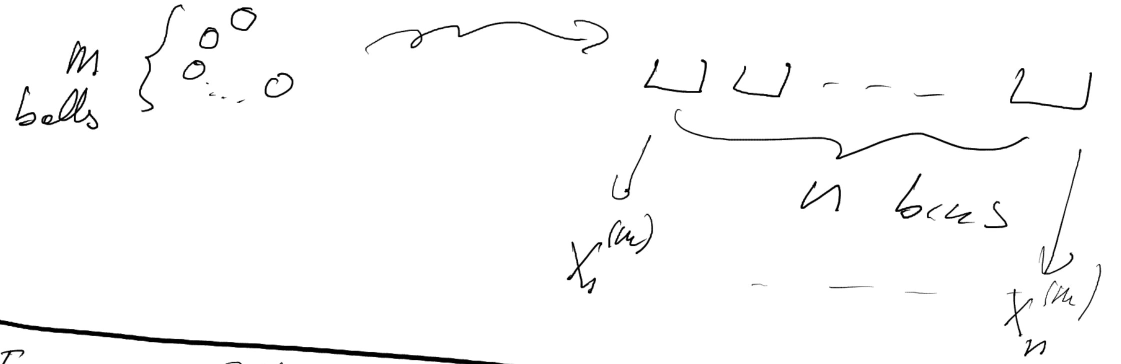
Balls & bins

Recall: we have m balls
each is thrown unif. at random
into one of n bins

$X_i^{(m)} :=$ # of balls in the i-th bin
$(i = 1, \dots, n)$



E.g.: $n = 365$, Q: $\max\limits_{i \le 365} X_i^{(m)} \ge 2$? ... birthday paradox

E.g.: analyze bucket sort

Lemma $m = n \Rightarrow \Pr\left[\exists i : X_i^{(m)} > \frac{3 \ln n}{\ln \ln n}\right] < \frac{1}{n}$ (n suff. large)

max is large

Proof $\Pr[X_1^{(m)} \ge M] \le \binom{n}{M} \frac{1}{n^M} \le \frac{1}{M!} \le \left(\frac{e}{M}\right)^M$ (OR Chernoff!)

$\Pr[\exists i \dots] \le n \left(\frac{e}{M}\right)^M \le \dots \le \frac{1}{n}$

$\square$

Lemma $m = n \Rightarrow \Pr\left[\max_i X_i^{(m)} < \frac{\ln n}{\ln \ln n}\right] \le \frac{1}{n}$ (n suff. large)

Proof The rest of this class.

Reason?

Poisson distribution

r pevné, $m, n \gg 1$

$$\Pr[X_1^{(m)} = r] = \binom{m}{r} \frac{1}{n^r} \left(1 - \frac{1}{n}\right)^{m-r} \doteq \frac{m^r}{r!} \cdot \frac{1}{n^r} \left(1 - \frac{1}{n}\right)^m \doteq \frac{1}{r!} \left(\frac{m}{n}\right)^r e^{-\frac{m}{n}}$$

Df Discrete Poisson rand. var. X with par. μ ($X \sim Pois(\mu)$) is any r.v. that sat. $\Pr[X = k] = \frac{e^{-\mu} \mu^k}{k!}$ ($k = 0, 1, 2, \dots$)

①_1 Def. is correct

①_2 $\mathbb{E} X = \mu$

①_3 $\operatorname{Var} X = \mu$

indep. r.v.

①_4 **Lemma** $X \sim Pois(\lambda)$, $Y \sim Pois(\mu) \Rightarrow X + Y \sim Pois(\lambda + \mu)$ (also for finitely many terms)

Pf $$\Pr[X+Y=n] = \sum_{k=0}^{n} \Pr[X=k] \cdot \Pr[Y=n-k] = \sum_k \frac{e^{-\lambda} \lambda^k}{k!} \cdot e^{-\mu} \frac{\mu^{n-k}}{(n-k)!} = \frac{e^{-(\lambda+\mu)}}{n!} \sum \binom{n}{k} \lambda^k \mu^{n-k} \quad \square$$

①_5 **Thm** (Limit of Bin. distr.) $X_n \sim Bin(n, p_n)$, $\lim n p_n = \lambda \in \mathbb{R}$, k fixed

$\Leftrightarrow \lim \Pr[X_n = k] = e^{-\lambda} \lambda^k / k!$ $\Rightarrow$ Applications

Poisson approximation

$X_1^{(m)}, \ldots, X_n^{(m)}$ — each $\overset{\text{approx.}}{\sim} \mathrm{Pois}\left(\frac{m}{n}\right)$
— not independent! $\quad$ (as $\sum X_i^{(m)} = m$)

$Y_1^{(m)}, \ldots, Y_n^{(m)}$ ... indep. r.v.'s, each $\sim \mathrm{Pois}\left(\frac{m}{n}\right)$

WANT: approximate $(X_1^{(m)}, \ldots, X_n^{(m)})$ by $\underbrace{(Y_1^{(m)}, \ldots, Y_n^{(m)})}_{\boldsymbol{Y}}$

Thm $\forall k, m, n$: distribution of $(Y_1^{(m)}, \ldots, Y_n^{(m)})$ conditioned on $\underbrace{\sum_{i=1}^n Y_i^{(m)}}_{\sim \mathrm{Pois}(m)} = k$
is the same as $\underbrace{(X_1^{(k)}, \ldots, X_n^{(k)})}_{\boldsymbol{X}}$.

Proof Just count ... pick $\underbrace{(k_1, \ldots, k_n)}_{\boldsymbol{k}}$ s.t. $\sum_{i=1}^n k_i = k$

$$\Pr[\boldsymbol{X} = \boldsymbol{k}] = \binom{k}{k_1 \ldots k_n} \cdot \frac{1}{n^k}$$

$$\Pr[\boldsymbol{Y} = \boldsymbol{k}] = \prod_{i=1}^n \left(e^{-\frac{m}{n}} \cdot \frac{\left(\frac{m}{n}\right)^{k_i}}{k_i!} \right) = e^{-m} \cdot \left(\frac{m}{n}\right)^k \cdot \frac{1}{k!} \cdot \binom{k}{k_1, \ldots, k_n}$$

$$\Pr\left[\sum Y_i = k\right] = \Pr[\mathrm{Pois}(m) = k] = e^{-m} \frac{m^k}{k!}$$

$$\Pr\left[\boldsymbol{Y} = \boldsymbol{k} \mid \sum Y_i^{(m)} = k\right] = \frac{\Pr[\boldsymbol{Y} = \boldsymbol{k}]}{\Pr\left[\sum_{i=1}^n Y_i^{(m)} = k\right]} = \Pr[\boldsymbol{X} = \boldsymbol{k}]$$

Thm $f(x_1, \dots, x_n) \geq 0$ — any fction

$$\mathbb{E}\left[f(X_1^{(m)}, \dots, X_n^{(m)})\right] \leq e\sqrt{m}\, \mathbb{E}\left[f(Y_1^{(m)}, \dots, Y_n^{(m)})\right]$$

Proof.

$$\mathbb{E}[f(Y)] = \sum_{k=0}^{\infty} \mathbb{E}\left[f(Y) \,\middle|\, \sum Y_i^{(m)} = k\right] \cdot \Pr\left[\sum Y_i^{(m)} = k\right]$$

$$\geq \mathbb{E}\left[f(Y) \,\middle|\, \sum Y_i^{(m)} = m\right] \cdot \Pr\left[\sum_{i=1}^{n} Y_i^{(m)} = m\right]$$

$$= \mathbb{E}[f(X)] \cdot \Pr[\text{Pois}(m) = m]$$

$$= e^{-m} \frac{m^m}{m!} > \frac{1}{e\sqrt{m}} \quad \text{— Stirling}$$

IF IN ADDITION $\mathbb{E}[f(X_1^{(m)}, \dots)]$ is monotone in m

— THEN $\mathbb{E}[f(X_1^{(m)}, \dots)] \leq 2\,\mathbb{E}[f(Y_1^{(m)}, \dots)]$

for $\mathbb{E}[f(X)]$ increasing in m

... "same proof"

$$\mathbb{E}[f(Y)] \geq \sum_{k \geq m} \underbrace{\mathbb{E}\left[f(Y) \,\middle|\, \sum Y_i = k\right]}_{= \mathbb{E}[f(X^{(k)})]} \cdot \Pr\left[\sum Y_i = k\right] \geq \mathbb{E}[f(X^{(m)})] \cdot \underbrace{\Pr\left[\sum Y_i^{(m)} \geq m\right]}_{\geq 1/2}$$

$$\Pr[\text{Pois}(m) \geq m] \geq \frac{1}{2} \quad ?$$

X … exact case

Y — Poisson case

Corollary Any event that takes place with probability p in the Poisson case takes place with probability $\leq pe\sqrt{m}$ in the exact case.

OR $\leq 2p$ for monotone f

Proof f … indic. ftion of the event $\begin{cases} 1 & \text{— event occurs} \\ 0 & \end{cases}$

prob. in the Poisson case $= E[f(Y)] = p$

prob. in the exact case $= E[f(X)] \leq e\sqrt{m}\, E[f(Y)] = pe\sqrt{m}$ $\blacksquare$

Thm $\Pr\left[\max_{i\le n} X_i^{(n)} < \frac{\ln n}{\ln\ln n}\right] < \frac{1}{n}$ (n suff. large) $(m=n)$

Proof $M := \frac{\ln n}{\ln\ln n}$

Poisson case: $\Pr\left[Y_1^{(m)} \geq M\right] \geq \Pr\left[\mathrm{Pois}(1) = M\right] = \frac{1}{e\cdot M!}$ $\quad 1 = \frac{m}{n}$

$$\Pr\left[\max Y_i^{(n)} < M\right] \leq \left(1 - \frac{1}{eM!}\right)^n \leq e^{-\frac{n}{eM!}}$$

pick M s.t. $M! \leq \frac{n}{2e\ln n}$ $\Longrightarrow$ Prob. in Poisson $\leq \frac{1}{n^2}$

$\Longrightarrow$ Pr. in exact case $\leq \frac{e\sqrt{n}}{n^2} \leq \frac{1}{n}$

$$M! \leq e\sqrt{M}\left(\frac{M}{e}\right)^M \leq M\left(\frac{M}{e}\right)^M$$

$$\ln M! \leq M\ln M - M + \ln M$$

$$= \frac{\ln n}{\ln\ln n}(\ln\ln n - \ln\ln\ln n) - \frac{\ln n}{\ln\ln n} + (\ln\ln n - \ln\ln\ln n)$$

$$\leq \ln n - \frac{\ln n}{\ln\ln n}$$

$$\leq \ln n - \ln\ln n - \ln(2e)$$

$$-\frac{\ln n \ln\ln\ln n}{\ln\ln n} + \ln\ln n - \dots \leq 0$$

$\sqrt{\ln n} > \ln\ln n$

$\blacksquare$

? $\mathbb{E}[f(X_1^{(n)}, \dots)]$ monot. in n?

$$f(x_1, \dots, x_n) = 1 \Leftrightarrow \max\{x_1, \dots, x_n\} < M$$

Power of 2 choices

So, max load is $\Theta\left(\frac{\log n}{\ln\ln n}\right)$ whp

.... but average is just 1 ball per bin!

Different model: $d \geq 2$ bins chosen unif. at random

We put ball in the one that is less full

Thm n balls, n bins, rule

$$\Pr\left[\max_i \tilde{X}_i^{(n)} \geq \frac{\ln\ln n}{\ln d} + O(1)\right] = o\left(\frac{1}{n}\right)$$

Application: hashing

$f: \mathcal{U} \to \{0, 1, \ldots, n-1\}$ "random" hashing ftion

representation of a subset of $\mathcal{U}$:

linked lists $L_0, \ldots, L_{n-1}$

add x : add x to $L_{f(x)}$ ---- time $O(1)$

find x : linear search in $L_{f(x)}$ ---- time $O(|L_i|) + O(1)$

set of size m ---- m balls, n bins ---- $|L_i| = X_i^{(m)}$

$$\mathbb{E}[\text{time to find}] = \frac{m}{n} + O(1)$$

$(m=n)$ max. time to find .. $\Theta\left(\frac{\ln n}{\ln \ln n}\right)$ whp