# STATISTICAL BIAS IN THE DISTRIBUTION OF PRIME PAIRS AND ISOLATED PRIMES

WALDEMAR PUSZKARZ

Abstract. Computer experiments reveal that twin primes tend to center on squareful multiples of 6 more often than on squarefree multiples of 6 compared to what should be expected from the ratio of the number of squareful multiples of 6 to the number of squarefree multiples of 6 equal $\pi^2/3-1$, or ca 2.290. For multiples of 6 surrounded by twin primes, this ratio is 2.427 (for the first $10^{10}$ primes), meaning that on average for every 1000 twin primes centered on squarefree multiples of 6, there are ca 137 twins that favor squareful multiples over squarefree multiples, a bias of ca 6.0%. The same kind of bias, though a bit weaker, ca 1.9%, exists for isolated primes.

## 1. Surprising effect

Except for the first two (2 and 3), all primes are of the form $6k-1$ or $6k+1$. This implies that twin primes (consecutive primes separated by 2) surround multiples of 6 (3 and 5 are the only exception).

All natural numbers are either squarefree or nonsquarefree (squareful). Unlike the former, the latter are divisible by a square greater than 1. All primes are squarefree.

Squarefree numbers have natural density of $6/\pi^2$, which gives $1-6/\pi^2$ for natural density of squareful numbers. Thus, the ratio of relative frequencies at which one expects to find these numbers in a sufficiently large sample of natural numbers is $\pi^2/6-1=0.6449...$, which amounts to saying that on average for every 100 squarefree numbers in such a sample there are about 64 squareful ones.

The ratio in question is different for multiples of 6. Since natural density of squarefree numbers divisible by 6 is $3/\pi^2$, this ratio is $R_0=\pi^2/3-1=2.2898...$ or **2.290 for practical purposes** (most numbers in this paper are rounded off to the 3rd decimal digit). What this means is that in a sufficiently large sample, on average for every 100 squarefree multiples of 6 there are about 229 squareful numbers divisible by 6.

If prime pairs are as likely to center on squarefree multiples of 6 as they are on nonsquarefree multiples of 6 (i.e., their distribution is unbiased in this respect), we should expect the same ratio for the multiples of 6 in their centers if calculations are performed on a large enough sample of prime pairs.

**However, this is not the case.**

This can be observed already in a sample of the first $10^6$ primes and the discrepancy persists (may even be getting slightly stronger) for larger samples. The largest we used consisted of the first $10^{10}$ primes and for it the ratio is $R_2=$ **2.427**.

2010 *Mathematics Subject Classification.* 11N05.

*Key words and phrases.* Primes, twin primes.

Defining the relative difference as the absolute value of the ratio of the difference between the experimental value and the theoretical one to the theoretical one, we find out that the relative difference in this case is ca 6.0%.

The same calculations applied to isolated primes (primes $p$ such that neither $p-2$ nor $p+2$ is prime) reveal a similar bias: such primes occur next to squareful multiples of 6 slightly more often than next to squarefree multiples of 6 compared to what would be expected in the non-biased distribution. In this case, the ratio is $R_1 = \mathbf{2.333}$ (for the sample of $10^{10}$ primes), and the relative difference is ca 1.9%.

The bias effect is smaller than for prime pairs, but too large to dismiss it as due to statistical noise.

To be sure this effect has indeed to do with primes and not squarefree numbers in general, we checked if this effect occurs for twin squarefree numbers centered on multiples of 6.

If twin primes are included in such test pairs, our ratio becomes 2.306 for the sample of the first $10^8$ multiples of 6 (and virtually unchanged for the sample of $10^9$), noticeably smaller than 2.427 and very close to the unbiased ratio, 2.290. But if we exclude twin primes from the test pairs, the effect goes away almost completely for the sample of $10^8$ as now we get 2.286, which leads to less than 0.2% in relative difference, a difference small enough to ascribe it to statistical fluctuations. Moreover, for the sample of $10^9$ primes, this ratio is 2.2889, less than 0.05% in relative difference, a very small difference indeed.

For isolated test squarefree numbers next to a multiple of 6 (to the left or right of it), the bias effect pretty much fails to manifest itself already in the sample of the first $10^8$ such multiples as we get 2.2907 (rounded off to the 4th decimal digit), which versus 2.2898 is less than 0.04% in relative difference. If the primes are excluded from these squarefree numbers, we obtain 2.2894, and the relative difference is now ca 0.02%, small enough to conclude that also for isolated primes, the bias effect is due to primes; it does not occur for other squarefree numbers. Moreover, for the sample of $10^9$ (primes excluded), the ratio is 2.2897, even closer to the unbiased theoretical value.

**Hence, to reiterate, the observed effect in both situations is most certainly a property of primes rather than a generic property of squarefree numbers.** To put it more precisely (if not pedantically), if there is any actual contribution to it from non-prime squarefree numbers, it is negligible compared to the contribution from primes.

Moreover, the effect appears pretty stable over several sample ranges with the range size growing by 10 for each data point we collected to determine the effect behavior (see the next section).

## 2. Excess functions

Let us define them as $\epsilon_1 = round(1000 * (R_1 - R_0))$ for isolated primes and $\epsilon_2 = round(1000 * (R_2 - R_0))$ for prime pairs, where $round(x)$ is tasked with rounding off $x$ to the nearest integer.

The data for $R_1$ and $R_2$ is obtained numerically while $R_0$ can be calculated analytically as done above. We obtained 5 data points for each of these functions (see the data section for more information) and they suggest (albeit quite weakly) that we may be dealing with slowly growing functions, at least in the case of $\epsilon_1$. More data is needed to be positive that the trends we suspect these functions may

be showing are not due to statistical fluctuations. It may very well be that the best approximation to these functions is a constant. This appears to be the most likely scenario for $\epsilon_2$ and it may be so asymptotically for $\epsilon_1$.

Below are the values of these functions at arguments that are consecutive powers of 10, starting at $10^6$ (we use the exponent values of $10^n$ to index the arguments).

For isolated primes,

$\epsilon_1(6) = 34$, $\epsilon_1(7) = 39$, $\epsilon_1(8) = 41$, $\epsilon_1(9) = 42$, $\epsilon_1(10) = 43$.

For prime pairs,

$\epsilon_2(6) = 136$, $\epsilon_2(7) = 134$, $\epsilon_2(8) = 135$, $\epsilon_2(9) = 136$, $\epsilon_2(10) = 137$.

These numbers represent the (average) excess of squareful numbers compared to the non-biased case for every 1000 squarefree numbers. For instance, $\epsilon_2(10)$ tells us that there are on average 137 more prime pairs centered on squareful multiples of 6 per 1000 squarefree multiples of 6 surrounded by primes than one would expect in the unbiased situation for the first $10^{10}$ primes.

## 3. Code and data

The effect discussed was first observed in Mathematica computer experiments performed on the first $10^6$ primes. The data for larger samples was obtained using PARI/GP, an open source software package for number theory.

What follows below is a sample of PARI/GP code used to obtain the data and the data. The data is indexed by exponents of range size, chosen to be powers of 10, from $10^6$ to $10^{10}$. The code for Mathematica can easily be produced from the PARI/GP code.

Our code counts all prime pairs (even though the first of them, $\{3, 5\}$, is not centered on a multiple of 6) and excludes 2 as an isolated prime. While 2 is sometimes treated as an isolated prime, it is actually less isolated from other primes than all odd primes save for 3. With 2 excluded, the number of isolated primes plus twice the number of pairs still add to a range size ($10^6$ through $10^{10}$), for 5 is counted twice as a member of two consecutive pairs that share it. These choices have no impact on our statistical results. We mention them for the sake of clarity.

In what follows, $a$ represents the number of all target objects (primes or test squarefree numbers), while $b$ only the number of such objects next to or centered on squarefree numbers. The ratios discussed above are calculated as $R = (a - b)/b$.

### Part A. Prime numbers

#### Twin primes

```
a=0; forprime(n=2, prime(10^8), isprime(n+2)&&a++); print1(a)
\\all twins
b=0; forprime(n=2, prime(10^8), isprime(n+2)&&issquarefree(n+1)
&&b++); print1(b) \\twins centered on a squarefree number
```

$a$: 86027 (6), 738597 (7), 6497407 (8), 58047180 (9), 524733511 (10).
$b$: 25113 (6), 215732 (7), 1897137 (8), 16944418 (9), 153121114 (10).

#### Isolated primes

```
a=0; forprime(n=3, prime(10^8), !isprime(n+2)&&!isprime(n-2)&&a++);
```

```
print1(a) \\all
b=0; forprime(n=3, prime(10^8), !isprime(n+2)&&!isprime(n-2)&&
((n%6==1&&issquarefree(n-1))||(n%6==5&&issquarefree(n+1)))&&b++);
print1(b) \\next to a squarefree number
```

$a$: 827946 (6), 8522806 (7), 87005186 (8), 883905640 (9), 8950532978 (10).
$b$: 249071 (6), 2560208 (7), 26123609 (8), 265275545 (9), 2685404943 (10).

**Part B. Test squarefree numbers**

**Squarefree twins (primes included) centered on a multiple of 6**

```
a=0; for(n=1, 10^8, issquarefree(6*n-1)&&issquarefree(6*n+1)&&
a++); print1(a) \\all
b=0; for(n=1, 10^8, issquarefree(6*n)&&issquarefree(6*n-1)&&
issquarefree(6*n+1)&&b++); print1(b) \\centered
on a squarefree number
```

$a$: 82962973 (8), 829630636 (9).
$b$: 25097397 (8), 250974031 (9).

**Squarefree twins (primes excluded) centered on a multiple of 6**

```
a=0; for(n=1, 10^8, issquarefree(6*n-1)&&!isprime(6*n-1)&&
issquarefree(6*n+1)&&!isprime(6*n+1)&&a++);
print1(a) \\all
b=0; for(n=1, 10^8, issquarefree(6*n)&&issquarefree(6*n-1)&&
!isprime(6*n-1)&&issquarefree(6*n+1)&&!isprime(6*n+1)&&b++);
print1(b) \\centered on squarefree numbers
```

$a$: 57015536 (8), 595982891 (9).
$b$: 17348734 (8), 181210143 (9).

**Isolated squarefree numbers (primes included) next to a multiple of 6**

```
a=0; for(n=1, 10^8, (issquarefree(6*n-1)||issquarefree(6*n+1))&&
a++); print1(a) \\number of cases a squarefree number is next to
a multiple of 6
b=0; for(n=1, 10^8, (issquarefree(6*n))&&(issquarefree(6*n-1)||
issquarefree(6*n+1))&&b++); print1(b) \\number of cases
a squarefree number is next to a squarefree multiple of 6
```

$a$: 99415124 (8).
$b$: 30211331 (8).

**Isolated squarefree numbers (primes excluded) next to a multiple of 6**

```
a=0; for(n=1, 10^8, (issquarefree(6*n-1)&&!isprime(6*n-1))||
```

```
(issquarefree(6*n+1)&&!isprime(6*n+1))&&a++); print1(a) \\number of
cases a non-prime squarefree number is next to a multiple of 6
b=0; for(n=1, 10^8, issquarefree(6*n)&&((issquarefree(6*n-1)&&
!isprime(6*n-1))||(issquarefree(6*n+1)&&!isprime(6*n+1)))&&b++);
print1(b) \\number of cases a non-prime squarefree number is
next to a squarefree multiple of 6
```

$a$: 94037859 (8), 948253019 (9).
$b$: 28588317 (8), 288245142 (9).

## 4. Conclusion

The results we reported above are quite basic, were obtained in an elementary fashion, and concern fundamental classes of numbers. It is therefore rather surprising that we found no mention of them in the literature of the subject. This leads us to believe that the statistical bias they describe was most likely unknown.

The study of biases in the distribution of prime numbers has recently been reinvigorated by the work of Lemke Oliver and Soundararajan on the phenomenon related to the one observed by Chebyshev already in 1853 and known as the Chebyshev bias (see [1] and references therein). However, the effect under consideration here is of different nature than those discussed in the paper mentioned.

Moreover, and more importantly, the Chebyshev bias is significantly smaller than our bias. Using the data for the first million primes from [1], we see that the deviation from the non-biased distribution (to be half a million of primes for either of two classes of primes that the Chebyshev effect concerns) is only 170, which compared to 500,000 gives a relative difference of less than 0.04%.

Let us contrast this with our bias for prime pairs. We have 86206 such pairs centered on multiples of 6 among the first $10^6$ primes. If these pairs were distributed in the unbiased way, the way non-prime squarefree twins are, the number of them surrounding squarefree multiples of 6 would be ca $round(86206/(R_0+1)) = 26149$ and not 25113 that we actually get. The deficit we observe, 1036, represents 3.96% of 26149, being two orders of magnitude larger than 0.04% computed above.

In statistics, the sample size may matter and comparing findings from samples of considerably different sizes may lead to erroneous conclusions. However, this is not an issue here. Let us note that the size of the sample of twin primes for the first $10^{10}$ primes is 524733511, which makes it comparable to the size of the sample of non-prime squarefree twins for the first $10^9$ multiples of 6, 595982891. Yet, for the latter sample, the ratio $R$ deviates from the expected $R_0$ by less than 0.05%, while for the former this deviation is ca 6.0%, a two orders of magnitude discrepancy.

It seems that no matter how we look at the data, it is telling us that the bias discussed here is real (and considerably large): primes do occur next to squareful multiples of 6 more often than next to squarefree multiples of 6 than one would expect from the unbiased distribution, and, in particular, more often than is the case for non-prime squarefree numbers.

## 5. Links

A text file with the PARI/GP code and an Excel spreadsheet with the data and results can be downloaded from the author's site.

**Acknowledgements.** The author is grateful to the developers of PARI/GP and Wolfram Mathematica whose software was indispensable to this research.

## References

[1] R. J. Lemke Oliver and K. Soundararajan, *Unexpected biases in the distribution of consecutive primes*, arXiv:1603.03720 [math.NT] (2016).

Los Angeles, CA, USA
*E-mail address*: psi_bar@yahoo.com