gap of 6 occurs.

3 Key Findings

3.1 Prime Gaps

The distribution of gaps between primes was discrete and skewed. Excluding the first anomaly (1), all prime gaps were even and formed a uniform sequence of multiples of 2, from 2 through 154, reflecting the intrinsic parity constraints of prime distributions beyond 2.

Prime Gap Formula

Given two consecutive prime numbers p_n and p_{n+1}, the gap g_n is calculated as:

$$g_n = p_{n+1} - p_n$$

For example, the gap between 3 and 2 is:

$$g_1 = 3 - 2 = 1$$

Note on Final Prime Gap Calculation

In computing the gap between consecutive primes, we subtract each prime number from the one that follows it:

$$g_n = p_{n+1} - p_n$$

However, the final prime in our dataset, $p_{1,000,000}$, has no subsequent prime $p_{1,000,001}$ available in the list. Attempting to calculate:

$$g_{1,000,000} = p_{1,000,001} - p_{1,000,000}$$

results in an invalid or undefined expression, since $p_{1,000,001}$ does not exist in the data. In Excel, this often leads to a negative number or an error, as the missing cell may default to zero or blank.

To handle this, we excluded the final cell from our gap analysis calculations to avoid distortion of summary statistics like mean, mode, and median.

3.2 Histogram of Gaps

The histogram showed a peak at 6, with frequencies decreasing for larger gaps. This aligns with known patterns in prime distribution.

3.3 Central Tendencies

- Mean gap: ≈ 15.49
- Median gap: 12

- Mode gap: 6

The relatively small difference between mean and median suggests slight asymmetry. However, the low mode may indicate a skewed tail, prompting deeper distribution analysis.

3.4 Cramér's Conjecture

Cramér's Conjecture, proposed in 1936, is one of the most widely referenced models in probabilistic number theory. It suggests that the maximal gap g_n between consecutive primes p_n and p_{n+1} satisfies:

$$g_n = p_{n+1} - p_n = \mathcal{O}(\log^2 p_n)$$

In simpler terms, while gaps between primes increase, they should do so no faster than $\log^2(p_n)$ asymptotically.

In our project, we computed the observed prime gaps for the first one million primes and plotted them against $\log^2(p_n)$ for each corresponding p_n. The results showed a clear upper bound behavior: **no observed prime gap exceeded** $\log^2(p_n)$.

This empirical support, while not a proof, aligns with Cramér's heuristic model and provides additional visual evidence that the conjecture holds over a significantly large sample.

3.5 Twin Primes

Twin primes are pairs of primes that differ by exactly 2:

$$(p, p+2)$$

These are a special subclass of primes that exhibit tight clustering, especially among smaller primes.

In our study, we identified a total of 86,027 twin prime pairs among the first one million primes. While the frequency of twin primes declines as numbers increase, their presence remains consistent — aligning with the Twin Prime Conjecture, which states:

There exist infinitely many primes p such that $p + 2$ is also prime.

We further observed that twin primes cluster most densely in the lower range (first 100,000 primes) but taper slowly, rather than disappearing, suggesting persistence at higher magnitudes.

3.6 Gap Modulo 6

Most primes greater than 3 are of the form $6k \pm 1$, due to divisibility rules. That is:

$$p > 3 \Rightarrow p \equiv 1 \text{ or } 5 \mod 6$$

This structure led us to examine the gaps between consecutive primes modulo 6. We calculated:

$$g_n \mod 6 \quad \text{for each prime gap } g_n$$

Our findings revealed that the vast majority of prime gaps were divisible by 6 — particularly:

$$\text{Gaps} \equiv 0 \mod 6$$

These made up over **400,000** of all observed gaps. This is consistent with the structural distribution of primes and provides evidence for repeating modular behavior in prime spacing.

3.7 Logarithmic Behavior

To understand how primes grow with respect to their index, we explored the function $\log(p_n)$ versus n, the index of the prime. This investigation is closely tied to the Prime Number Theorem, which estimates the number of primes less than x as:

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

Plotting $\log(p_n)$ against n revealed an initially steep curve that gradually bends to the right — a behavior expected from the logarithmic nature of prime spacing.

Interpretation of the Log Curve

In the region where $\log(p_n) \approx 10$ to 12, we noted:

$$p_n \approx e^{10} \approx 22,000, \quad p_n \approx e^{12} \approx 162,754$$

This transition zone marks a noticeable curvature, showing that while $\log(p_n)$ continues to increase, its rate of growth slows — a hallmark of logarithmic behavior. This flattening effect supports the idea that prime gaps become wider at a logarithmic rate as n increases.

3.8 Q-Q Plot Analysis

Motivated by feedback from Professor Spayd, we conducted a quantile-quantile (Q-Q) analysis on the normalized prime gaps:

$$\text{Ratio} = \frac{g_n}{\log^2(p_n)}$$

We used this normalization to account for the logarithmic growth of gaps and compared the resulting distribution to a normal distribution.

The Q-Q plot revealed a prominent U-shape — with points deviating from the red reference line at both ends. This shape indicates significant deviation from normality, especially in the tails. Thus, the distribution of normalized prime gaps exhibits heavy tails and skewness, contrary to Gaussian assumptions.

This deeper statistical comparison aligns with the growing belief that while prime gaps grow, their distribution is far from normal, perhaps fitting heavy-tailed or exponential-like distributions.

3.9 Empirical PDF and CDF

To better understand the shape and frequency of prime gaps, we constructed:

- **PDF (Probability Density Function):** Shows the relative likelihood of various prime gaps.
- **CDF (Cumulative Distribution Function):** Indicates the cumulative proportion of gaps less than or equal to a given value.

PDF Findings: The histogram of gaps displayed a heavy right-skew, with a strong peak at 6. This reinforces the observation that smaller even gaps (like 2, 4, 6) dominate.

CDF Findings: The CDF rose steeply for gaps below 30, meaning a large proportion of gaps occur in this range. Beyond that, the CDF flattens, indicating sparsity among larger gaps — consistent with prime dispersion.

This dual perspective offered an intuitive and statistical confirmation of the tight clustering of small prime gaps and supported further exploration of their empirical behavior.

4 Tools Used

- Python: NumPy, matplotlib, seaborn, collections, csv
- Excel: Frequency tables, gap computations, early summaries
- Jupyter Notebook: Interactive analysis and plots
- LaTeX: Report generation

5 Conclusion

This research provides substantial computational and statistical support for several core ideas in analytic number theory. By analyzing the first one million prime numbers, we uncovered both expected patterns and new insights that enrich our understanding of prime behavior.

First, our results empirically support **Cramér's Conjecture**, as no observed prime gap exceeded the bound of $\log^2(p_n)$ across the entire data range. This reinforces the conjectured upper limit on prime gaps and aligns well with probabilistic models of prime distribution.

Secondly, our identification of **86,027 twin prime pairs** confirms that twin primes remain prevalent even at higher ranges, though they become less frequent. This sustained presence supports the **Twin Prime Conjecture** and offers a compelling case for further study of tight prime clustering.

Third, the analysis of **gap modulo 6** behavior revealed that the majority of gaps are divisible by 6, a result consistent with the structural form of primes ($6k \pm 1$). This modular regularity is a clear consequence of deeper divisibility constraints and reinforces known arithmetic properties of primes.

We also observed strong evidence of **logarithmic behavior** in prime growth. The curve of $\log(p_n)$ versus index flattens noticeably as p_n increases — an expected outcome based on the **Prime Number Theorem**, which describes the thinning density of primes among larger integers.

In terms of statistical centrality, we calculated the **mean gap** to be approximately 15.49, with a **median of 12**, and a clear **mode at 6**. This strong mode at 6 — consistent across different prime ranges — suggests a natural central tendency in prime gaps, even in the face of increasing variability at higher magnitudes.

Further analysis using a **Q-Q plot**, as recommended by Professor Spayd, revealed that normalized gaps do not follow a normal distribution. The U-shaped pattern indicated heavy tails and skewed behavior, suggesting that prime gaps follow a more complex distribution, possibly heavy-tailed or exponential in nature.

Finally, our construction of the **empirical PDF and CDF** of prime gaps provided an intuitive and visual understanding of gap concentration. The PDF confirmed right-skewness, while the CDF highlighted that most gaps fall below 30, confirming a dense cluster of small gaps.

Taken together, these findings demonstrate the power of combining classical theory with computational methods. This investigation not only supported long-standing conjectures but also laid the groundwork for future exploration into advanced topics such as prime gap distributions, density functions, and the deeper statistical nature of prime behavior at scale.

Acknowledgments

Special thanks to a Professor for insightful feedback on distributional comparisons, especially the use of Q-Q plots and deeper statistical visualization (PDF and CDF). Her support helped elevate the rigor of this investigation.

Future Directions

- Expand to 10 million primes
- Investigate the Hardy-Littlewood k-tuple conjecture
- Analyze prime gaps in blocks (e.g., every 100k)
- Test higher-order moments (skewness, kurtosis)

- Submit as an undergraduate poster or research article

References

[1] Cramér, H. (1936). "On the order of magnitude of the difference between consecutive prime numbers." *Acta Arithmetica* 2: 23–46.

[2] Hardy, G.H. and Wright, E.M. (2008). *An Introduction to the Theory of Numbers*. Oxford University Press.

[3] Granville, A. (1995). "Harald Cramér and the distribution of prime numbers." *Scandinavian Actuarial Journal*, 1: 12–28.

[4] Waskom, M. (2021). "Seaborn: Statistical data visualization." `https://seaborn.pydata.org`

[5] Apostol, T.M. (1976). *Introduction to Analytic Number Theory*. Springer.