L2: Statistical Phenomena

Aug 20, 2025

Data Mining

Jeff Phillips

Data $X = \{x_1, x_2, \dots x_n\} \underset{\text{iid}}{\sim} D(\theta)$

param.

independent identically Distributed.

Central Limit Theorem

as $n \to \infty$, get better estimates

$X \subset [m] = \{0, 1, 2, \ldots, m-1\}$

- IP addresses $m = 10^{16}$
- words in English $m = 100{,}000$
- people in USA $m = 360$ M.

$D(\theta) = \text{Unif}([m])$

$$\Pr(x_i = j) = \begin{cases} \frac{1}{m} & \text{if } j \in [m] \\ 0 & \text{otherwise} \end{cases}$$

Hash Function

Choose $h_a \sim \mathcal{H}$

$h : \Sigma^* \to [m]$ Deterministic

SHA1

$$\Pr_{h_a \sim \mathcal{H}}\left[h_a(\text{string}) = j \right] = \begin{cases} \frac{1}{m} & \text{if } j \in [m] \\ 0 & \text{o.w.} \end{cases}$$

$a = \text{salt}$

input

salt

$$v = \text{hash}(\underbrace{\text{string} + a}_{\text{concat}})$$

Birthday Paradox

m = unique days year

365

Jan	26, 12, 1
Feb	26
Mar	6, 19
Apr	30, 5
May	8, 14, 25, 11
June	26, 21, 23
July	10
Aug	9, 16, 22, 30, 29, 15, 24
Sep	21, 15, 29, 11
Oct	17, 24
Nov	
Dec	

$Pr(\text{collision } n \text{ trials})$

$Pr(\text{coll } n=1) = 0$

$Pr(\text{coll } n=2) = \frac{1}{365}$

$Pr(\text{coll } n=3) = 1-\left(1-\frac{1}{365}\right)^3$

$Pr(\text{coll } n) = 1-\Big(\underbrace{1-\frac{1}{365}}_{0.997}\Big)^{\binom{n}{2}}$ $\quad \binom{n}{2} \approx n^2/2$

when first collision

$n \approx \sqrt{2m}$

if $n=23$ $\quad \binom{23}{2} = 253$

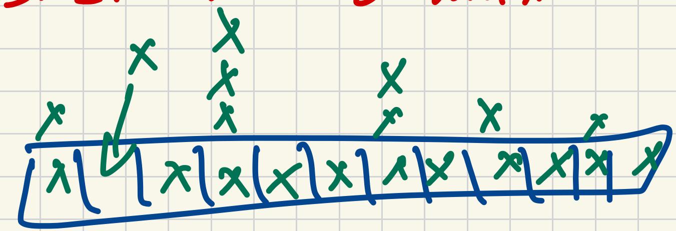
$1-(0.997)^{253} \approx 0.532$

What is wrong?

- birthday are not uniform?
- leap year $n = 365$ Feb 29
  also not unif
- math correlation
  in pairs
- sample independent?
  (twins)

Coupon Collectors

Size m domain $\to$ see all distinct elements $j \in [m]$



How many trials until see all $[m]$

- at least m.

$2m$

$m \log m$

m^2

m^3

2^m

$$\gamma = 0.577\ldots$$

$$E[\#\text{trials}] = m(\gamma + \ln m)$$

t_j = # trials between jth new coupon and the $(j+1)$th new coupon

$$t_0 = 1$$

$$E[t_j] = \frac{m}{m-j}$$

$$j = \frac{m}{2} \qquad \frac{m}{m-j} = \frac{m}{m-\frac{m}{2}} = \frac{m}{m/2} = 2$$

$$j = m-1 \qquad \frac{m}{m-(m-1)} = m$$

m-th Harmonic num

$$T = \sum_{j=0}^{m-1} t_j \qquad E[T] = \sum_{j=0}^{m-1} E[t_j] = \sum_{j=0}^{m-1} \frac{m}{m-j}$$

$$= m \cdot \sum_{j=0}^{m-1} \frac{1}{m-j} = m \cdot \underbrace{\sum_{j=1}^{m} \frac{1}{j}}_{H_m}$$

Summarize

- To use <u>math</u> we need simple models.
- Models are wrong, but sometimes ok.
- non-trivial phenomena
  - collision $\approx \sqrt{2m}$
  - see all $\approx m \log m$