Statistical bias in the distribution of prime pairs and isolated primes

Waldemar Puszkarz

Abstract. Computer experiments reveal that primes tend to occur next to squareful numbers more often than next to squarefree numbers compared to what one should expect from a non-biased distribution. The effect is more pronounced for prime pairs than 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 (asymptotic) 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...

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).

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 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**.

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 a non-biased distribution. In this case, the ratio is **R_{1}=2.333** (for the sample of 10^10 primes), and the relative difference is ca 2.0%.

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 2.2898, equal to the non-biased theoretical value up to the first 4 decimal digits, an excellent agreement 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 less than 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.

The results reported 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. That is not to say that this bias cannot be explained theoretically, but that considering how basic it is, it should have been part of our knowledge of the distribution of primes for some time now.

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})].

The data for R_{1} and R_{2} is obtained numerically while R_{0} is calculated analytically as done above. We obtained 5 data points for each of these functions (see the data section for more information) and these points suggest (albeit quite weakly) that we may be dealing with slowly growing functions. More data points are needed to be positive that the trends we suspect these functions show are not due to statistical fluctuations. It may be that these functions are actually constant.

Here 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 label the arguments).

epsilon_{1}(6)=34 , epsilon_{1}(7)=39, epsilon_{1}(8)=41, epsilon_{1}(9)=42, epsilon_{1}(10)=43

epsilon_{2}(6)=136 , epsilon_{2}(7)=134, epsilon_{2}(8)=135, epsilon_{2}(9)=136, epsilon_{2}(10)=137

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 the PARI/GP code used to obtain the data and the data. 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. 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

Twins

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

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

Isolated primes

a=0; forprime(n=prime(2), prime(10^8), !isprime(n+2)&&!isprime(n-2)&&a++); print1(a) \\all
b=0; forprime(n=prime(2), 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 (10^6), 8522806 (10^7), 87005186 (10^8), 883905640 (10^9), 8950532978 (10^10)
b: 249071 (10^6), 2560208 (10^7), 26123609 (10^8), 265275545 (10^9), 2685404943 (10^10)

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 (10^8), 829630636 (10^9)
b: 25097397 (10^8), 250974031 (10^9)

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 (10^8), 595982891 (10^9)
b: 17348734 (10^8), 181210143 (10^9)

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

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

a: 99415124 (10^8)
b: 30211331 (10^8)

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 squarefree number that is not prime 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 squarefree number that is not prime is next to a squarefree multiple of 6
```

a: 94037859 (10^8) , 948253019 (10^9)
b: 28588317 (10^8) , 288245142 (10^9)

Acknowledgments. The author is grateful to the developers of Mathematica and PARI/GP. Without their software, this research would not have been possible.

Version 0.1 (2/11/2018)
Los Angeles, CA
Email address: psi_bar@yahoo.com