26 Jan 2026 The Birthday Paradox

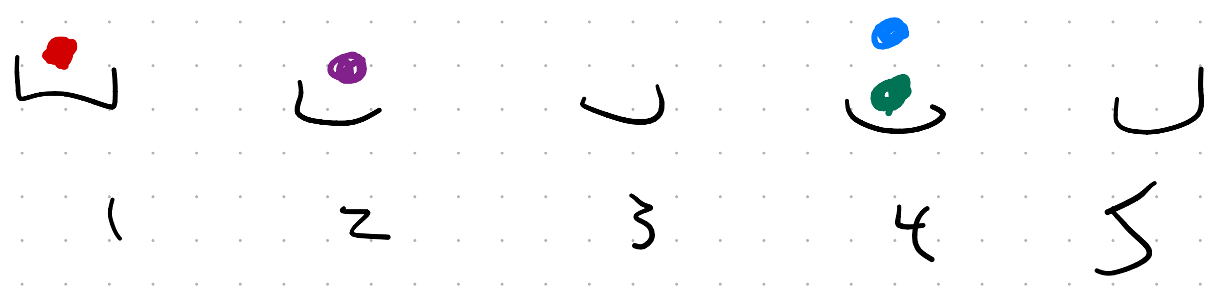
Announcements

1. Textbook posted on Canvas ("Textbook" module) and website ("Lecture" tab, see assigned readings)
2. Office hours posted on website + Canvas.
   (David's Fri. OH canceled this week)
3. Weds quiz covers **ONLY TODAY'S LECTURE.** Generally it will cover the prior Wed + Mon.

After m balls have been thrown into n bins (independently, each uniformly random) what does the n-dimensional vector of bin occupancy look like?

Occupancy vector $\vec{v} \in \mathbb{R}^n$ has

v_i = # balls in bin i



$m = 4$

$n = 5$

$$v = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 2 \\ 0 \end{bmatrix}$$

Questions,

- How likely that no collisions?

$$(v \in \{0,1\}^n)$$

"Birthday Paradox"

- How likely that each bin occupied?

$$(v \in \mathbb{R}^n_{>0})$$

"Coupon Collector Problem"

- How likely that the loads/occupancies are well balanced?

$$\frac{v_{max}}{v_{min}} < 1+\varepsilon \qquad (\varepsilon > 0)$$

"Load Balancing"

Say m balls, n bins.

$\Pr(\text{no collision})$

$$= \Pr(\text{ball 2 diff bin from ball 1})$$

$$\cdot \Pr(\text{ball 3 diff from } 1,2 \mid \text{no collision yet})$$

$$\cdot \Pr(\text{ball 4} \ldots 1,2,3 \mid \text{no collision yet})$$

$$\vdots$$

$$\cdot \Pr(\text{ball } m \text{ diff from } 1,2,\ldots,m-1 \mid \text{no collision yet})$$

$$= \left(\frac{n-1}{n}\right)\cdot\left(\frac{n-2}{n}\right)\cdots\left(\frac{n-m+1}{n}\right)$$

When $n=365$, $m=23$, this is $< \frac{1}{2}$.

Factorization above is justified by repeated application of the rule

$$\Pr(\mathcal{E}_a \cap \mathcal{E}_b) = \Pr(\mathcal{E}_a)\cdot\Pr(\mathcal{E}_b \mid \mathcal{E}_a).$$

$$\Pr(\underbrace{\mathcal{E}_1 \cap \ldots \mathcal{E}_{k-1}}_{\mathcal{E}_a} \cap \underbrace{\mathcal{E}_k}_{\mathcal{E}_b}) = \Pr(\mathcal{E}_1)\cdot\Pr(\mathcal{E}_2 \mid \mathcal{E}_1)\cdot\Pr(\mathcal{E}_3 \mid \mathcal{E}_2 \cap \mathcal{E}_1) \cdots \Pr(\mathcal{E}_k \mid \mathcal{E}_{k-1} \cap \mathcal{E}_{k-2} \cap \ldots \cap \mathcal{E}_1).$$

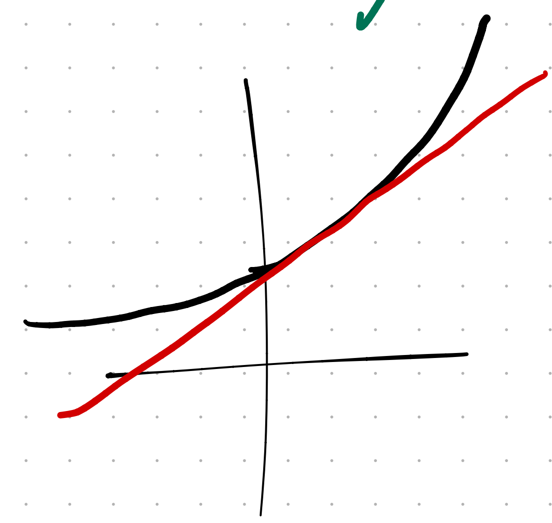
In the above, $\mathcal{E}_k$ = no collisions of balls $1,\ldots,k$.

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

The most important inequality in rand. algs...

$$\forall x \in \mathbb{R} \qquad 1 + x \le e^x$$

(equality iff $x = 0$)



$$\Pr(\text{no collision}) < \exp\left(-\sum_{k=1}^{m-1} \frac{k}{n}\right)$$

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

RHS is $\le \frac{1}{2}$ when

$$-\frac{m(m-1)}{2n} \le \ln\left(\frac{1}{2}\right)$$

$$m^2 - m \ge 2n \ln(2)$$

$$m \ge \sqrt{2n \ln(2) + \frac{1}{4}} + \frac{1}{2}$$

At $n = 365$ this equals 22.99994...

Good enough for most purposes:

$\Pr(\text{no collision}) < \frac{1}{2}$ when $m \ge \sqrt{2n}$

Also good enough for many purposes

$\Pr(\text{no collision}) < \frac{1}{2}$ when $m = \Omega(\sqrt{n})$

How to bound $\Pr(\text{no collision})$ from below

$$1 - x \geq e^{-x-x^2} \qquad 0 \leq x \leq \frac{1}{2}$$

Why true?

$$\ln(1-x) \geq -x - x^2$$
$$\parallel$$
$$-x - \frac{1}{2}x^2 - \frac{1}{3}x^3 - \frac{1}{4}x^4 - \ldots$$

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

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

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

$$\frac{m(m-1)(2m-1)}{6n^2} = \ln(2) \cdot \frac{2m-1}{3n}$$

$\Theta\left(\frac{m}{n}\right) = \Theta\left(\frac{1}{\sqrt{n}}\right).$

$$\frac{1}{2} \geq \prod_{k=1}^{m-1} \left(1 - \frac{k}{n}\right) \geq \frac{1}{2} \cdot \exp\left(-\frac{c}{\sqrt{n}}\right)$$

The true smallest m such that $Pr(\text{no collision}) \leq \frac{1}{2}$ lies between

$\sqrt{2n \ln(2)} - 2$ and $\sqrt{2n \ln(2) + \frac{1}{4}} + \frac{1}{2}$.

CRYPTOGRAPHIC HASH FUNCTIONS.

E.g. MD5, SHA-1, SHA-2.

Compress file of arbitrary length to a bit-string of predetermined length

E.g. 224 to 512 bits for SHA-2.

For a hash with b-bit output, there are $n = 2^b$ potential values.

So, if you randomly guess

m distinct inputs,

you will succeed in

finding hash collision

w. prob. $> \frac{1}{2}$

provided $m > \sqrt{2n} = \sqrt{2 \cdot 2^b}$

$$= 2^{\frac{b+1}{2}}.$$