Chapter 6

The Birthday Paradox, Occupancy and the Coupon Collector Problem

By Sariel Har-Peled, March 19, 2024[①]

I built on the sand
And it tumbled down,
I built on a rock
And it tumbled down.
Now when I build, I shall begin
With the smoke from the chimney

Leopold Staff, Foundations

6.1. Some needed math

Lemma 6.1.1. *For any positive integer n, we have:*

(i) $1 + x \le e^x$ *and* $1 - x \le e^{-x}$, *for all* x.

(ii) $(1 + 1/n)^n \le e \le (1 + 1/n)^{n+1}$.

(iii) $(1 - 1/n)^n \le \frac{1}{e} \le (1 - 1/n)^{n-1}$.

(iv) $(n/e)^n \le n! \le (n + 1)^{n+1}/e^n$.

(v) For any $k \le n$, *we have:* $\left(\frac{n}{k}\right)^k \le \binom{n}{k} \le \left(\frac{ne}{k}\right)^k$.

Proof: **(i)** Let $h(x) = e^x - 1 - x$. Observe that $h'(x) = e^x - 1$, and $h''(x) = e^x > 0$, for all x. That is $h(x)$ is a convex function. It achieves its minimum at $h'(x) = 0 \implies e^x = 1$, which is true for $x = 0$. For $x = 0$, we have that $h(0) = e^0 - 1 - 0 = 0$. That is, $h(x) \ge 0$ for all x, which implies that $e^x \ge 1 + x$, see Figure 6.1.

(ii, iii) Indeed, $1 + 1/n \le \exp(1/n)$ and $(1 - 1/n)^n \le \exp(-1/n)$, by (i). As such

$$(1 + 1/n)^n \le \exp(n(1/n)) = e \qquad \text{and} \qquad (1 - 1/n)^n \le \exp(n(-1/n)) = \frac{1}{e},$$

which implies the left sides of (ii) and (iii). These are equivalent to

$$\frac{1}{e} \le \left(\frac{n}{n+1}\right)^n = \left(1 - \frac{1}{n+1}\right)^n \qquad \text{and} \qquad e \le \left(1 + \frac{1}{n-1}\right)^n,$$

which are the right side of (iii) [by replacing $n + 1$ by n], and the right side of (ii) [by replacing n by $n + 1$].

(iv) Indeed,

$$\frac{n^n}{n!} \le \sum_{i=0}^{\infty} \frac{n^i}{i!} = e^n,$$

by the Taylor expansion of $e^x = \sum_{i=0}^{\infty} \frac{x^i}{i!}$. This implies that $(n/e)^n \le n!$, as required.

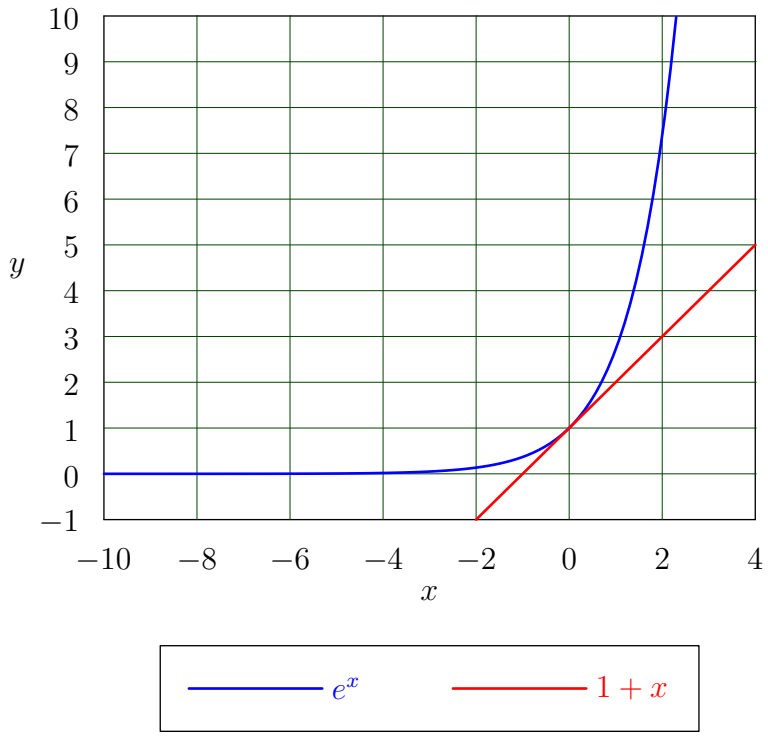

Figure 6.1

As for the righthand side. The claim holds for $n = 0$ and $n = 1$. Let $f(n) = (n+1)^{n+1}/e^n$, and observe that by (ii), we have

$$\frac{f(n)}{f(n-1)} = \frac{(n+1)^{n+1}/e^n}{n^n/e^{n-1}} = \frac{n}{e}\left(\frac{n+1}{n}\right)^{n+1} = \frac{n}{e}\left(1+\frac{1}{n}\right)^{n+1} \geq n\frac{e}{e} = n.$$

Thus, we have

$$\frac{(n+1)^{n+1}}{e^n} = f(n) = \frac{f(n)}{f(n-1)} \cdot \frac{f(n-1)}{f(n-2)} \cdots \frac{f(1)}{f(0)} \geq n \cdot n - 1 \cdots 1 = n!$$

(v) For any $k \leq n$, we have $\frac{n}{k} \leq \frac{n-1}{k-1}$ since $kn - n = n(k-1) \leq k(n-1) = kn - k$. As such, $\frac{n}{k} \leq \frac{n-i}{k-i}$, for $1 \leq i \leq k-1$. As such,

$$\left(\frac{n}{k}\right)^k \leq \frac{n}{k} \cdot \frac{n-1}{k-1} \cdots \frac{n-i}{k-i} \cdots \frac{n-k+1}{1} = \frac{n!}{(n-k)!k!} = \binom{n}{k}.$$

As for the other direction, we have $\binom{n}{k} \leq \frac{n^k}{k!} \leq \frac{n^k}{(k/e)^k} = \left(\frac{ne}{k}\right)^k$, by (iii). ■

6.2. The birthday paradox

Consider a group of n people, and assume their birthdays are uniformly distributed no the dates in the year (this assumption is not quite true, but close enough). We are interested in the question of how large n has to be till we get a collision – that is, two people with the same birthday. Intuitively, since the year has $m = 364$ days, the probability of person to land on a specific birthday is $p = 1/364$. So the natural guess would be that n needs to be approximately 364. Surprisingly, the answer is much smaller.

Lemma 6.2.1. *Let $X_1, \ldots, X_n$ be n variables picked uniformly, randomly and independently from $[\![m]\!] = \{1, \ldots, m\}$. Then, the expected number of collisions is $\binom{n}{2}/m$.*

Proof: Let $Y_{i,j} = 1 \iff X_i = X_j$. We have that $\mathbb{E}\Big[Y_{i,j}\Big] = \mathbb{P}\Big[Y_{i,j} = 1\Big] = 1/m$. Thus, the expected number of collisions is

$$\mathbb{E}\left[\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} Y_{i,j}\right] = \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \mathbb{E}\Big[Y_{i,j}\Big] = \binom{n}{2}\frac{1}{m}.$$ ■

As such, for birthdays, for $m = 364$, and $n = 28$, we have that the expected number of collisions is

$$\binom{28}{2}\frac{1}{364} = \frac{378}{364} > 1.$$

This seems weird, but is it the truth?

Lemma 6.2.2. *Let $X_1, \ldots, X_n$ be n variables picked uniformly, randomly and independently from $[\![m]\!] = \{1, \ldots, m\}$. Then, the probability that no collision happened is at most $\exp\Big(-\binom{n}{2}/m\Big)$.*

Proof: Let $\mathcal{E}_i$ be the event that X_i is distinct from all the values in $X_1, \ldots, X_{i-1}$. Let $\mathcal{B}_i = \cap_{k=1}^{i} \mathcal{E}_k = \mathcal{B}_{i-1} \cap \mathcal{E}_i$ be the event that all of $X_1, \ldots, X_i$ are distinct. Clearly, we have

$$\mathbb{P}[\mathcal{E}_i \mid \mathcal{B}_{i-1}] = \mathbb{P}[\mathcal{E}_i \mid \mathcal{E}_1 \cap \cdots \cap \mathcal{E}_{i-1}] = \frac{m-(i-1)}{m} = 1 - \frac{i-1}{m} \leq \exp\left(-\frac{i-1}{m}\right).$$

Observe that

$$\mathbb{P}[\mathcal{B}_i] = \mathbb{P}[\mathcal{B}_{i-1}]\frac{\mathbb{P}[\mathcal{E}_i \cap \mathcal{B}_{i-1}]}{\mathbb{P}[\mathcal{B}_{i-1}]} = \mathbb{P}[\mathcal{B}_{i-1}]\,\mathbb{P}[\mathcal{E}_i \mid \mathcal{B}_{i-1}] \leq \exp\left(-\frac{i-1}{m}\right)\mathbb{P}[\mathcal{B}_{i-1}] \leq \prod_{k=1}^{i} \exp\left(-\frac{k-1}{m}\right).$$

$$= \exp\left(-\sum_{k=1}^{i}\frac{k-1}{m}\right) = \exp\left(-\frac{i(i-1)}{2}\frac{1}{m}\right) = \exp\left(-\binom{i}{2}/m\right).$$

Which implies the desired claim for $i = n$. ■

6.3. Occupancy Problems

Problem 6.3.1. We are throwing m balls into n bins randomly (i.e., for every ball we randomly and uniformly pick a bin from the n available bins, and place the ball in the bin picked). There are many natural questions one can ask here:

(A) What is the maximum number of balls in any bin?
(B) What is the number of bins which are empty?
(C) How many balls do we have to throw, such that all the bins are non-empty, with reasonable probability?

Theorem 6.3.2. *With probability at least $1 - 1/n$, no bin has more than $k^* = \left\lceil \frac{3\ln n}{\ln\ln n} \right\rceil$ balls in it.*

Proof: Let X_i be the number of balls in the ith bins, when we throw n balls into n bins (i.e., $m = n$). Clearly,

$$\mathbb{E}[X_i] = \sum_{j=1}^{n} \mathbb{P}[\text{The } j\text{th ball fall in } i\text{th bin}] = n \cdot \frac{1}{n} = 1,$$

by linearity of expectation. The probability that the first bin has exactly i balls is

$$\binom{n}{i}\left(\frac{1}{n}\right)^i\left(1-\frac{1}{n}\right)^{n-i} \le \binom{n}{i}\left(\frac{1}{n}\right)^i \le \left(\frac{ne}{i}\right)^i\left(\frac{1}{n}\right)^i = \left(\frac{e}{i}\right)^i$$

This follows by Lemma 6.1.1 (iv).

Let $C_j(k)$ be the event that the jth bin has k or more balls in it. Then,

$$\mathbb{P}[C_1(k)] \le \sum_{i=k}^{n}\left(\frac{e}{i}\right)^i \le \left(\frac{e}{k}\right)^k\left(1+\frac{e}{k}+\frac{e^2}{k^2}+\ldots\right) = \left(\frac{e}{k}\right)^k \frac{1}{1-e/k}.$$

For $k^* = c\ln n/\ln\ln n$, we have

$$\begin{aligned}\mathbb{P}[C_1(k^*)] &\le \left(\frac{e}{k^*}\right)^{k^*}\frac{1}{1-e/k^*} \le 2\exp\Big(k^*(1-\ln k^*)\Big) \le 2\exp\left(-\frac{k^*\ln k^*}{2}\right)\\ &\le 2\exp\left(-\frac{c\ln n}{2\ln\ln n}\underbrace{\ln\frac{c\ln n}{\ln\ln n}}_{\approx \ln\ln n}\right) \le 2\exp\left(-\frac{c\ln n}{4}\right) \le \frac{1}{n^2},\end{aligned}$$

for n and c sufficiently large.

Let us redo this calculation more carefully (yuk!). For $k^* = \lceil(3\ln n)/\ln\ln n\rceil$, we have

$$\begin{aligned}\mathbb{P}[C_1(k^*)] &\le \left(\frac{e}{k^*}\right)^{k^*}\frac{1}{1-e/k^*} \le 2\left(\frac{e}{(3\ln n)/\ln\ln n}\right)^{k^*} = 2\exp\Big(\underbrace{1-\ln 3}_{<0} - \ln\ln n + \ln\ln\ln n\Big)^{k^*}\\ &\le 2\exp((-\ln\ln n + \ln\ln\ln n)k^*)\\ &\le 2\exp\left(-3\ln n + 6\ln n\frac{\ln\ln\ln n}{\ln\ln n}\right) \le 2\exp(-2.5\ln n) \le \frac{1}{n^2},\end{aligned}$$

for n large enough. We conclude, that since there are n bins and they have identical distributions that

$$\mathbb{P}[\text{any bin contains more than } k^* \text{ balls}] \le \sum_{i=1}^{n} C_i(k^*) \le \frac{1}{n}.$$ ■

Exercise 6.3.3. Show that when throwing $m = n\ln n$ balls into n bins, with probability $1 - o(1)$, every bin has $O(\log n)$ balls.

6.3.1. The Probability of all bins to have exactly one ball

Next, we are interested in the probability that all m balls fall in distinct bins. Let X_i be the event that the ith ball fell in a distinct bin from the first $i-1$ balls. We have:

$$\begin{aligned}\mathbb{P}[\cap_{i=2}^m X_i] &= \mathbb{P}[X_2]\prod_{i=3}^{m}\mathbb{P}\Big[X_i \;\Big|\; \cap_{j=2}^{i-1}X_j\Big] \le \prod_{i=2}^{m}\left(\frac{n-i+1}{n}\right) \le \prod_{i=2}^{m}\left(1-\frac{i-1}{n}\right)\\ &\le \prod_{i=2}^{m} e^{-(i-1)/n} \le \exp\left(-\frac{m(m-1)}{2n}\right),\end{aligned}$$

thus for $m = \lceil \sqrt{2n} + 1 \rceil$, the probability that all the m balls fall in different bins is smaller than $1/e$.

This is sometime referred to as the *birthday paradox*, which was already mentioned above. You have $m = 30$ people in the room, and you ask them for the date (day and month) of their birthday (i.e., $n = 365$). The above shows that the probability of all birthdays to be distinct is $\exp(-30 \cdot 29/730) \leq 1/e$. Namely, there is more than 50% chance for a birthday collision, a simple but counter-intuitive phenomena.

6.4. The Coupon Collector's Problem

There are n types of coupons, and at each trial one coupon is picked in random. How many trials one has to perform before picking all coupons? Let m be the number of trials performed. We would like to bound the probability that m exceeds a certain number, and we still did not pick all coupons.

Let $C_i \in \{1, \ldots, n\}$ be the coupon picked in the ith trial. The jth trial is a success, if C_j was not picked before in the first $j-1$ trials. Let X_i denote the number of trials from the ith success, till after the $(i+1)$th success. Clearly, the number of trials performed is

$$X = \sum_{i=0}^{n-1} X_i.$$

Furthermore, X_i has a geometric distribution with parameter p_i, that is $X_i \sim \text{Geom}(p_i)$, with $p_i = (n-i)/n$. The expectation and variance of X_i are

$$\mathbb{E}[X_i] = \frac{1}{p_i} \qquad \text{and} \qquad \mathbb{V}[X_i] = \frac{1 - p_i}{p_i^2}.$$

Lemma 6.4.1. *Let X be the number of rounds till we collection all n coupons. Then, $\mathbb{V}[X] \approx \left(\pi^2/6\right)n^2$ and its standard deviation is $\sigma_X \approx (\pi/\sqrt{6})n$.*

Proof: The probability of X_i to succeed in a trial is $p_i = (n-i)/n$, and X_i has the geometric distribution with probability p_i. As such $\mathbb{E}[X_i] = 1/p_i$, and $\mathbb{V}[X_i] = q/p^2 = (1-p_i)/p_i^2$.

Thus,

$$\mathbb{E}[X] = \sum_{i=0}^{n-1} \mathbb{E}[X_i] = \sum_{i=0}^{n-1} \frac{n}{n-i} = nH_n = n(\ln n + \Theta(1)) = n \ln n + O(n),$$

where $H_n = \sum_{i=1}^{n} 1/i$ is the nth Harmonic number.

As for variance, using the independence of $X_0, \ldots, X_{n-1}$, we have

$$\begin{aligned}
\mathbb{V}[X] &= \sum_{i=0}^{n-1} \mathbb{V}[X_i] = \sum_{i=0}^{n-1} \frac{1-p_i}{p_i^2} = \sum_{i=0}^{n-1} \frac{1-(n-i)/n}{\left(\frac{n-i}{n}\right)^2} = \sum_{i=0}^{n-1} \frac{i/n}{\left(\frac{n-i}{n}\right)^2} = \sum_{i=0}^{n-1} \frac{i}{n}\left(\frac{n}{n-i}\right)^2 \\
&= n \sum_{i=0}^{n-1} \frac{i}{(n-i)^2} = n \sum_{i=1}^{n} \frac{n-i}{i^2} = n\left(\sum_{i=1}^{n} \frac{n}{i^2} - \sum_{i=1}^{n} \frac{1}{i}\right) = n^2 \sum_{i=1}^{n} \frac{1}{i^2} - nH_n \approx \frac{\pi^2}{6} n^2,
\end{aligned}$$

since $\lim_{n\to\infty} \sum_{i=1}^{n} \frac{1}{i^2} = \pi^2/6$, we have $\lim_{n\to\infty} \dfrac{\mathbb{V}[X]}{n^2} = \dfrac{\pi^2}{6}$. ■

This implies a weak bound on the concentration of X, using Chebyshev inequality, we have

$$\mathbb{P}\left[X \geq n \ln n + n + t \cdot n \frac{\pi}{\sqrt{6}}\right] \leq \mathbb{P}\left[\left|X - \mathbb{E}[X]\right| \geq t\sigma_X\right] \leq \frac{1}{t^2},$$

Note, that this is somewhat approximate, and hold for n sufficiently large.

6.5. Notes

The material in this note covers parts of [MR95, sections 3.1, 3.2, 3.6]

References

[MR95] R. Motwani and P. Raghavan. *Randomized Algorithms*. Cambridge, UK: Cambridge University Press, 1995.