Hitting Primes via Dice Rolls

Zhang Ruiyang

Abstract

This notes is prepared for a talk at the UCL Undergraduate Math Colloquium in October 2022 under the same title by the author. This notes is mostly based on the paper by Noga Alon and Yaakov Malinovsky [1].

A standard, fair dice has six faces, each with a number from 1 to 6, and the chance of getting any one of the six faces is going to be 1/6. Imagine that we roll the dice once and get a number. This number could be prime (say when it gives 2), and then we have 'hit' a prime in one roll. If the number is not a prime (say we get a 4 instead), we will roll the dice again. If the second roll gives us a number that makes the sum of the numbers from the dice roll a prime (say we get a 4 and then a 3, which has a sum of 7 and this is a prime), we have hit a prime in two rolls. We will repeat rolling dice until the sum of all the numbers we have obtained is a prime, then we stop and record the number of the rolls, and claim that we have hit a prime in that number of rolls.

The main question of this study is thus: **how many rolls, on average, do we need to hit a prime?**

This quantity is obviously random, so we will consider its expectation and variance to get a sense of its performance. It will be shown later that the expectation of this quantity is around 2.43, and the variance is around 6.24.

This notes is divided into three parts. The first part will give a more quantifiable description of the question, and outline the proof. The second part will study the finite portion of the overall quantity, while the third part will aim to obtain a bound on the infinite tail portion.

1 Problem Setup

Let $X_1, X_2, \ldots$ be independent and identically distributed (i.i.d.) random variables that take 1 to 6 with equal probability, each representing a roll of dice. We also consider the partial sums S_n that is defined to be

$$S_n := \sum_{i=1}^{n} X_i.$$

This is the sum of the first n dice rolls, and we would like to know when will be the first time this sum is a prime. Let $\mathcal{P}$ denote the set of all primes. The quantity of our interest here is defined as follows:

$$\tau := \min\{n \geq 1 \mid S_n \in \mathcal{P}\}.$$

The two associated quantities that we would like to estimate are $\mathbb{E}[\tau]$ and $\mathrm{Var}[\tau]$.

We will consider the quantity $\mathbb{E}[\tau]$ first. This is not too easy to compute and estimate directly, but we can make the following rewriting to make it more penetrable.

$$\begin{aligned}
\mathbb{E}[\tau] &= \sum_{x=1}^{\infty} x\mathbb{P}(\tau = x) \\
&= \mathbb{P}(\tau = 1) + 2\mathbb{P}(\tau = 2) + 3\mathbb{P}(\tau = 3) + \dots \\
&= \left[\sum_{x=1}^{\infty} \mathbb{P}(\tau = x)\right] + \left[\sum_{x=2}^{\infty} \mathbb{P}(\tau = x)\right] + \left[\sum_{x=3}^{\infty} \mathbb{P}(\tau = x)\right] + \dots \\
&= \mathbb{P}(\tau \geq 1) + \mathbb{P}(\tau \geq 2) + \mathbb{P}(\tau \geq 3) + \dots \\
&= \sum_{x=1}^{\infty} \mathbb{P}(\tau \geq x).
\end{aligned}$$

This is a much nicer expression, and we would like to obtain a similar result for $\mathrm{Var}[\tau]$. Since $\mathrm{Var}[\tau] = \mathbb{E}[\tau^2] - \mathbb{E}[\tau]^2$, we just need to obtain $\mathbb{E}[\tau^2]$. Firstly, we have the following identity

$$n^2 = \sum_{k=1}^{n} (2k-1) \qquad \text{for all} n \in \mathbb{Z}_+.$$

This can be easily checked, and it gives us the following.

$$\begin{aligned}
\mathbb{E}[\tau^2] &= \sum_{x=1}^{\infty} x^2\mathbb{P}(\tau = x) \\
&= \mathbb{P}(\tau = 1) + 2^2\mathbb{P}(\tau = 2) + 3^2\mathbb{P}(\tau = 3) + \dots \\
&= \mathbb{P}(\tau = 1) + \sum_{k=1}^{2} (2k-1)\mathbb{P}(\tau = 2) + \sum_{k=1}^{3} (2k-1)\mathbb{P}(\tau = 3) + \dots \\
&= \sum_{x=1}^{\infty} (1 \times 2 - 1) \cdot \mathbb{P}(\tau = x) + \sum_{x=2}^{\infty} (2 \times 2 - 1) \cdot \mathbb{P}(\tau = x) + \sum_{x=3}^{\infty} (3 \times 2 - 1) \cdot \mathbb{P}(\tau = x) + \dots \\
&= (1 \times 2 - 1) \cdot \mathbb{P}(\tau \geq 1) + (2 \times 2 - 1) \cdot \mathbb{P}(\tau \geq 2) + (3 \times 2 - 1) \cdot \mathbb{P}(\tau \geq 3) + \dots \\
&= \sum_{x=1}^{\infty} (2x-1)\mathbb{P}(\tau \geq x).
\end{aligned}$$

Thus, we have the following expressions

$$\mathbb{E}[\tau] = \sum_{k=1}^{\infty} p(k) \qquad \mathbb{E}[\tau^2] = \sum_{k=1}^{\infty} (2k-1)p(k)$$

where $p(k) := \mathbb{P}(\tau \geq k)$.

These two expressions involve an infinite sum. One way to approximate it is by finding the exact value of the first many terms while finding an upper bound for the remaining tail terms. By the way this random variable τ is defined, it should not be surprising for $p(k) \to 0$ as $k \to \infty$. This forms the strategy of our proof.

We will find a way to compute the sum of the first 1000 terms in the next section, and will try to bound the remaining terms in the last section. The choice of 1000 is slightly arbitrary, since

any sufficiently large value will make the following work valid, and a larger value will only result in a slightly more precise estimation.

Before moving on to the next section, we will first show the results of some Monte Carlo simulations of this problem, to give us a rough sense of what the quantity of interest should be. The following tables are lifted from the original paper [1].

number of repetitions	$mean(\tau)$	$variance(\tau)$	$\max(\tau)$
10^6	2.4316	6.2735	49
$2 * 10^6$	2.4274	6.2572	67
$3 * 10^6$	2.4305	6.2372	70
$5 * 10^6$	2.4287	6.2418	64
10^7	2.4286	6.2463	65

Figure 1: Monte Carlo Simulations

2 Estimation of the Main Part

The quantity that we would like to compute in this section is $\sum_{k=1}^{1000} p(k)$, and also $\sum_{k=1}^{1000}(2k-1)p(k)$. This might be a hard task at first, but if we can find a way to compute $p(k)$ iteratively, then we could write out some computer programme to help us calculate these quantities.

It turns out that there is such an auxiliary function that can be used to define $p(k)$ iteratively. For $k \geq 1$ and $n \neq \mathcal{P}$, with $k \leq n \leq 6k$, we let $p(k,n)$ denote

$$p(k,n) := \mathbb{P}[X_1 + \ldots, X_k = n \text{ and } X_1 + \ldots, X_i \neq \mathcal{P} \ \forall \ i < k].$$

In words, this quantity $p(k,n)$ denotes the probability that the sum of the first k rolls has sum n and it has not hit a prime yet. We will compute these quantities for all $1 \leq k \leq K$ where K is a fixed number, and $1 \leq n \leq 6K$.

It is easy to note that $p(1,1) = p(1,4) = p(1,6) = 1/6$ while $p(1,n) = 0$ for any other possible n. Next, for $k = 2, \ldots, K$, and any $n \notin \mathcal{P}$ that is between k and $6k$, we have

$$p(k,n) = \frac{1}{6} \sum_{\substack{i \in \{1,\ldots,6\} \\ n-i \notin \mathcal{P}}} p(k-1, n-i)$$

while $p(k,n)$ for other n is set to be 0. Notice that for a fixed k, we have $p(k) = \sum_n p(k,n)$.

A computer programme can thus be written to compute all the $p(k,n)$ for $k = 1, 2, \ldots, K$ and the corresponding n. Here, we will compute until $k = 1000$, since we are interested in computing the sum of the first 1000 terms of $p(k)$. Thus, we have

$$\sum_{k=1}^{1000} p(k) \approx 2.42850, \qquad \sum_{k=1}^{1000} (2k-1)p(k) \approx 12.14038.$$

These numbers are completely deterministic.

3 Estimation of the Tail Part

We have established the sum of the first 1000 terms for the infinite sum of both $\mathbb{E}[\tau]$ and $\mathrm{Var}[\tau]$ in the previous section. Here, we will try to bound the remaining terms.

We write $\mathbb{E}[\tau] = E_K + E_R$ where $E_K = \sum_{k=1}^{K} p(k)$ and $E_R = \sum_{k>K} p(k)$. Similarly, we have $\mathbb{E}[\tau^2] = E_K^{(2)} + E_R^{(2)}$ where $E_K^{(2)} = \sum_{k=1}^{K}(2k-1)p(k)$ and $E_R^{(2)} = \sum_{k>K}(2k-1)p(k)$. We know E_{1000} and $E_{1000}^{(2)}$. The goal of this section is to give an upper bound for E_R and $E_R^{(2)}$, which will then complete this proof.

Proposition 1. *For any k and $n \notin \mathcal{P}$, we have*

$$p(k,n) < \frac{1}{3}\left(\frac{5}{6}\right)^{\pi(n)}$$

where $\pi(n) := \#\{x \le n, x \in \mathcal{P}\}$ is the prime counting function.

Proof. We will prove this by induction on k, and we will ignore those n that makes $p(k,n) = 0$. For $k = 1$, we have

$$\begin{aligned} p(1,1) &= \frac{1}{6} < \frac{1}{3}\left(\frac{5}{6}\right)^{0} = \frac{1}{3} \\ p(1,4) &= \frac{1}{6} < \frac{1}{3}\left(\frac{5}{6}\right)^{2} = \frac{25}{108} \\ p(1,6) &= \frac{1}{6} < \frac{1}{3}\left(\frac{5}{6}\right)^{3} = \frac{125}{648}. \end{aligned}$$

Assume that the desired statement holds for all $k \le m-1$. Then, we consider $k = m$.

Suppose there are q primes between $\{n-6, \dots, n-1\}$, then we have $\pi(n-i) \ge \pi(n) - q$ for all $n - i \notin \mathcal{P}$. So, we have

$$\begin{aligned} p(m,n) &= \frac{1}{6} \sum_{\substack{i \in \{1,\dots,6\} \\ n-i \notin \mathcal{P}}} p(m-1, n-i) \\ &< \frac{1}{6}\left(\frac{1}{3}\left(\frac{5}{6}\right)^{\pi(n-i)}\right) \cdot (6-q) \quad \text{by induction hypothesis} \\ &\le \frac{1}{6}\left(\frac{1}{3}\left(\frac{5}{6}\right)^{\pi(n)-q}\right) \cdot (6-q) \\ &= \left(1 - \frac{q}{6}\right)\frac{1}{3}\left(\frac{5}{6}\right)^{\pi(n)-q} \\ &\le \left(1 - \frac{1}{6}\right)^{q}\frac{1}{3}\left(\frac{5}{6}\right)^{\pi(n)-q} \quad \text{by Bernoulli inequality} \\ &= \frac{1}{3}\left(\frac{5}{6}\right)^{\pi(n)} \end{aligned}$$

as desired, thus completing the induction. □

Next, we would like to give a lower bound to $\pi(n)$ to improve on the upper bound from the above result. Notice that using the prime number theory, for $n > 1000$, we have $\pi(n) > 0.9\frac{n}{\log n}$. So we have the following result.

Corollary 1. *For any k and $n \notin \mathcal{P}$, we have*

$$p(k,n) < \frac{1}{3}\left(\frac{5}{6}\right)^{0.9n/(\log n)}.$$

This allows us to bound R_{1000}. We have

$$\begin{aligned}
R_{1000} &= \sum_{k>1000} p(k) = \sum_{k=1001}^{\infty} \sum_{n\in\{k,k+1,\ldots,6k\}} p(k,n) \\
&< \sum_{k=1001}^{\infty} \sum_{n=k}^{6k} \frac{1}{3}\left(\frac{5}{6}\right)^{0.9n/(\log n)} \\
&= \sum_{n=1001}^{\infty} \sum_{k=\max(1001,n/6)}^{n} \frac{1}{3}\left(\frac{5}{6}\right)^{0.9n/(\log n)} \qquad \text{via double sum} \\
&< \sum_{n=1001}^{\infty} (n-1000)\frac{1}{3}\left(\frac{5}{6}\right)^{0.9n/(\log n)}.
\end{aligned}$$

Let $f(n) = (n-1000)\frac{1}{3}\left(\frac{5}{6}\right)^{0.9n/(\log n)}$. For $n \geq 1000$, $f(n)$ is uniquely maximised at 1050, and for all $n \geq 1050$, we observe that

$$\frac{f(n+13\log n)}{f(n)} < \frac{1}{2}.$$

This means that we can break the terms after 1050 into intervals of length $13\log(1050)$ and the function will be roughly halved after every interval. This means, we have

$$\begin{aligned}
R_{1000} &< \sum_{n=1001}^{\infty} (n-1000)\frac{1}{3}\left(\frac{5}{6}\right)^{0.9n/(\log n)} \\
&= \sum_{n=1001}^{\infty} f(n) \\
&< 50f(1050) + f(1050)(13\log(1050))\sum_{j=0}^{\infty}\frac{1}{2^j} \\
&< 50f(1050) + 2f(1050)(13\log(1050)) \\
&< 7\times 10^{-8}.
\end{aligned}$$

This implies, the accuracy of E_{1000} as an approximation of $\mathbb{E}[\tau]$ is accurate up to 10^{-8}, which is great, and thus completes the exposition of this notes. The remainder term bound of the variance can be derived similarly.

References

[1] Alon, N., Malinovsky, Y. (2022). Hitting a prime in 2.43 dice rolls (On average). *arXiv: 2209.07698*.