1~1000000 之間質數間距分佈、等差數列結構分析

張家瑋

前言

質數作為數論中的基本元素，其重要性類似於原子在物理中的地位。它們看似孤立，但卻構成了整個數學世界的基礎結構。對質數的分布特性進行分析，不僅是純數學的研究問題，也與密碼學、隨機過程及計算數學等領域密切相關。

本研究聚焦於 1 至 1,000,000 之間的質數，系統性地探討相鄰質數間距（prime gaps）的分佈特性，並分析質數在等差數列中的結構規律。我們試圖回答以下問題：哪些等差數列更容易產生小間距？不同餘類下質數的分布是否存在規律性？以及如何透過統計方法量化質數分布的「聚散特性」？

透過對大規模質數數據的分析，本研究不僅提供對質數行為的新見解，也建立了「剩餘類 → 間距分布」的經驗規則，為後續數論研究提供量化方法與直觀工具。

總體而言，本研究希望通過數據分析和統計方法，深化對質數分布規律的理解，並為數論及相關應用領域提供新的研究視角和工具。

預先知識

在進行實證分析之前，需要對幾個關鍵定理有所了解。首先，狄利克雷定理（Dirichlet's theorem）【1】指出，對於任意互質（即 gcd(a,d)=1 ）正整數 a 與 d，數列：

$$a, a+d, a+2d, a+3d, ...$$

裡面存在無窮多個質數。這意味著只要 a 與 d 互質，這個數列就保證有無窮多個質數出現。 為研究質數在模 m 下的餘類分布提供了理論基礎。透過這一理論，我們可以合理地對質數進行餘類分類，分析不同餘類下的質數間距分布，而非僅依賴經驗觀察。其次，Green–Tao 定理（Green–Tao theorem）【2】指出，在質數中存在任意長度的等差數列。該定理揭示，質數不僅隨機分布，同時存在高度規律的結構，支持本研究對特定模數下質數間距規律的探索。這些理論使我們能夠將統計分析與數論理論相結合，從而更全面地理解質數的內在規律。而在此定理中也可以發覺，等差數列的公差都是奇數，因為若公差=d，第一項是奇數 p_1，第二項=p_1+d 會是偶數，因為奇數+奇數=偶數（不是質數除了 2 以外），又因 p_1+d=奇數+偶數=奇數，p_1+2d=奇+2*偶數=奇數，此時都是奇數，所以可能就是質數。

研究方法

為了分析 1 到 1,000,000 之間質數的分布，我採用經典的 埃拉托斯特尼篩法（Sieve of Eratosthenes）【3】 生成質數序列。透過這一方法，可以有效且精確地獲取大範圍內的質數集合，作為後續統計分析的基礎。

程式碼：

def sieve_of_eratosthenes(n): # 建立一個布林陣列，初始假設全部都是質數

sieve = [True] * (n + 1) sieve[0] = sieve[1] = False # 0 和 1 不是質數

p = 2

while p * p <= n: # 只需要篩到 sqrt(n)，所以如果 n 是一百萬，只需要到一百萬的平方根=1000

    if sieve[p]: # 如果 p 還是質數

        # 把 p 的倍數全部設為 False (從 p*p 開始)

        # rnage 的參數是 start（開始，包含這個值),stop（結束，不包含這個值),step（每次增加多少）

            for multiple in range(p * p, n + 1, p):

                sieve[multiple] = False

```python
        p += 1

    # 回傳所有質數
    return [i for i, is_prime in enumerate(sieve) if is_prime]


n = 1_000_000
primes = sieve_of_eratosthenes(n)
print(f"1 到 {n} 的質數共有 {len(primes)} 個")
print("前 20 個質數:", primes[:20])
print("最後 20 個質數:", primes[-20:])
```

輸出：

1 到 1000000 的質數共有 78498 個
前 20 個質數: [2, 3, 5, 7, 11, 13, 17, 19, 23, 29, 31, 37, 41, 43, 47, 53, 59, 61, 67, 71]
最後 20 個質數: [999671, 999683, 999721, 999727, 999749, 999763, 999769, 999773, 999809, 999853, 999863, 999883, 999907, 999917, 999931, 999953, 999959, 999961, 999979, 999983]

接著開始對 mod 進行分類：

mod 6 分類

```
mod6 = {1: [], 5: []}

for p in primes:

    if p > 3:

        r = p % 6

        if r in mod6:

            mod6[r].append(p)
```

mod 30 分類

```
mod30_rems = [1,7,11,13,17,19,23,29]

mod30 = {r: [] for r in mod30_rems}

 for p in primes:

    if p > 5:

         r = p % 30
```

```
        if r in mod30:
            mod30[r].append(p)

印出每個分類的質數個數

print("mod 6  分類數量：", {k: len(v) for k,v in mod6.items()})

print("mod 30  分類數量：", {k: len(v) for k,v in mod30.items()})
```

輸出：

mod 6 分類數量： {1: 39231, 5: 39265}

mod 30 分類數量： {1: 9807, 7: 9812, 11: 9810, 13: 9824, 17: 9809, 19: 9788, 23: 9840, 29: 9805}

等差數列分析

我們知道在質數中存在任一長度的等差數列，在研究中對質數進行了模數分類，這裡以 Mod 30 為例，將質數按照與 30 互質的餘數進行分組。這些餘數分別為 [1, 7, 11, 13, 17, 19, 23, 29]。對每個餘數類，我們統計了該類質數的數量，並計算其相鄰質數之間的差值（gap)，以分析其在等差數列結構下的分布特徵。

具體數據如下：

- **餘數 1**：質數數量 9807，前 10 個 gap 示例為 [30, 90, 30, 30, 30, 30, 60, 90, 120, 30]
- **餘數 7**：質數數量 9812，前 10 個 gap 示例為 [30, 30, 30, 30, 30, 120, 30, 30, 30, 30]
- **餘數 11**：質數數量 9810，前 10 個 gap 示例為 [30, 30, 30, 30, 60, 60, 30, 30, 90, 30]
- **餘數 13**：質數數量 9824，前 10 個 gap 示例為 [30, 30, 30, 60, 30, 30, 60, 30, 60, 60]
- **餘數 17**：質數數量 9809，前 10 個 gap 示例為 [30, 60, 30, 30, 30, 30, 30, 60, 30, 120]
- **餘數 19**：質數數量 9788，前 10 個 gap 示例為 [60, 30, 30, 60, 30, 120, 30, 30, 30, 60]
- **餘數 23**：質數數量 9840，前 10 個 gap 示例為 [30, 30, 30, 60, 60, 30, 30, 60, 30, 60]
- **餘數 29**：質數數量 9805，前 10 個 gap 示例為 [30, 30, 60, 30, 60, 30, 90, 30, 30, 30]

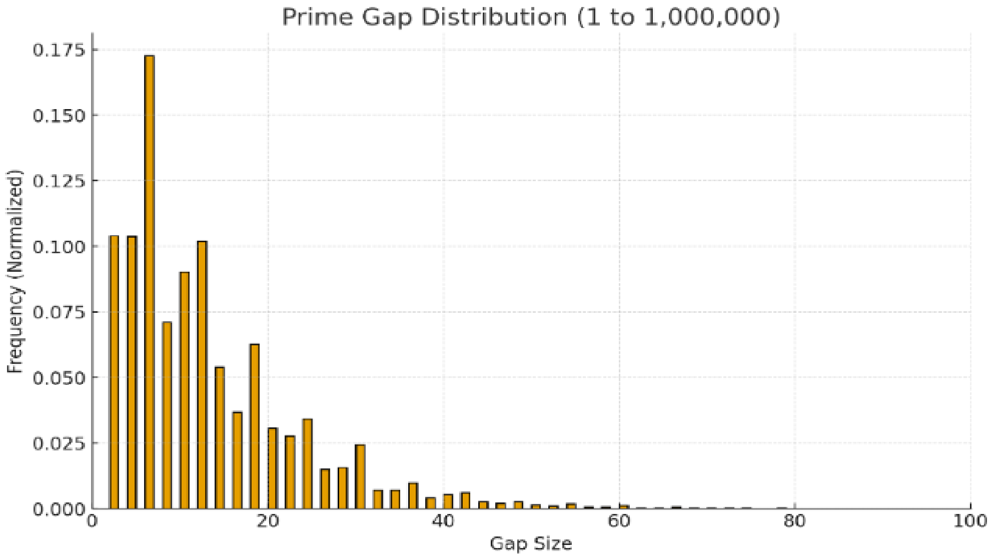
從上述資料中可以觀察到，雖然每個餘數類的質數數量大致相近，但 gap

分布呈現明顯差異。例如，餘數 7 的 gap 多為 30，說明質數在該餘數類中更為「擠在一起」，而餘數 19 則存在較多的 60 或 120 gap，分布較為分散。這表明，在 Mod 30 的等差數列結構下，質數的排列既有規律性，也存在局部波動。

所以這裡我們就可以說：「mod 30 的剩餘類 r=7，比其他類更容易產生小 gap。」

小 Gap

在這裡，定義了一個小 gap，這個小 gap 的意思是在一剩餘類 r 的分佈中，後一項減去前一項的差，例如 6k+1 當 k=1 的時候商會是 7，當 k=2 的時候商會是 13，此時小 gap 就是 13-7=6，或 30k+1 當 k=1 的時候商是 31，當 k=2 的時候商是 61，此時小 gap 就是 61-31=30，而有趣的是如果現在不考慮這種剩餘類 r 的等差數列的話，單純從質數之間的 gap 來來計算，可以知道一件事，即所有質數的 gap 中最常出現的會是 6，因為除了 2 和 3，所有質數都 ≡ 1 或 5 (mod 6)，換句話說，質數一定落在「6k ± 1」的位置，例如 5=6k-1，7=6k+1，11=6k-1,13=6k+1，這個限制來自於：如果一個數字是 6k, 6k±2, 6k+3，它必定能被 2 或 3 整除，所以不可能是質數。



回到餘數類 r 的等差數列分佈，在此研究中可以發現一種規律，即模數越大 → 等差序列越稀疏 → 相鄰質數 gap 越大 → 小 gap 出現越少 → gap 分布越不規律，

============================

參數設定

============================

N = 10000

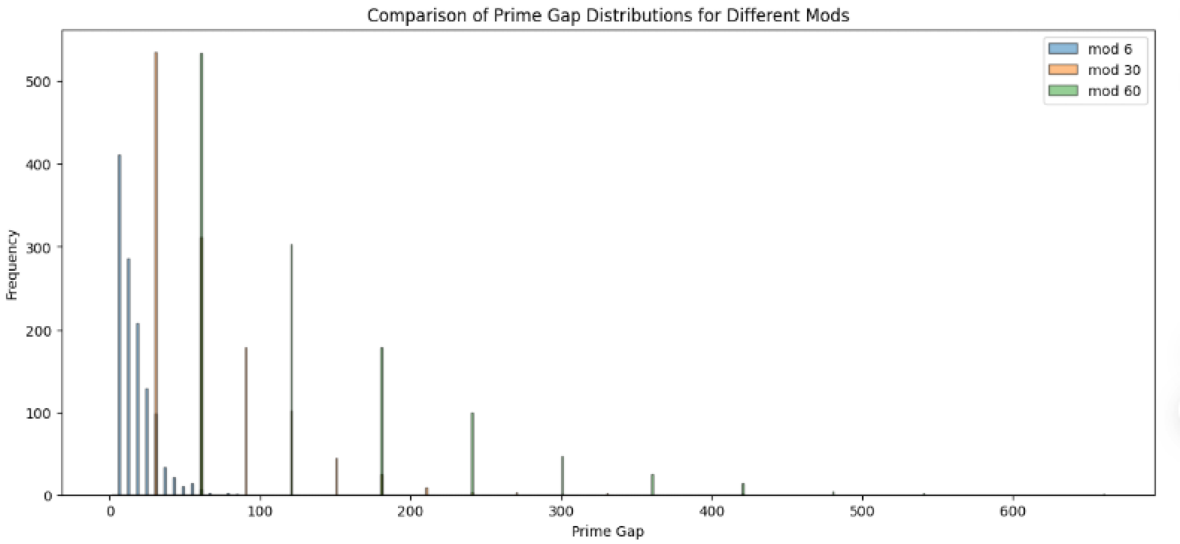
mods_to_plot = [6, 30, 60]

important_gaps = None

============================

```
生成質數序列

==============================

primes = list(primerange(2, N))

  print(f"總質數數量: {len(primes)}")

==============================

計算 gap 分布

==============================

gap_data_per_mod = {}

for m in mods_to_plot:

        coprime_remainders = [r for r in range(1, m) if gcd(r, m) == 1]

          all_gaps = []

        for r in coprime_remainders:

                seq = [p for p in primes if p % m == r]

                gaps = [seq[i+1] - seq[i] for i in range(len(seq)-1)]

                all_gaps.extend(gaps)

        gap_data_per_mod[m] = all_gaps

==============================

畫 histogram 對比

==============================

plt.figure(figsize=(14,6))
```

```
for m in mods_to_plot:
    gaps = gap_data_per_mod[m]
    plt.hist(gaps, bins=range(1, max(gaps)+2), alpha=0.5, label=f'mod {m}', edgecolor='black')
plt.xlabel('Prime Gap')
plt.ylabel('Frequency')
plt.title('Comparison of Prime Gap Distributions for Different Mods')
plt.legend()
plt.show()
```



在研究等差數列的結構中，其中發現了一條規律，關乎於小 gap 定義與模數之間的關係，設一序列需要比較的模數數數列裡面，例如：1~10 或 11~20，此時小 gap 定義=10 或 20 的話，也就是小 gap 是數列裡面的最大數，那麼對於模數>6 且是奇數的時候，沒有任何一個類 r 的 gap 分佈是<

小 gap 的，此時小 gap 就是模數數列的上確界（supremum），且如果小 gap 定義=1 或 11 的話，也就是小 gap 是數列裡面的最小數，可以得知沒有任何一個類 r 的 gap 分佈是<小 gap 的，那麼此時的小 gap 定義就是模數數列的下確界（infimum），那麼是否所有的模數>6 且是奇數的時候，沒有任何一個類 r 的 gap 分佈是<小 gap 的這件事情（意指無窮質數），還是需要有待觀察驗證的。

有趣的是，如果此時的小 gap 定義脫離了模數數列的話，上確界的約束效果就會失效，所以加入我們有一個 2~30 的模數數列，如果小 gap 定義為 20 的話，在 11~20 之間能夠遵循上確界的約束，但是在 2~10 之間，這種約束效果就消失了，而 21~30 自然的不會有小於小 gap 的類 r 的 gap 分佈存在。

質數在不同模數 (mod m) 下的**餘數類（r）之間的 gap 分佈特性**

在這裡我們取所有小於 1,000,000 的質數，分別對模數進行比較進行比較，若兩個連續質數之差 ≤ g_th（我們定義的小 gap），就記為「小 gap」。

定義：

$$R_m = \{r \in \{1,2, \ldots, m-1\} \mid \gcd(r, m) = 1\}$$

對每個餘數類 r∈R_m ，定義質數序列：

$$P_r = \{ p_i \mid p_i \; is\, prime\, number\, and\, p_i \bmod m = r \}$$

gap 序列：

$$G_r = \{ g_i = p_{i+1} - p_i \mid p_i , p_{i+1} \in P_r \}$$

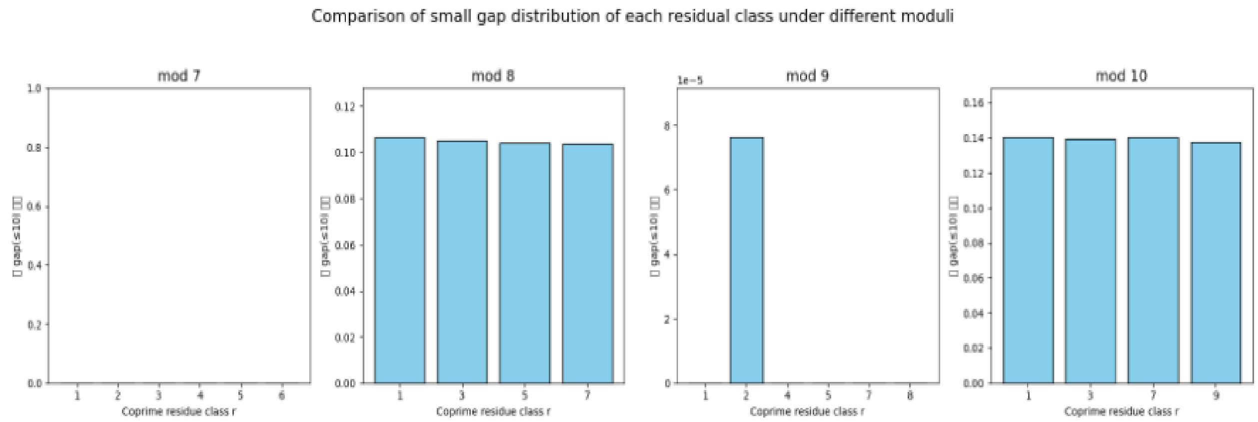
小 gap 比例：

$$ratio_r = \frac{|\{ g_i \in G_r \mid g_i \leq g_{th} \}|}{|G_r|}$$

其中 g_th = small_gap_threshold。

最常出現的 gap：

$$\dot{g_r} = \arg \max \; count(g)_{g \in G_r}$$



在上圖中，g_th=10，可以看到小 gap 比例在不同的餘數類 r 底下是大致平均的，而 mod > 6 且是奇數的條件下，沒有餘數類 r 的 gap < g_th 的條件也成立，而有些時候在 mod > 6 且是奇數的條件下，會出現 r=2 的異常值情況。

而在此研究中，也發現對於相同模數下但不同餘數類 r 的情況下，最常出現的小 gap 數字都是一樣的。

模 7 結果：

r= 1: 小 gap 比例=0.0000, 最常出現 gap=42 出現 2609 次

r= 2: 小 gap 比例=0.0000, 最常出現 gap=42 出現 2571 次

r= 3: 小 gap 比例=0.0000, 最常出現 gap=42 出現 2637 次

r= 4: 小 gap 比例=0.0000, 最常出現 gap=42 出現 2566 次

r= 5: 小 gap 比例=0.0000, 最常出現 gap=42 出現 2602 次

r= 6: 小 gap 比例=0.0000, 最常出現 gap=42 出現 2609 次

模 8 結果：

r= 1: 小 gap 比例=0.1067, 最常出現 gap=24 出現 3303 次

r= 3: 小 gap 比例=0.1052, 最常出現 gap=24 出現 3413 次

r= 5: 小 gap 比例=0.1043, 最常出現 gap=24 出現 3373 次

r= 7: 小 gap 比例=0.1038, 最常出現 gap=24 出現 3368 次

模 9 結果：

r= 1: 小 gap 比例=0.0000, 最常出現 gap=18 出現 2729 次

r= 2: 小 gap 比例=0.0001, 最常出現 gap=18 出現 2692 次

r= 4: 小 gap 比例=0.0000, 最常出現 gap=18 出現 2737 次

r= 5: 小 gap 比例=0.0000, 最常出現 gap=18 出現 2712 次

r= 7: 小 gap 比例=0.0000, 最常出現 gap=18 出現 2745 次

r= 8: 小 gap 比例=0.0000, 最常出現 gap=18 出現 2836 次

模 10 結果：

r= 1: 小 gap 比例=0.1401, 最常出現 gap=30 出現 4098 次

r= 3: 小 gap 比例=0.1393, 最常出現 gap=30 出現 4070 次

r= 7: 小 gap 比例=0.1404, 最常出現 gap=30 出現 4076 次

r= 9: 小 gap 比例=0.1374, 最常出現 gap=30 出現 4038 次

這裡可以發現兩個事實：同一餘數類的相鄰質數之差必然是模數的倍數，（因為如果$p \equiv r \ (mod\ m)$且下一個同類的質數$q \equiv r \ (mod\ m)$，那$q - p \equiv 0 \ (mod\ m)$)，例如：

$$m = 6$$

假設取兩個質數：

$$p = 11, q = 17$$

各自除以 m=6:

$$11 \div 6 = 1 \dots 5$$
$$17 \div 6 = 2 \dots 5$$

兩個質數的餘數都是 5，因此它們屬於同一餘數類 r = 5

$$q - p = 17 - 11 = 6$$
$$q - p \equiv 0 \ (mod\ 6)$$

如果兩個質數在模 6 下餘數相同（都等於 5），那麼它們的差一定是 6 的倍數。

另外一點是關於從 6 開始的結構限制：除去 2 和 3 以後的質數都落在 {1, 5} (mod6)（等價地，所有大於 3 的質數都與 6 互質），因此任兩個大

質數的差會是 6 的倍數（也就是會同時被 2 與 3 整除）。

任何整數除以 6，餘數只能是：

$$Z_6 = \{0, 1, 2, 3, 4, 5\}$$

質數不可以被 2 或 3 整除，所以餘數是 0, 2, 3, 4 的數都會被 2 或 3 整除（不是質數），所以剩下能成為質數的餘數只有(1, 5)。

如果兩個質數都在同一餘數類：

$$p \equiv 1 \ (mod\ 6),\ q \equiv 1 \ (mod\ 6)$$

那它們的差：

$$q - p \equiv 0 \ (mod\ 6)$$

差是 6 的倍數，舉個例子 5, 11, 17, 23, 29, 35, 41，取其中兩個相減都是 6 的倍數，如果質數大於 3 且取兩個相同餘數類（都餘 1 或都餘 5）的質

數，它們的差會是 6 的倍數。

對 mod11 來說：lcm(11,6)=66，所以同一餘數類中，可以出現的最小 gap 就是 66。對 mod13：lcm(13,6)=78，因此最常出現的 gap 很自然會是 78。對 mod12：lcm(12,6)=12，所以最常見的 gap 是 12。

而每個類裡的 gap 分布、甚至最常見的 gap 值都差不多，這裡即是 Dirichlet 定理。

若 a 與 m 互質，則算術級數 a, a+m, a+2m, a+3m, ...裡一定有無限多個質數。

並且當你考慮所有與 m 互質的餘數類時（例如 mod 10 的 1, 3, 7, 9），它們裡的質數出現頻率在「無窮大」的極限下是平均分配的。

$$\pi(x;\, m,\, r) \;\approx\; \frac{1}{\varphi(m)}\pi(x)$$

mod =10, 與 10 互質的餘數類是 1, 3, 7, 9,所以$\varphi(10) = 4$。

$$\pi(x;\ 10,\ 1)\ \approx\ \pi(x;\ 10,\ 3)\ \approx\ \pi(x;\ 10,\ 7)\ \approx\ \pi(x;\ 10,\ 9)\ \approx\ \frac{1}{4}\pi(x)$$

$\pi(x;\ m,\ r)$ 表示小於等於 x 的質數中，餘 r 的個數，$\pi(x)$ 是小於等於 x 的所有質數個數，$\varphi(m)$是 Eulerφ 函數，代表與 m 互質的整數數量。這就像是所有質數像是被平均灑在各個模 m 的互質餘數類裡。當 mod10，餘 1 的質數約佔 25%，餘 3 的質數同樣約是 25%，7 和 9 同理，當取樣到一百萬以內的質數時，雖然還有些微隨機起伏，但已經很接近均勻分布。

根據 Dirichlet 定理，當質數越大，它們在每個互質餘數類裡的分布會越平均，所以同模數下不同 r 的質數差（gap）行為也會越相似，也是最常出現的 gap 一樣、出現次數接近的根本原因。

Iota

在這裡引入了經驗隱數 ι _m,r 來量化特定模數 m 下互質餘數類 r 的小 gap 行為相對於全域平均的偏差。具體而言，對於模 m 下的餘數類 r, 其

小 gap 比例定義為 p_m,r ，表示該餘數類相鄰質數差值不超過預設閾值 small_gap_threshold 的比例。同時，我們計算全域小 gap 比例 p_m,global 作為所有互質餘數類小 gap 比例的平均值，進而定義隱數：

$$p_{m,global} = \frac{1}{R}\sum_{r} p_{m,r}$$

$$\iota_{m,r} = p_{m,r} - p_{m,global}$$

此指標直觀反映了該餘數類與整體趨勢之間的相對偏離：當 ι_m,r > 0 時，餘數類 r 相對全域更容易出現小 gap；反之， ι_m,r < 0 則顯示該類質數之間的 gap 普遍偏大，幾乎不出現小 gap。值得注意的是，對於全域平均小 gap 比例本身極小的情況， ι_m,r 小於 1e-4 的時候（有時候在 r=2 時，會是 1e-3），即可有效推斷該餘數類大部分相鄰質數差值均大於所定義的閾值。換言之，經驗隱數 ι_m,r 不僅提供了對特定餘數類小 gap 出現概率的相對量化，更可作為判斷其 gap 分布特性的重要指標。此方法在分析不同模數下質數分布結構及其局部行為差異時，具有直觀且可操作的統計意義。

符號	意義
(m)	模數（modulus）
(r)	餘數類（remainder class）
(R)	有效餘數類的數量（通常 = m，但可能少）
($p_{m,r}$)	餘數 r 下的「小 gap 比例」
($p_{m,global}$)	所有有效餘數類的平均小 gap 比例
($\iota_{m,r}$)	該餘數相對於平均的偏差

這就像統計裡的「樣本值 − 平均值 = 殘差（residual）」。如果 $\iota_{m,r} > 0$:這個 r 比平均更常出現小 gap，如果 $\iota_{m,r} < 0$:這個 r 比平均更少出現小 gap，所有 $\iota_{m,r}$ 加起來的平均=0。

ι 的作用有兩個層面：

相對擠度。 $\iota > 0 \rightarrow$ 這個餘數類的質數相對其他餘數類更「擠」，並且小 gap 出現頻率高， $\iota < 0 \rightarrow$ 相對鬆散，並且小 gap 出現頻率低。

小 gap 出現的絕對情況：如果整個模數的 small-gap 平均 很大 → 表示整個模數下小 gap 很常見，如果整體平均很小 → 表示小 gap 很少， ι 則告訴你某個餘數類的情況相對於平均是高還是低。

引用

【1】 Dirichlet, G.L. (1837) Beweis des Satzes, dass jede unbegrentze arithmetische Progression, deren erstes Glied und Differenz ganze Zahlen ohne gemeinschaftlichen Factor sind, unendlich viele Primzahlen enthält. Abhandlungen der Königlichen Preussischen Akademie der Wissenschaften zu Berlin, 48, 45-71.

【2】 Green, B., & Tao, T. (2004). The primes contain arbitrarily long arithmetic progressions. *Annals of Mathematics*, 167(2), 481 – 547.

【3】 Eratosthenes of Cyrene (ca. 200 BCE). Sieve of Eratosthenes, described in Nicomachus of Gerasa’ s *Introduction to Arithmetic*.