An estimate on the density of semiprimes pq having $p \leq q < p^2$ in an interval of fixed width

Jan van Delden

July 5, 2021

A semiprime is a natural number which is the product of two (possibly equal) prime numbers. It is the purpose of this article to give some insight into the number of semiprimes pq with $p \leq q < p^2$ within an interval of fixed width. Heuristics are developed for the density of semiprimes in an interval of fixed width and an adaptation to correct for $q < p^2$ is discussed.

An asymptotic formula for the number of semiprimes less than a given bound is borrowed from On Distribution of Semiprime Numbers [1], the presented formula for the distribution of strong semiprimes is adapted.

1 Introduction

In order to be able to count the semiprimes of the form pq with $p \leq q < p^2$ and $pq \leq n$ we need to distinguish between two situations. Given p we either have $pq \leq n$ for all $q < p^2$ or we reach the bound n sooner. An example might illustrate this.

For $n = 200$ we find:

p	*bound* q	*bound* pq	q
2	$p^2 = 4$	$p^3 = 8$	$\{2, 3\}$
3	$p^2 = 9$	$p^3 = 27$	$\{3, 5, 7\}$
5	$p^2 = 25$	$p^3 = 125$	$\{5, 7, 11, 13, 17, 19, 23\}$
7	$\lfloor n/p \rfloor = 28$	$n = 200$	$\{7, 11, 13, 17, 19, 23\}$
11	$\lfloor n/p \rfloor = 18$	$n = 200$	$\{11, 13, 17\}$
13	$\lfloor n/p \rfloor = 15$	$n = 200$	$\{13\}$

We may limit p by using $p^2 \leq n$. If $p^3 \leq n$ we may count all available q in $p \leq q < p^2$. If $p^3 > n$ the available q and hence the semiprimes pq are defined by $p \leq q \leq \lfloor n/p \rfloor$.

2 Definitions

The number of primes $p \in \mathbb{P}$ less or equal to a given bound x is defined as

$\pi(x) = \#\{p \in \mathbb{P} | p \leq x\}$

The number of semiprimes less than or equal to x can be expressed by

$$\pi_2(x) = \sum_{p_k \leq \sqrt{x}} (\pi(\lfloor x/p_k \rfloor) - \pi(p_k) + 1) = \sum_{p_k \leq \sqrt{x}} (\pi(\lfloor x/p_k \rfloor) - k + 1) \tag{2.1}$$

The number of semiprimes less than or equal to x for which $p \leq q < p^2$ may be split into

$$\pi_{2,a}(x) = \sum_{p_k \leq \sqrt[3]{x}} (\pi(p_k^2) - \pi(p_k) + 1) = \sum_{p_k \leq \sqrt[3]{x}} (\pi(p_k^2) - k + 1) \tag{2.2}$$

[1] On Distribution of Semiprime Numbers, by Sh.T. Ishmukhametov and F.F. Sharifullina, 01-31-2013

where we are able to count all available semiprimes of the form pq with $p \le q < p^2$, and

$$\pi_{2,b}(x) = \sum_{\sqrt[3]{x} < p_k \le \sqrt{x}} (\pi(\lfloor x/p_k \rfloor) - \pi(p_k) + 1) = \sum_{\sqrt[3]{x} < p_k \le \sqrt{x}} (\pi(\lfloor x/p_k \rfloor) - k + 1) = \pi_2(x) - \pi_2(\sqrt[3]{x^2}) \tag{2.3}$$

where the number of available q are limited by $\lfloor x/p_k \rfloor$ instead of p^2. The last equality is only for demonstration purposes.

3 Goal

We are interested in the number of semiprimes of the form pq with $p \le q < p^2$ in an interval of the form $A < pq \le B$ for natural numbers A, B.

$$\Delta\pi_2(A,B) = \underbrace{\pi_{2,a}(B) - \pi_{2,a}(A)}_{q \text{ limited by } p^2} + \pi_{2,b}(B) - \pi_{2,b}(A) \tag{3.1}$$

The first term, counting semiprimes for all q with $q < p^2$ cancels or equals 1 if we choose the width of the interval $W = B - A$ small compared to A. Set $p_{max,A} = \text{prevprime}(\lfloor \sqrt[3]{A} \rfloor)$ and $p_{max,B} = \text{prevprime}(\lfloor \sqrt[3]{B} \rfloor)$. We would like these primes to coincide but that would be too much to hope for in general; $p_{max,A}$ might be slightly smaller than $\sqrt[3]{A}$ and the next prime, $\text{nextprime}(p_{max,A})$, might still be smaller than or equal to $\sqrt[3]{B}$. All we can do is make sure that $p_{max,B} <= \text{nextprime}(p_{max,A}) <= p_{max,A} + 2$, i.e. enforce a maximal prime gap 2.

$$W \le 2 + 6\left(\sqrt[3]{A} + 1\right)^2 \vee A \ge \left(-1 + \sqrt{(W-2)/6}\right)^3 \implies 0 \le \pi_{2,a}(B) - \pi_{2,a}(A) \le 1 \tag{3.2}$$

For instance $W = 10^6 \implies A \ge 6.75 \cdot 10^7$ ensures that we may compute (3.1) by only considering the second term (with a maximum error equal to 1). In general this correction equals

$$\pi_{2,a}(B) - \pi_{2,a}(A) = \sum_{\lceil \sqrt[3]{A} \rceil < p_k \le \lfloor \sqrt[3]{B} \rfloor} (\pi(p_k^2) - k + 1) \tag{3.3}$$

In order to estimate the number of semiprimes pq in an interval of the form $[A, B]$, with A large enough, we may thus focus on $\pi_{2,b}(x)$ (2.3). This formula is hardly practical for large values of x. Instead we'll approximate two functions $g(y), g^*(y)$, where $g(y)$ is the probability of finding a semiprime less than or equal to y and $g^*(y)$ the probability of finding a semiprime pq for which $p > \sqrt[3]{y}$. For $g(y)$ I'll follow the text given by Sh.T. Ishmukhametov and F.F. Sharifullina and for $g^*(y)$ I'll derive an adaptation to their formula (which I believe has an error), but will otherwise closely follow their lead.

4 Estimates for $g(y)$ and $g^*(y)$

Let y be fixed and $p \le \lfloor \sqrt{y} \rfloor$ be an arbitrary prime. Consider the real number $v = y/p$. The number v is an integer with probability $1/p$ and prime with probability

$$g_p(y) \approx 1/(p \ln(y/p)) = 1/(p(\ln(y) - \ln(p)))$$

i.e., $g_p(y)$ equals the probability that y is semiprime and one of its divisors is p. We may thus approximate $g(y)$ by the sum

$$g(y) \approx \sum_{p \le \sqrt{y}} \frac{1}{p(\ln(y) - \ln(p))} \tag{4.1}$$

the summing is over prime p. In turn, we may approximate $g^*(y)$ by the sum

$$g^*(y) \approx \sum_{\sqrt[3]{y} < p \le \sqrt{y}} \frac{1}{p(\ln(y) - \ln(p))} \tag{4.2}$$

where we also assume that p is prime.

Notice that in doing so we used that v is prime with probability

$$\frac{\pi(v)}{v} \approx \frac{1}{\ln(v)}$$

without considering a possible estimate of the error that is involved. For large enough v we have

$$\frac{1}{\ln(v)}\left(1+\frac{C_1}{\ln(v)}\right) \underset{v\geq 599}{\leq} \frac{\pi(v)}{v} \underset{v>1}{\leq} \frac{1}{\ln(v)}\left(1+\frac{C_2}{\ln(v)}\right)$$

with $C_1 = 1, C_2 = 1.2762$. Sharper estimates may be found in *Estimates of some functions over primes without R.H.*, by *Pierre Dusart* [2]. Notice that for $\sqrt[3]{y} < p \leq \sqrt{y}$ we have

$$\frac{1.5}{\ln(y)} < \frac{1}{\ln(v)} \leq \frac{2}{\ln(y)} \tag{4.3}$$

5 Preparation, Abel's summation theorem and Mertens' formula

The given summand in (4.1) and (4.2) may be recast to

$$\frac{1}{p_n(\ln(y)-\ln(p_n))} = \frac{\ln(p_n)}{p_n} \cdot \frac{1}{\ln(p_n)(\ln(y)-\ln(p_n))} \tag{5.1}$$

Where the prime p is equipped with an indexed. In doing so we may rewrite (4.1) for fixed y into the form

$$g(y) \approx \sum_{p_n \leq \sqrt{y}} a_n \varphi_y(p_n) \tag{5.2}$$

where

$$a_n = \frac{\ln(p_n)}{p_n}$$

and

$$\varphi_y(t) = \frac{1}{\ln(t)(\ln(y)-\ln(t))} = \frac{1}{\ln(y)}\left(\frac{1}{\ln(t)} + \frac{1}{\ln(y)-\ln(t)}\right) \tag{5.3}$$

is a function of t having a continuous derivative for $0 < t < y$. This region is quite a bit larger than needed, since we would like to apply $\varphi_y(t)$ for $2 \leq t \leq \sqrt{y}$. With $p_n = \lambda_n, x = \sqrt{y}$ we may apply:

Abel's summation theorem *Let $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n \leq \cdots$ be a sequence or real numbers such that $\lim_{n\to\infty} \lambda_n = \infty$, and $\{a_n\}, n \in \mathbb{N}$ be a sequence of complex numbers. Let $A(x) = \sum_{\lambda_n < x} a_n$ and $\varphi(x)$ be a complex valued function, defined for $x \geq 0$ having a continuous derivative for $x > 0$. Then*

$$\sum_{n=1}^{k} a_n \varphi(\lambda_n) = A(x)\varphi(x) - \int_{\lambda_1}^{x} A(t)\varphi'(t)\,\mathrm{d}t \tag{5.4}$$

where $\lambda_k \leq x < \lambda_{k+1}$.

In order to be able to compute the right hand side we need to find an alternative expression for

$$A(x) = \sum_{p_n < x} a_n = \sum_{p_n < x} \frac{\ln(p_n)}{p_n}$$

We may approximate $A(x)$ by using Mertens' first theorem

$$A(x) = \ln(x) + C + O(1/\ln(x))$$

We'll replace $A(x)$ in (5.4) by

$$A^*(x) = \ln(x) + R(x), |R(x)| < 2$$

instead[3], where I'll replace $R(x)$ by a suitable constant R to simplify the computation. Here I'll deviate from Sh.T. Ishmukhametov and F.F. Sharifullina, for two reasons. The main reason is that I don't need the asymptotic behaviour of $R(x)$, a second reason is that their related investigation of $I_3(x)$ contains a serious error.

[2] Estimates of some functions over primes without R.H., by Pierre Dusart, 02-02-2010

[3] Mertens' proof on Mertens' theorem, by Mark T. Villarino, (2.3.3), 04-28-2005

6 Asymptotic analysis of $g(y), g^*(y)$

We'll investigate

$$h(x) = A^*(x)\varphi_y(x) - \int_2^x A^*(t)\varphi_y'(t)\,\mathrm{d}t \tag{6.1}$$

where, for some constant $R < 2$

$$A^*(x)\varphi_y(x) = \frac{\ln(x) + R}{\ln(x)(\ln(y) - \ln(x))} \tag{6.2}$$

and

$$\varphi_y'(t) = \frac{1}{\ln(y)}\left(-\frac{1}{t\ln^2(t)} + \frac{1}{t(\ln(y) - \ln(t))^2}\right) \tag{6.3}$$

And we have

$g(y) \approx h(\sqrt{y}),\ g^*(y) \approx h(\sqrt{y}) - h(\sqrt[3]{y})$

Substitution of $x = \sqrt{y}, \ln(x) = \ln(y)/2$ in (6.2) gives

$$A^*(\sqrt{y})\varphi_y(\sqrt{y}) = \frac{2}{\ln(y)} + \frac{4R}{\ln^2(y)} \tag{6.4, g}$$

Substitution of $x = \sqrt[3]{y}, \ln(x) = \ln(y)/3$ in (6.2) gives

$$A^*(\sqrt[3]{y})\varphi_y(\sqrt[3]{y}) = \frac{3}{2\ln(y)} + \frac{9R}{2\ln^2(y)}$$

The contribution of (6.2) for $g^*(y)$ is thus

$$A^*(\sqrt{y})\varphi_y(\sqrt{y}) - A^*(\sqrt[3]{y})\varphi_y(\sqrt[3]{y}) = \frac{1}{2\ln(y)} - \frac{R}{2\ln^2(y)} \tag{6.4, g^*}$$

The integral in (6.1) may be split as

$$\int_2^x A^*(t)\varphi_y(t)\,\mathrm{d}t = \underbrace{\frac{1}{\ln(y)}\int_2^x -\frac{1}{t\ln(t)} + \frac{\ln(t)}{t(\ln(y) - \ln(t))^2}\,\mathrm{d}t}_{I_1} + R\underbrace{\int_2^x \varphi_y'(t)\,\mathrm{d}t}_{I_2}$$

Substitution of $z = \ln(t), \mathrm{d}z = \mathrm{d}t/t$ gives

$$I_1 = \frac{1}{\ln(y)}\int_{\ln(2)}^{\ln(x)} -\frac{1}{z} + \frac{z}{(\ln(y) - z)^2}\,\mathrm{d}z = \frac{1}{\ln(y)}\int_{\ln(2)}^{\ln(x)} -\frac{1}{z} + \frac{z - \ln(y) + \ln(y)}{(\ln(y) - z)^2}\,\mathrm{d}z$$

$$= \frac{1}{\ln(y)}\Big[-\ln(z) + \ln(\ln(y) - z)\Big]_{\ln(2)}^{\ln(x)} + \Big[\frac{1}{\ln(y) - z}\Big]_{\ln(2)}^{\ln(x)}$$

$$= \frac{\ln(\ln(2)) - \ln(\ln(x)) + \ln(\ln(y) - \ln(x)) - \ln(\ln(y) - \ln(2))}{\ln(y)} + \frac{1}{\ln(y) - \ln(x)} - \frac{1}{\ln(y) - \ln(2)} \tag{6.5}$$

We may use a Taylor series approximation $\ln(1 + \alpha) = \alpha + O(\alpha^2)$ to obtain

$$\ln(\ln(y) - \ln(2)) = \ln\left(\ln(y)\left(1 - \frac{\ln(2)}{\ln(y)}\right)\right) = \ln(\ln(y)) - \frac{\ln(2)}{\ln(y)} + O\left(\frac{1}{\ln^2(y)}\right)$$

Substition of $x = \sqrt{y}, \ln(x) = \ln(y)/2, \ln(\ln(y) - \ln(x)) = \ln(\ln(x))$ into (6.5) and the above asymptotics give:

$$I_1(\sqrt{y}) = \frac{\ln(\ln(2)) - \ln(\ln(y)) + \ln(2)/\ln(y) + O(1/\ln^2(y))}{\ln(y)} + \frac{2}{\ln(y)} - \frac{1}{\ln(y) - \ln(2)} = -\frac{\ln(\ln(y))}{\ln(y)} + O\left(\frac{1}{\ln(y)}\right) \tag{6.5, g}$$

For $g^*(y)$ we need to compute $I_1(\sqrt{y}) - I_1(\sqrt[3]{y})$ and may drop the contributions of $\ln(\ln(2)), \ln(y) - \ln(2)$ in (6.5). We are left with

$$I_1(\sqrt{y}) - I_1(\sqrt[3]{y}) = \frac{\ln(\ln(y)/3) - \ln(2\ln(y)/3)}{\ln(y)} + \frac{2}{\ln(y)} - \frac{3}{2\ln(y)} = -\frac{\ln(2)}{\ln(y)} + \frac{1}{2\ln(y)} \qquad (6.5,\ g^*)$$

Similarly we find, use (5.3):

$$I_2 = \varphi_y(x) - \varphi_y(2) = \frac{1}{\ln(x)(\ln(y) - \ln(x))} - \frac{1}{\ln(2)(\ln(y) - \ln(2))} \qquad (6.6)$$

Substitute $x = \sqrt{y}$ and find its contribution to $g(y)$

$$R \cdot I_2(\sqrt{y}) = R\left(\frac{2}{\ln^2(y)} - \frac{1}{\ln(2)(\ln(y) - \ln(2))}\right) = O\left(\frac{1}{\ln(y)}\right) \qquad (6.6,\ g)$$

Similarly we find the contribution to $g^*(y)$

$$R \cdot (I_2(\sqrt{y}) - I_2(\sqrt[3]{y})) = -\frac{5R}{2\ln^2(y)} = O\left(\frac{1}{\ln^2(y)}\right) \qquad (6.6,\ g^*)$$

If we collect (6.4,g),(6.5,g) and (6.6,g) we obtain

$$g(y) = \frac{\ln(\ln(y))}{\ln(y)} + O\left(\frac{1}{\ln(y)}\right) \qquad (6.7)$$

If we collect (6.4,g^*),(6.5,g^*) and (6.6,g^*) we obtain

$$g^*(y) = \frac{\ln(2)}{\ln(y)} + O\left(\frac{1}{\ln^2(y)}\right) \qquad (6.8)$$

In the article by Sh.T. Ishmukhametov and F.F. Sharifullina, a formula $\tilde{g}^*(y)$ is derived for the probability of finding a strong semiprime pq, $p \le q$, $\sqrt[4]{y} < p \le \sqrt{y}$. To be specific, the probability at the right hand side of

$$g^*(y) \approx \sum_{\sqrt[4]{y} < p \le \sqrt{y}} \frac{1}{p(\ln(y) - \ln(p))}$$

is estimated by the following asymptotics, valid for $y < 10^{10}$

$$\tilde{g}^*(y) = \frac{\ln(\ln(y))}{\ln(y)} - \frac{2.65}{\ln(y)} + \frac{13.7}{\ln^2(y)}$$

A slight adaptation in the derivation of $g^*(y)$ in (6.8) and dropping $O(1/\ln^2(y))$ gives a simple estimate

$$\tilde{g}_1^*(y) = \frac{\ln(3)}{\ln(y)}$$

A comparison of the aforementioned three expressions, where the first three rows in the table are borrowed from the article by Sh. T. Ishmukhametov and F.F. Sharifullina

y	10^3	10^4	10^5	10^6	10^7	10^8	10^9	10^{10}
$g^*(y)$	0.125	0.107	0.086	0.073	0.066	0.059	0.052	0.0470
$\tilde{g}^*(y)$	0.183	0.115	0.085	0.070	0.061	0.055	0.050	0.0470
$\tilde{g}_1^*(y)$	0.159	0.119	0.095	0.080	0.068	0.060	0.053	0.0477

The approximations to $g^*(y)$ behave quite similar. Using $\tilde{g}_1^*(y)$ there is no need of estimating additional terms of the form $C_k/\ln^k(y)$ by applying a regression technique using the same (or similar) data as a source.

7 The density of semiprimes in an interval of fixed width

The number of semiprimes less than or equal to x for which $p \leq q < p^2$, see (2.2),(2.3), may be written as

$\pi_2^*(x) = \pi_{2,a}(x) + \pi_{2,b}(x)$

These semiprimes inside an interval $[y - W/2, y + W/2]$ may be computed, see (3.1), by

$\Delta\pi_2(y - W/2, y + W/2) = \pi_2^*(y + W/2) - \pi_2^*(y - W/2)$

If we choose W sufficiently small, see (3.2).

$y \geq W/2 + (-1 + \sqrt{(W-2)/6})^3$

we may bound $\delta = \pi_{2,a}(y + W/2) - \pi_{2,a}(y - W/2) \leq 1$ and find

$\Delta\pi_2(y - W/2, y + W/2) = \pi_{2,b}(y + W/2) - \pi_{2,b}(y - W/2) + \delta$

We also defined the probability of finding a semiprime pq having $\sqrt[3]{y} < p \leq \sqrt{y}$, see (2.3)

$$\frac{\pi_{2,b}(y)}{y} \approx g^*(y)$$

Where the left hand side is exact and the right hand side is an approximation, see (4.2).

$$g^*(y) \approx \sum_{\sqrt[3]{y} < p \leq \sqrt{y}} \frac{1}{p(\ln(y) - \ln(p))}$$

Although not exact, it is my belief that if we propagate the bounds on the error term given by Pierre Dusart and use (4.3) that the final asymptotics on the density will remain as given later on.

The number of semiprimes less than y for which $p \leq q < p^2$, for y sufficiently large is equal to

$$\Delta\pi_2(y - W/2, y + W/2) \approx (y + W/2)g^*(y + W/2) - (y - W/2)g^*(y - W/2) + \delta \tag{7.1}$$

If we divide by the width W of the interval the average density is equal to

$$\frac{\Delta\pi_2(y - W/2, y + W/2)}{W} \approx \frac{(y + W/2)g^*(y + W/2) - (y - W/2)g^*(y - W/2) + \delta}{W}$$

Discard δ, as a curiosity the right hand side tends to the derivative of $yg^*(y)$ if $W/y \to 0$. Apply a Taylor series for $W/y \to 0$. We may use (6.8)

$$g^*(y) = \frac{\ln(2)}{\ln(y)} + O\left(\frac{1}{\ln^2(y)}\right)$$

and we'll find an approximation to the density

$$\frac{\Delta\pi_2(y - W/2, y + W/2)}{W} \approx \frac{\ln(2)}{\ln(y)} + O\left(\frac{1}{\ln^2(y)}\right) + O((W/y)^2) \tag{7.2}$$

The last term is included in order to be complete. The other big-o term can be estimated by using our bound on R and a further correction by using (4.3). We could have found the main term by setting $g^*(y - W/2) = g^*(y) = g^*(y + W/2)$ in (7.1) as well. As a corollary we found

$$\frac{\pi_2^*(y + W/2) - \pi_2^*(y - W/2)}{\pi(y + W/2) - \pi(y - W/2)} \approx \ln(2)$$

This constant matches experimental results quite well; this comparison may be found at

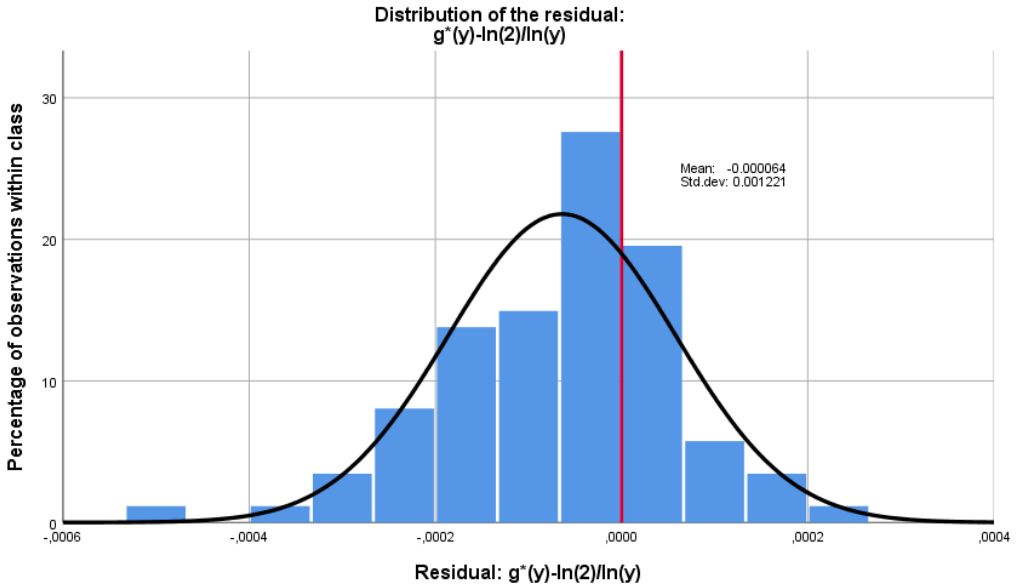
Conjecture 90, by Alain Rochelli. The prime puzzles and problem connection[4] is moderated by Carlos Rivera.

[4] The prime puzzles and problem connection

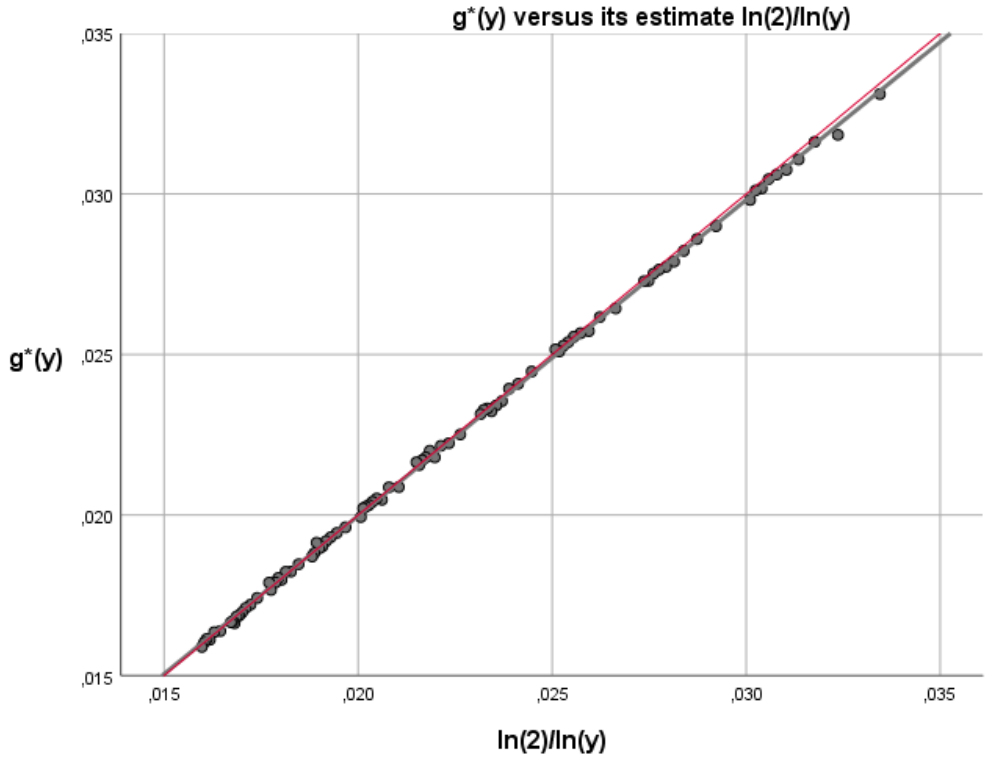
8 Visual comparison

Intervals $[y-W/2, y+W/2]$ having width $W = 2\cdot 10^6$ are used to compute the density of semiprimes of the form pq with $p \le q < p^2$ (the left hand side of (7.2)), with $y = m10^k + W/2, k \in \{9, 10, \ldots, 18\}, m \in \{1, 2, \ldots, 9\}$, where I limited $m \le 7$ for $k = 18$. A total of 87 intervals. In this section $g^*(y)$ represents the actual density.

A distribution of the residual $g^*(y) - \dfrac{\ln(2)}{\ln(y)}$

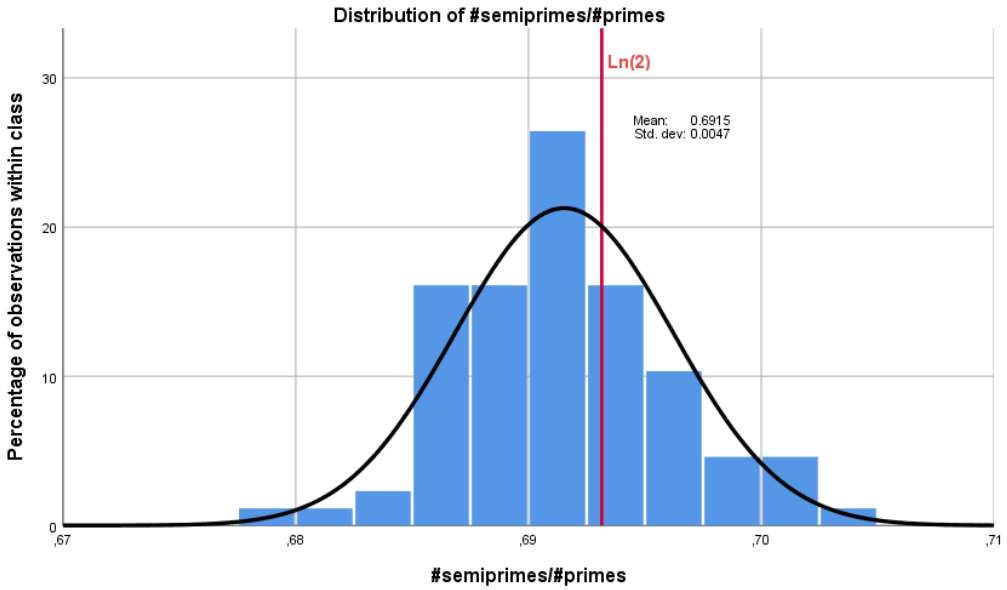


The requested density is slightly underestimated, the distribution is skewed to the left. A normal distribution was not expected, since the error is exptected to have size $O(1/\ln^2(y))$ and should decrease with increasing y. In order to capture the dependence on y a graph where $g^*(y)$ is graphed versus $\ln(2)/\ln(y)$.

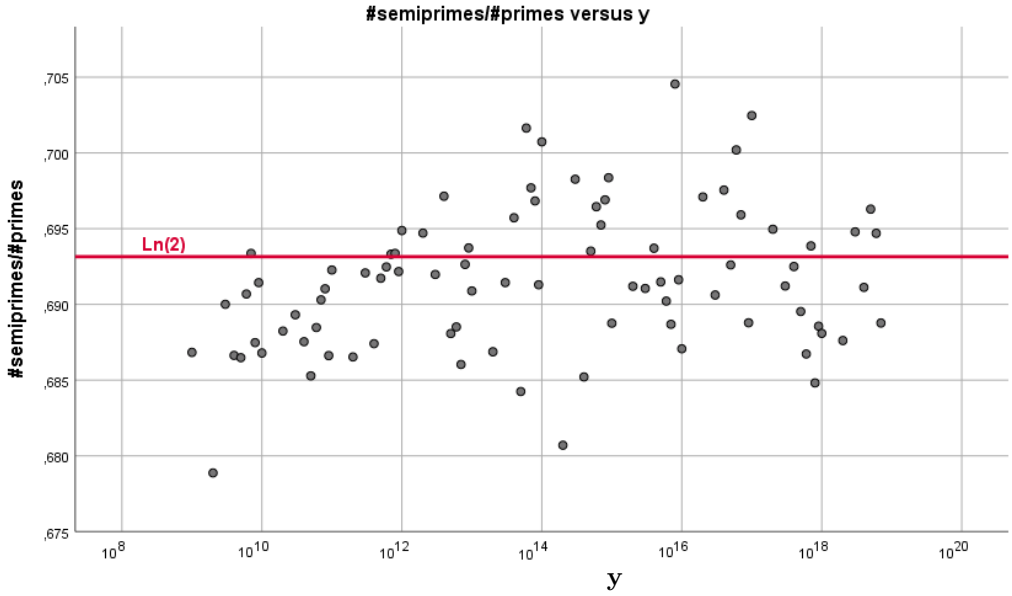


The black line represents the best fit due to linear regression and the red line has equation $g^*(y) = \ln(2)/\ln(y)$. For large values of y, i.e. for small values of $\ln(2)/\ln(y)$, the black and red line are rather close, which indicates an improvement of our estimate for larger values of y.

Similarly we may compare the fraction of the number of semiprimes and the number of primes within these intervals with ln(2).



The value ln(2) overestimates the fraction #semiprimes/#primes. Close, but no cigar. In order to investigate the dependence on y, a second graph of #semiprimes/#primes versus y is provided



The displayed domain for y is much too small to conclude that this fraction tends to ln(2) or to say anything sensible with regard to a possible decrease in the displayed deviations with increasing y. Due to Dusart we are able to bound the denominator, the #primes, however similar bounds are necessary for the numerator. The asymptotic behaviour $O(1/\ln^2(y))$ is insufficient to say anything more definitive.