L2: Statistical Phenomena

Data Mining
Aug 27, 2026
Jeff Phillips

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

parameters

independently and identically distributed

statistics $n \to \infty$

Central Limit Theorem

each $X_i \in [m] = \{0, 1, 2, \dots m-1\}$

- IP addresses $m = 10^{16}$
- words in English $m = 100{,}000$
  tokens $300{,}000$
- people world $m = 6B$
  US $m = 360M$

$$X_i \underset{\text{iid}}{\sim} D(\theta) = \text{Unif}([m])$$

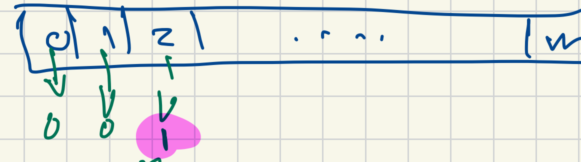
each $j \in [m]$ $\Pr[X_i = j] = \frac{1}{m}$

$j \notin [m]$ $\Pr[X_i = j] = 0$

- natural : Hash Table
- represents no structure.

Hash Tables

array $A = [m]$



deterministic

hash function $h: \underbrace{\Sigma^*}_{\text{string}} \to [m]$

Uniform hash $\Pr\left[h[j] = h[j']\right] = \frac{1}{m}$ $\quad j, j' \in \Sigma^*$, $j \neq j'$

salt s ← randomly choosen string

$h[j \text{ concat } s]$, $h[j' \text{ concat } s]$

SHA-1

Birthday Paradox

$m = 365$ (days of year)

Jan	25, 12, 17, 27
Feb	2, 17, 28, 16
Mar	30, 23, 31
Apr	12, 6
May	14, 23, 13, 29, 10, 31
June	2, 16, 28
July	1, 18, 28
Aug	18, 16, 10, 19
Sep	22, 13
Oct	8
Nov	25, 13, 20
Dec	26, 24, 14

38

$\Pr(\text{coll } n \text{ samples})$ $\qquad m = 365$

$\Pr[n=1] = 0$ $\qquad \binom{n}{2}$ n choose 2

$\Pr[n=2] = \frac{1}{365} = \frac{1}{m}$ $\qquad \frac{n(n-1)}{2} \approx \frac{n^2}{2}$

$\Pr[n=3] = 1-\left(1-\frac{1}{365}\right)^3 = \left(\frac{1}{365}\right)\left(\frac{1}{364}\right)$

$\Pr[n] \approx 1-\left(1-\frac{1}{365}\right)^{\binom{n}{2}} \approx 1-\left(1-\frac{1}{365}\right)^{\frac{n^2}{2}} = 0.5 \Rightarrow n = 23$

$1-(0.997)^{n^2/2}$ $\qquad n = 38$

$1-(0.997)^{38^2/2} = 1-0.1142$

0.882

$\Pr[\text{coll for } n] = 0.5$

$n = \sqrt{2m}$

$m = 20{,}000$

$2m = 40{,}000$ $\qquad \sqrt{2m} = 200$

What was wrong w/ this analysis?

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

- birthdays not uniform
- no leap year $m = 366$ (Feb 29)?
- n gets large? $n > m$
- twins

Coupon Collector

m coupons

collect them all.

How many packs (n) to open to get all m?

$n \overset{?}{=} 2m$

$= m$

$= 2^m$

r_i = number trials to get i^{th} distinct pokemon

$t_i = r_i - r_{i-1}$ = num trials to go from $(i-1)$ to (i)

$r_1 = t_1 = 1$

$$E[t_i] = \frac{1}{\Pr[\text{new}]} = \frac{1}{\left(\frac{m-i}{m}\right)} = \frac{m}{m-i}$$

$$r_m = T = \sum_{i=1}^{m} t_i$$

$$= \sum_{i=1}^{m} \frac{m}{m-i} = \boxed{m \sum_{j=1}^{m} \frac{1}{j}} \quad j = m - i$$

$$\Longrightarrow T = m(0.577 + \ln m)$$

H_m = Harmonic number

$= \underset{0.577}{\gamma} + \ln m + o(1)$

t_i: $\quad i = [0, m/2] \qquad E[t_i] \approx 1 \cdot 2 \Rightarrow 2$

$E[t_i] \quad i = [m/2, 3m/4] \Rightarrow 4$

$[3m/4, 7m/8] \Rightarrow 8$