哥德巴赫猜想研究

哥德巴赫猜想说，一个足够大的偶数可以表示成两个素数之和。

关于素数分布的已知规律

素数定理

当N趋于无穷时，小于N的素数数量，可以近似表示为$\frac{N}{\log(N)}$；对数积分$\mathrm{li}(x)=\int_0^x \frac{\mathrm{d}x}{\log x}$是对$\pi(x)$的更好的估计。

或者等价地说，N附近的素数密度约为$\frac{1}{\log(N)}$。

数值结果

任何小于4×10^{18}的偶数都经过计算机程序验证符合哥德巴赫猜想。

猜测

哈代-李特尔伍德猜测说，大整数N可以表示成的2个素数之和的方式可以近似表示为

$$2\Pi_2\left(\prod_{p|N;p\geq 3}\frac{p-1}{p-2}\right)\frac{N}{\left(\log(N)\right)^2}$$

其中$\Pi_2\approx 0.6601\ldots$是哈代-李特尔伍德的孪生素数常数。

从直觉上说，对于一个随机数x，若$1\leq x\leq N$，则x是素数的概率约为$\frac{1}{\log(N)}$；$N-x$是素数的概率同样约为$\frac{1}{\log(N)}$；若把这两个事件近似看作独立事件，则它们同时发生的概率约为$\frac{1}{(\log(N))^2}$。由此得到N表示为2个素数之和的方式，应该约等于$\frac{N}{(\log(N))^2}$。这在数量级上，跟哈代-李特尔伍德猜测相吻合。

埃拉托斯特尼筛法

埃拉托斯特尼筛法是一种古老的求出N以内的所有素数的方法。

它的过程如下

1. 按顺序写下2至N的整数
2. 划掉2的倍数（除了2本身）
3. （下一个没有划掉的数是3）划掉3的倍数（除了3本身）
4. （下一个没有划掉的数是5）划掉5的倍数（除了5本身）
5. 重复以上步骤
6. 当进行到$\sqrt{N}$时，剩下的数就已经全都是素数了。

埃拉托斯特尼筛法的变形，求哥德巴赫猜想的解的近似数量

1. 写下2至$N-2$的整数
2. 划掉所有偶数
3. 划掉所有3的倍数；若N除以3的余数是r，划掉所有$3k+r$形式的数
4. 划掉所有5的倍数；若N除以5的余数是r，划掉所有$5k+r$形式的数
5. 划掉所有7的倍数；若N除以7的余数是r，划掉所有$7k+r$形式的数
6. 划掉所有11的倍数；若N除以11的余数是r，划掉所有$11k+r$形式的数
7. 重复以上步骤
8. 当进行到$\sqrt{N}$时，剩下的数，一定满足哥德巴赫猜想的条件。

举例

若$N=42$,
则经过步骤2，剩下的数为$3,5,7,9,11,13,15,17,19,21,23,25,27,29,31,33,35,37,39$
经过步骤3，剩下的数为$5,7,11,13,17,19,23,25,29,31,35,37$
经过步骤4，剩下的数为$11,13,19,23,29,31$
下一个素数已经大于$\sqrt{N}$，停止

因为上述过程筛掉了所有的$\sqrt{N}$内的素数的倍数，剩下的必然都是素数
又因为上述过程的对称性，因此对于每一个剩下的数x，$N-x$也一定存在其中；

所以剩下的数（如果数量不为0），一定是对于偶数N的哥德巴赫猜想的解。

注意，上述过程会筛掉一部分合乎规则的解。比如对于$N=42$，N可以表示为$5+37$，而这个解被筛掉了。
故上述方式，可能会低估哥德巴赫问题的解的数量。

估算数量

以上的每一个步骤，都可以用一个简单的方式来估算筛掉的数的数量

1. 一开始的数的数量是$N-3$
2. 划掉所有偶数——划掉的数约为剩下的数的$\frac{1}{2}$

3. 划掉3的倍数；划掉$3k+r$形式的数——若N是3的倍数，划掉的数约为剩下的数的$\frac{1}{3}$；若N不是3的倍数，划掉的数约为剩下的数的$\frac{2}{3}$。
4. 划掉5的倍数；划掉$5k+r$形式的数——若N是5的倍数，划掉的数约为剩下的数的$\frac{1}{5}$；若N不是5的倍数，划掉的数约为剩下的数的$\frac{2}{5}$。

   …………

由此可以得到一个估算公式

$$\left(\prod_{p\le\sqrt{N}}(1-\frac{F(p,N)}{p})\right)(N-3)$$

其中，$F(p,N)$的定义为，“若$p\mid N$，则$F(p,N)=1$；否则$F(p,N)=2$”。

比如，当N是42时，上述公式给出的估计是7.8（上述过程的结果是6，真实的解数量为8）
当N是10^6时，上述公式给出的估计是11541.7（上述过程的结果是10764，真实的解数量为10804）

当N是奇数时，上述公式给出的估计是0。

估算素数数量

上面的逻辑，也可以用来估算$\pi(N)$。

$$\pi(N)\approx\left(\prod_{p\le\sqrt{N}}\left(1-\frac{1}{p}\right)\right)N+\pi(\sqrt{N})$$

当用每一个素数p去筛剩余的数时，可以预计剩下的数近似变为$1-\frac{1}{p}$。这个连乘去掉了所有小于$\sqrt{N}$的素数，所以最后加上$\pi(\sqrt{N})$。

比如，当$N=100$时，上述公式给出的结果是26.86，而$\pi(100)=25$。
当$N=10^6$时,上述公式给出的结果是81133.26，而$\pi(10^6)=78498$。
当$N=10^8$时,上述公式给出的结果是6089698.25，而$\pi(10^8)=5761455$。
当$N=10^{10}$时,上述公式给出的结果是487538770.51，而$\pi(10^{10})=455052511$。

上述估算公式的误差分析

素数个数

上述公式估算素数数量时，隐含用到一个假设，“‘x是p_1的倍数’、‘x是p_2的倍数’……是独立事件”。若这些事件相互之间并不独立，则计算数量时，不能直接用比例相乘。事实上，它们确实不是独立事件。因此上述公式有误差是可以预计的。

通过数值方法，可以直接计算出误差的量级。

N	$\pi(N)$	估算结果	相对误差
10^2	25	26.86	-6.9%
10^3	168	163.85	2.5%
10^4	1229	1228.17	-0.067%
10^5	9592	9716.94	-1.29%
10^6	78498	81133.26	-3.25%
10^7	664579	696459.96	-4.58%
10^8	5761455	6089698.25	-5.39%
10^9	50847534	54170223.35	-6.13%
10^{10}	455052511	487538770.51	-6.66%
10^{11}	4118054813	4433073529.48	-7.11%
10^{12}	37607912018	40638288669.65	-7.46%

哥德巴赫猜想解的个数

通过数值方法，也可以得出哥德巴赫猜想的解的个数

N	筛得的结果个数	估算结果	相对误差	比例
10^2	10	9.24	8.25%	-1.20
10^3	48	41.27	16.31%	6.52
10^4	250	255.24	-2.05%	30.50
10^5	1600	1640.72	-2.48%	1.92
10^6	10764	11541.72	-6.74%	2.07
10^7	77526	85284.53	-9.10%	1.99
10^8	582562	652588.80	-10.73%	1.99
10^9	4547836	5165199.09	-11.95%	1.95
10^{10}	36399500	41842797.05	-13.01%	1.95

上表添加了“比例”这一列，用于比较这两种公式的相对误差。
可以看到，当N的值很大时，估算素数个数的公式和估算哥德巴赫猜想的解的个数的公式，都倾向于多估；
并且这两者的相对误差的比例，慢慢稳定在2附近。

梅滕斯定理

梅滕斯第三定理说，当N很大时，$\prod_{p\le N}(1-\frac{1}{p})$近似为$\frac{e^{-\gamma}}{\log N}$，其中$\gamma$是欧拉-马斯刻若尼常数。

将其应用于$\left(\prod_{p\le\sqrt{N}}\left(1-\frac{1}{p}\right)\right)N$，得到一个估算公式

$$\left(\prod_{p\le\sqrt{N}}\left(1-\frac{1}{p}\right)\right)N\approx\frac{2e^{-\gamma}N}{\log N}$$

其中，$2e^{-\gamma}\approx 1.122918967$

但是根据素数定理，更准确的估算公式应该是$\frac{N}{\log N}$

比较这两个公式，可以预计，当N趋于正无穷时，上述公式的“相对误差”应该是$\frac{1}{2e^{-\gamma}}-1\approx 10.946\%$

即，随着N的增大，我们预计$A(N)=\pi(N)$与$B(N)=\left(\prod_{p\le\sqrt{N}}1-\frac{1}{p}\right)N$之间的相对误差将会趋近于“$\frac{A(N)}{B(N)}-1\approx 10.94\%$”

若上述“相对误差大约是2倍”的规律能够一直持续，则公式$\left(\prod_{p\le\sqrt{N}}(1-\frac{F(p,N)}{p})\right)N$的相对误差应该会趋近于约“$22\%$”

哈代-李特尔伍德的公式和上述公式的比较

哈代的公式为

$$2\Pi_2\left(\prod_{p|N;p\ge 3}\frac{p-1}{p-2}\right)\frac{N}{(\log N)^2}$$

我们的公式为

$$\left(\prod_{p\le\sqrt{N}}1-\frac{F(p,N)}{p}\right)N$$

其中，$F(p,N)$定义为“如果$p \mid N$，则$F(p,N)=1$；否则，$F(p,N)=2$”

它可以变形为

$$\frac{N}{2}\prod_{3\le p\le\sqrt{N}}\left(1-\frac{2}{p}\right)\prod_{p|N;3\le p\le\sqrt{N}}\frac{p-1}{p-2}$$

当N很大的时候，$\prod_{3\le p\le N}\left(1-\frac{2}{p}\right)$可以表示成$\frac{4\Pi_2 e^{-2\gamma}}{(\log N)^2}$

上式变形为

$$\frac{8\Pi_2 e^{-2\gamma}N}{(\log N)^2}\left(\prod_{p|N;3\le p\le\sqrt{N}}\frac{p-1}{p-2}\right)$$

比较这两个公式，发现它们在一些地方很相似。如果我们暂时忽略“N的因子中大于$\sqrt{N}$的那一部分对结果的影响”，则这两个公式是成比例的，比例大约为“1比$4\ e^{-2\gamma}$”，“相对误差”约为$\frac{1}{4\ e^{-2\gamma}}-1\approx 20.69\%$。这与我们前面根据“相对误差大约是2倍”估算的“$22\%$”是很接近的。或许我们前面观察到的“误差大约是2倍”有一定的内在原因。因为$2\ e^{-\gamma}=1.1229\ldots$，它的值接近1，而$4\ e^{-2\gamma}$是$2\ e^{-\gamma}$的平方。