27 Jan 2025

The Birthday Paradox

"Balls and bins" problems involve a sequence of random samples $X_1, X_2, \ldots, X_m$ each drawn uniformly at random (independently) from $[n] = \{1, \ldots, n\}$.

Two viewpoints:

- simultaneous: $X_1, \ldots, X_m$ sampled all at once.
- sequential: picture sequence $X_1, \ldots$ generated one element at a time. Ask questions about the (random) time when some property is achieved.

Occupancy vector: n-dimensional vector counting # balls in each bin.

E.g. $(X_1, X_2, \ldots, X_8) = (2,1,3,3,1,1,3,2)$

occupancy vector $\begin{bmatrix} 3 \\ 2 \\ 3 \end{bmatrix}$

Questions.

- Birthday paradox: when does occupancy vector leave the set $\{0,1\}^n$? (First time we have a bin with 2 or more balls in it.) (Never later than $m = n+1$, probably much earlier.)

- Coupon collector: when does min(occupancy) exceed $\emptyset$?
- Load balancing: when does $\frac{\max(\text{occupancy})}{\min(\text{occupancy})}$ become less than $1+\varepsilon$?

Fact. The collision probability for m balls in n bins, i.e. $\Pr(\exists\, i \neq j \text{ s.t. } X_i = X_j)$, can be calculated as follows.

$$\Pr(\text{collision}) = 1 - \Pr(\text{no collision})$$

$$= 1 - \prod_{j=1}^{m} \Pr\left(X_j \text{ is not equal to any } X_i \text{ s.t. } i<j \;\middle|\; X_1, \ldots, X_{j-1} \text{ all distinct}\right)$$

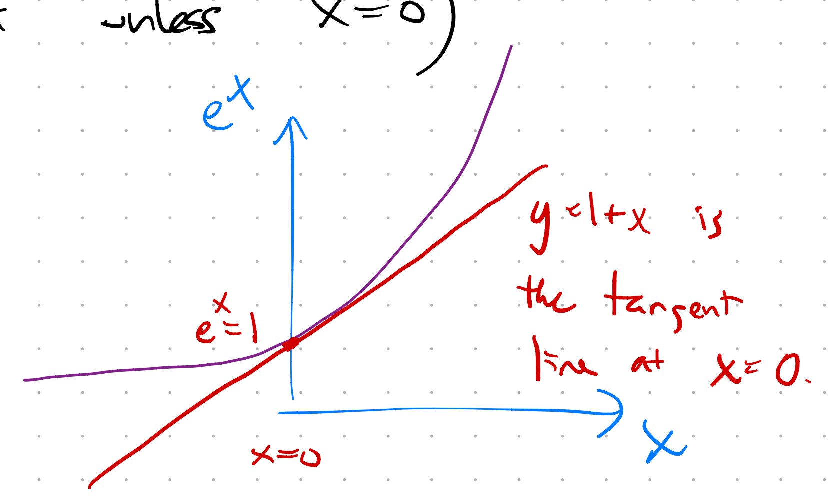
$$= 1 - \prod_{j=1}^{m} \left(1 - \frac{j-1}{n}\right) \qquad (\text{substitute } k = j-1)$$

$$= 1 - \prod_{k=1}^{m-1} \left(1 - \frac{k}{n}\right)$$

The most useful inequality in the analysis of randomized algorithms:

$$\forall x \in \mathbb{R} \qquad 1 + x \leq e^x = \exp(x)$$

(the inequality is strict unless $x=0$)



$$\Rightarrow \quad 1-\frac{k}{n} < e^{-k/n}$$

$$\prod_{k=1}^{m-1}\left(1-\frac{k}{n}\right) < \prod_{k=1}^{m-1} e^{-k/n} = \exp\left(-\sum_{k=1}^{m-1}\frac{k}{n}\right)$$

$$= \exp\left(-\frac{m(m-1)}{2n}\right)$$

To make Pr(collision) $> \frac{1}{2}$ we want

Pr(no collision) $< \frac{1}{2}$, and we've shown

$$\Pr(\text{no collision}) < \exp\left(-\frac{m(m-1)}{2n}\right)$$

choose smallest m that makes this $\leq \frac{1}{2}$.

$$\exp\left(-\frac{m(m-1)}{2n}\right) \leq \frac{1}{2}$$

$$\exp\left(\frac{m(m-1)}{2n}\right) \geq 2$$

$$\frac{m(m-1)}{2n} \geq \ln(2)$$

$$m^2 - m \geq 2n\ln(2).$$

$$m^2 - m + \frac{1}{4} \geq 2n\ln(2) + \frac{1}{4}$$

tight within additive $\pm O(1)$

$$m - \frac{1}{2} \geq \sqrt{2n\ln(2) + 1/4}$$

$$\boxed{m \geq \underbrace{\sqrt{2n\ln(2) + 1/4}} + \frac{1}{2}} = O(\sqrt{n})$$

Plug in $n = 365$ to you get $m \geq 22.99994\ldots$

We have estimated $\Pr(\text{collision}) > 1 - \exp\left(-\frac{m(m-1)}{2n}\right)$.

How close is RHS to true probability?

$$??? < 1 - \frac{k}{n} < e^{-k/n}$$

<u>Fact.</u> For $0 < x < \frac{1}{2}$, $1 - x > e^{-x-x^2}$

<u>Proof.</u> $\ln(1-x) = -\sum_{i=1}^{\infty} \frac{x^i}{i} = -x - \frac{x^2}{2} - \frac{x^3}{3} - \frac{x^4}{4} \ldots$

$$> -x - \frac{x^2}{2} - \frac{x^3}{2} - \frac{x^4}{2} \ldots$$

$$= -x - \frac{1}{2}x^2\left(1 + x + x^2 + x^3 + \ldots\right)$$

$$= -x - \frac{x^2}{2(1-x)} > -x - x^2$$

Exponentiate both sides $\Rightarrow$ $1 - x > e^{-x-x^2}$.

$$\exp\left(-\frac{k}{n} - \frac{k^2}{n^2}\right) < 1 - \frac{k}{n} < e^{-k/n}$$

So $\Pr(\text{no collision}) > \prod_{k=1}^{m-1} \exp\left(-\frac{k}{n} - \frac{k^2}{n^2}\right)$

$$= \exp\left(-\sum_{k=1}^{m-1} \frac{k}{n} - \sum_{k=1}^{m-1} \frac{k^2}{n^2}\right)$$

$$= \exp\left(-\frac{m(m-1)}{2n} - \frac{m(m-1)(2m-1)}{6n^2}\right)$$

Upper and lower bounds differ by

$$\exp\left(-\frac{m(m-1)(2m-1)}{6n^2}\right)$$

Recall we chose $m \approx \sqrt{2n\ln(2)}$, $\frac{m(m-1)}{2n} \approx \ln(2)$

$$\exp\left(-\boxed{\frac{m(m-1)}{2n}}\frac{(2m-1)}{3n}\right) \approx \exp\left(-\frac{(2m-1)\ln(2)}{3n}\right)$$

$$\approx \exp\left(-\frac{2\sqrt{2n\ln 2}\cdot\ln 2}{3n}\right)$$

$$= \exp\left(-\frac{(2\ln 2)^{3/2}}{3}\cdot\frac{1}{\sqrt{n}}\right)$$

$$> \exp\left(-\frac{1}{\sqrt{n}}\right) > 1-\frac{1}{\sqrt{n}}$$

$$\left(1-\frac{1}{\sqrt{n}}\right)\exp\left(-\frac{m(m-1)}{2n}\right) < \Pr(\text{no collision}) < \exp\left(-\frac{m(m-1)}{2n}\right)$$

APPLICATION 1. Cryptography

Cryptographic hash functions are functions $h(x)$ that are:

1. Easy to compute, deterministic
2. Output values in $\{0,1\}^k$ security parameter.
3. Believed to be computationally hard to find $x_0 \neq x_1$ with $h(x_0) = h(x_1)$.

Often modeled as if $h(x_1), h(x_2), \ldots, h(x_m)$ for any fixed m-tuple of inputs $x_1 \ldots x_m$ are computationally indistinguishable from indep. random unif samples from $\{0,1\}^k$.

If they truly were indep random samples, then $h(x_1), \ldots, h(x_m)$ would be m balls thrown randomly into $n = 2^k$ bins.

The "birthday attack" — hashing $h(x_i)$ for independent random $x_1, \ldots, x_m$ until a collision occurs — takes $\approx \sqrt{2n} = 2^{(k+1)/2}$ trials to succeed.

Crypto hash functions considered secure if nobody knows how to find collision using $\ll 2^{\frac{k+1}{2}}$ tries.

<u>E.g.</u> SHA-1 has 160-bit output.

$k = 160$, $\frac{k+1}{2} \approx 80$

Xiaoyun Wang + collaborators: found collision using 2^{63} trials, more than 10^5 faster than random guessing.

SHA-2 is now standard, uses k ranging from 224 to 512.