UU: Data Mining

# L2: Statistical Phenomena

Jan 8, 2025

## Data

$X = \{x_1, x_2, \dots, x_n\}$

$X \underset{iid}{\sim} D(\theta)$

distribution, parameter $\theta$

observations of the world

each $x_i$ is one observation

$X \underset{iid}{\sim} D(\theta)$

drawn from

iid: independently and identically distributed

Assume domain of $D(\theta)$ and data $X$

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

models: IP addresses $m \approx 16^{10}$

words

$m = 100{,}000$

all people in USA = 3.8 million

assume $X_i \sim D(\theta)$

$\Pr[X_i = j] = \frac{1}{m}$ if $j \in [m]$

$= 0$ o.w.

random hash function

$h : \Sigma^t \to [m]$

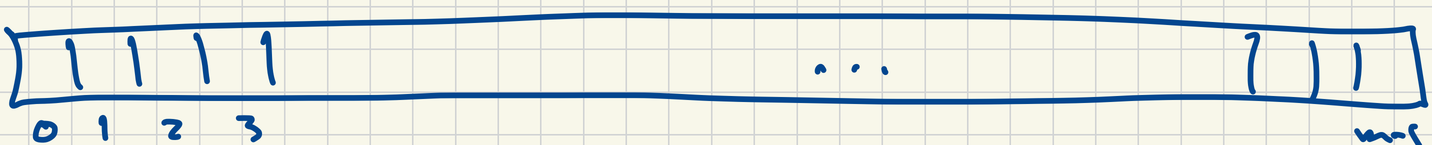
$\Pr_{h \sim \mathcal{H}}[h(i) = h(j)] = \frac{1}{m}$ $\quad i \neq j$

h: deterministic

$h_a \sim \mathcal{H}$ $\quad$ a: salt $\leftarrow$ chosen at random

$h_a = f(\text{string}, a)$

f: complex $\quad$ string $\in \Sigma^t$

$[m]$ :


Complex Deterministic Function

$$f : (\Sigma_1^t, salt) \to [m]$$

- SHA-1 : hard to invert $f$

$\boxed{f^{-1} ?}$

$\to [m] \quad m = 2^{160}$

- Multiplication Hashing $\quad x = \Sigma^* = \mathbb{R}$

$$h_a(x) = \lfloor m \cdot frac(x \cdot a) \rfloor$$

$$\lfloor (x \cdot a / 2^8) \rfloor$$

$frac(17.23) = 0.23$

$\lfloor 17.23 \rfloor = 17$

~~Do not~~ do modular hashing

$$h(x) = x \bmod m$$

# Birthday Paradox

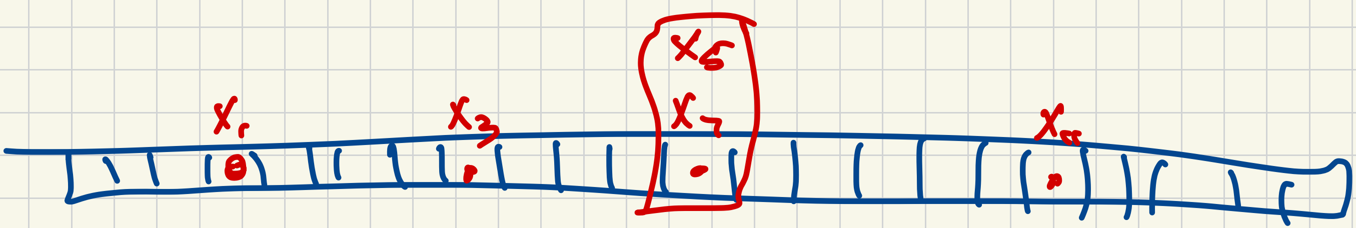
$m = 365$ $\quad n = 23$

Draw $x_i \sim X$ $\qquad x_i \in [m]$ $\quad \Pr[x_i = j] = \frac{1}{m}$

$x_1, x_2, \dots x_n$

When do I expect that two $x_i = x_j$
$i \neq j$

Answer: $n \approx \sqrt{2m}$



| | |
|---|---|
| Jan | |
| Feb | 14 |
| Mar | 20 |
| Apr | |
| May | 23 |
| Jun | 3, 26 |
| Jul, 3 | 21 |
| Aug | |
| Sep | |
| Oct | |
| Nov | |
| Dec | |

x6

21, 20, 2 collision of n = 6

19, 2, 18, 12

31, 19, 24

29, 27, 23

24, 14, 22, 30, 29, 7, 21

20, 17

25, 1, 15, 5, 11

17, 31, 5

4, 21, 9, 24

25, 11, 2, 7, 21, 12

22, 20

8, 14, 12

x46

$\Pr[\text{coll}, n=1] = 0$

$n = 23$

$\Pr[\text{coll}, n=2] = \frac{1}{m}$

$\Pr[\text{coll}, n=3] \approx 1-\left(1-\frac{1}{m}\right)^3$

$\binom{n}{2} = \#\text{ pairs}$

$\Pr[\text{coll}, n] \approx 1-\left(1-\frac{1}{m}\right)^{\binom{n}{2}}$

$= \frac{n(n-1)}{2} \approx \frac{n^2}{2}$

$\approx 1-\left(1-\frac{1}{m}\right)^{n^2/2}$

$\approx 0.048$

$m=365$

$n=6$

$\approx 1 - 0.997^{253} = 0.532$

$n=23$

# What is wrong w/ this analysis?

- bias: more b-days in Spring.

  leap year $m = 366$ (Feb 29)

  not real b-day

- set of pairs not iid.

What happens when $n = m+1$

$$1 - \left(1 - \frac{1}{m}\right)^{\binom{m+1}{2}} \approx 1 - 0.0000007$$

$$1 - \left(\frac{m-1}{m}\right)\left(\frac{m-2}{m}\right)\left(\frac{m-3}{m}\right)\cdots(0) = 1$$

## Coupon Collectors

$X_1, \dots, X_n \overset{\sim}{iid} \mathrm{Unif}([m])$

How many draws $n$ until I see **all** $j \in [m]$ so some $X_i = j$?

$n \approx 4m^2$?

$\approx 5m$

$m \log m = m(0.577 + \ln m)$

$$E[n] = m \boxed{\sum_{i=1}^{m} \frac{1}{i}} \Rightarrow H_m \quad \text{harmonic number}$$

$$\approx (0.577 + \ln m)$$