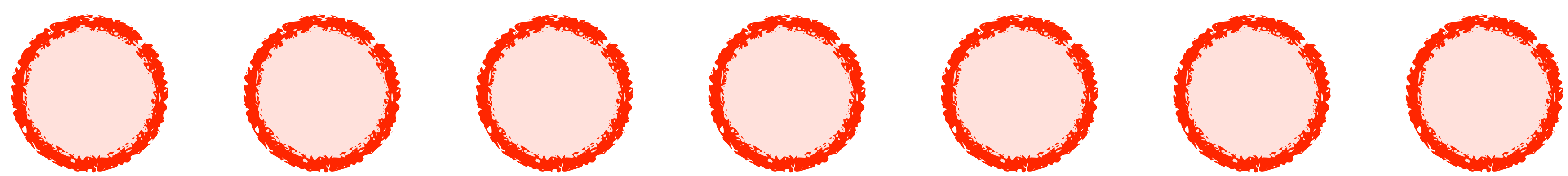
Advanced Algorithms

南京大学

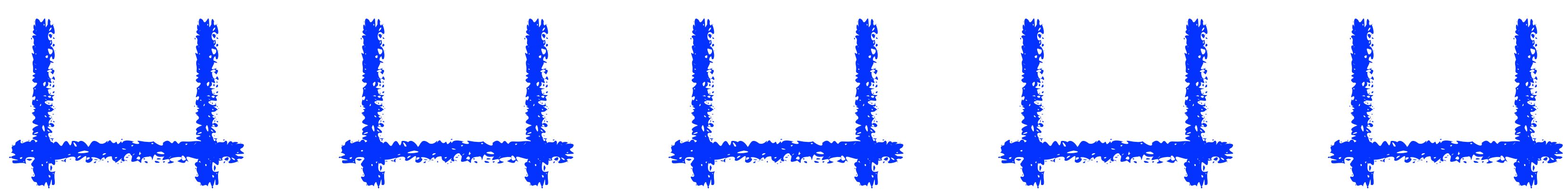
尹一通

Balls and Bins

m balls



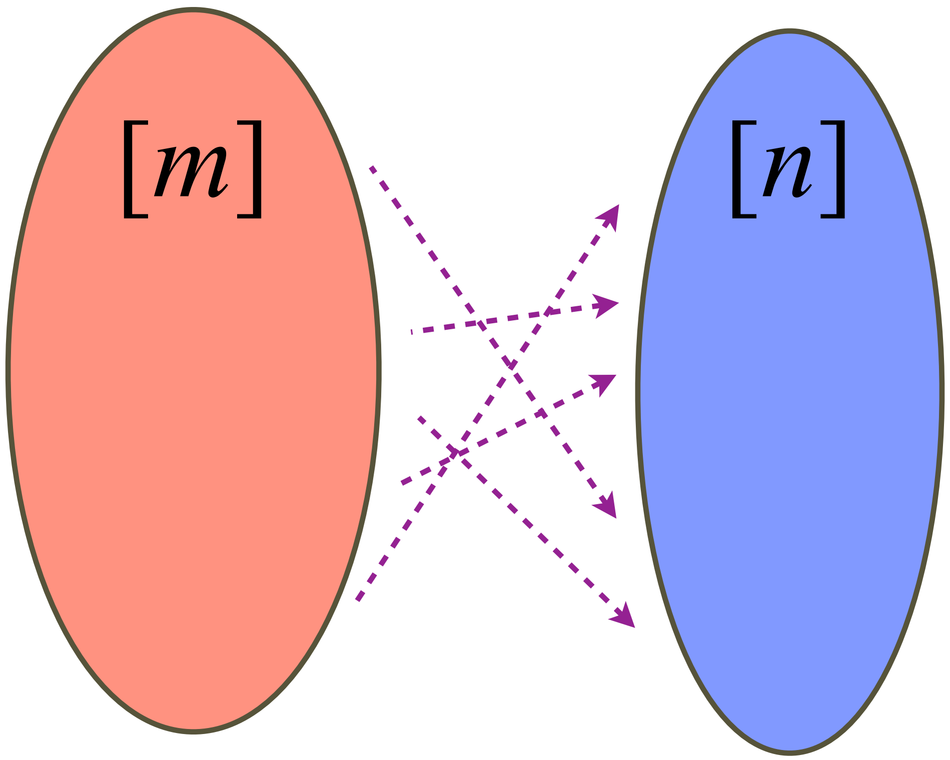
uniformly & independently



n bins

birthday problem, coupon collector problem, occupancy problem, ...

Random function



uniformly random function

balls-into-bins:

$$\Pr[\text{assignment}] = \underbrace{\frac{1}{n} \cdot \frac{1}{n} \cdots \frac{1}{n}}_{m} = \frac{1}{n^m}$$

random function:

$$\Pr[\text{assignment}] = \frac{1}{|[m] \to [n]|} = \frac{1}{n^m}$$

1-1	birthday problem
on-to	coupon collector
pre-images	occupancy problem

Birthday Paradox

Paradox:
(i) a statement that leads to a contradiction;
(ii) a situation which defies intuition.

birthday paradox:

Assumption: birthdays are uniformly & independently distributed.

In a class of m>57 students, with >99% probability, there are two students with the same birthday.

m-balls-into-n-bins:
$\mathcal{E}$: there is no bin with > 1 balls.

Birthday Paradox

m-balls-into-n-bins:
$\mathcal{E}$: there is no bin with > 1 balls.

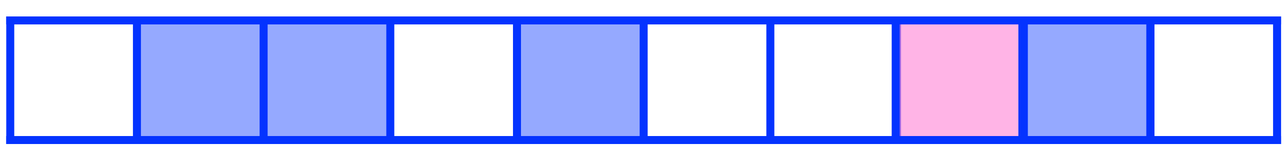
$$\Pr[\mathcal{E}] = \prod_{k=0}^{m-1}\left(1-\frac{k}{n}\right)$$

suppose balls are thrown one-by-one:

$$\Pr[\mathcal{E}] = \Pr[\text{no collision for all } m \text{ balls}]$$

$$= \prod_{k=0}^{m-1} \Pr[\text{no collision for the } (k+1)\text{th ball} \mid \text{no collision for the first } k \text{ balls}]$$

chain rule



Birthday Paradox

m-balls-into-n-bins:
$\mathcal{E}$: there is no bin with > 1 balls.

$$\Pr[\mathcal{E}] = \prod_{k=0}^{m-1}\left(1 - \frac{k}{n}\right)$$

Taylor's expansion: $e^{-k/n} \approx 1 - k/n$

$$\prod_{k=1}^{m-1}\left(1 - \frac{k}{n}\right) \approx \prod_{k=1}^{m-1} e^{-\frac{k}{n}}$$

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

$$= e^{-m(m-1)/2n}$$

$$\approx e^{-m^2/2n}$$

Birthday Paradox

m-balls-into-n-bins:
$\mathcal{E}$: there is no bin with > 1 balls.

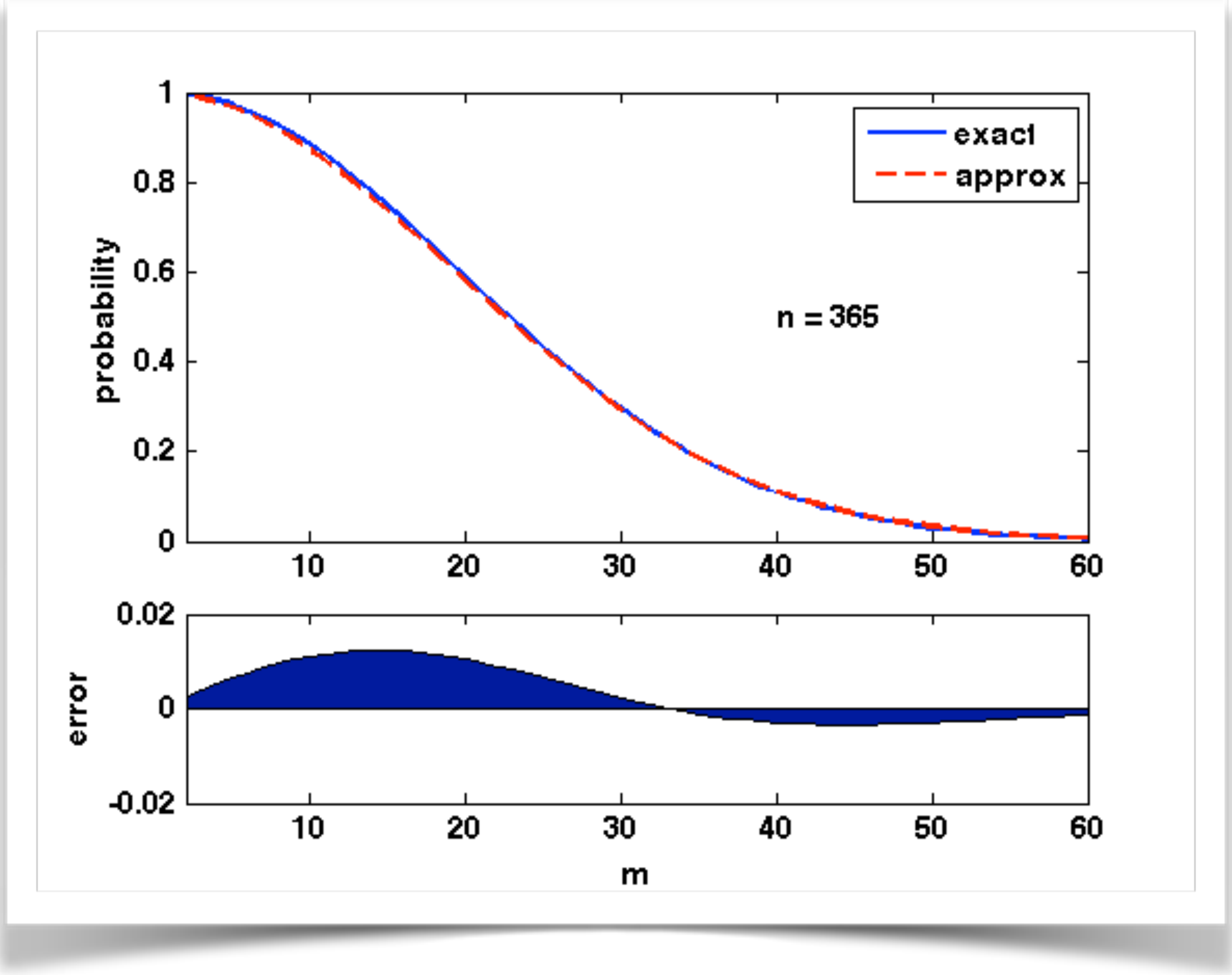
$$\Pr[\mathcal{E}] = \prod_{k=0}^{m-1}\left(1 - \frac{k}{n}\right)$$

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

$$\text{for } m = \sqrt{2n \ln \frac{1}{\epsilon}},$$

$$\Pr[\mathcal{E}] \approx \epsilon$$

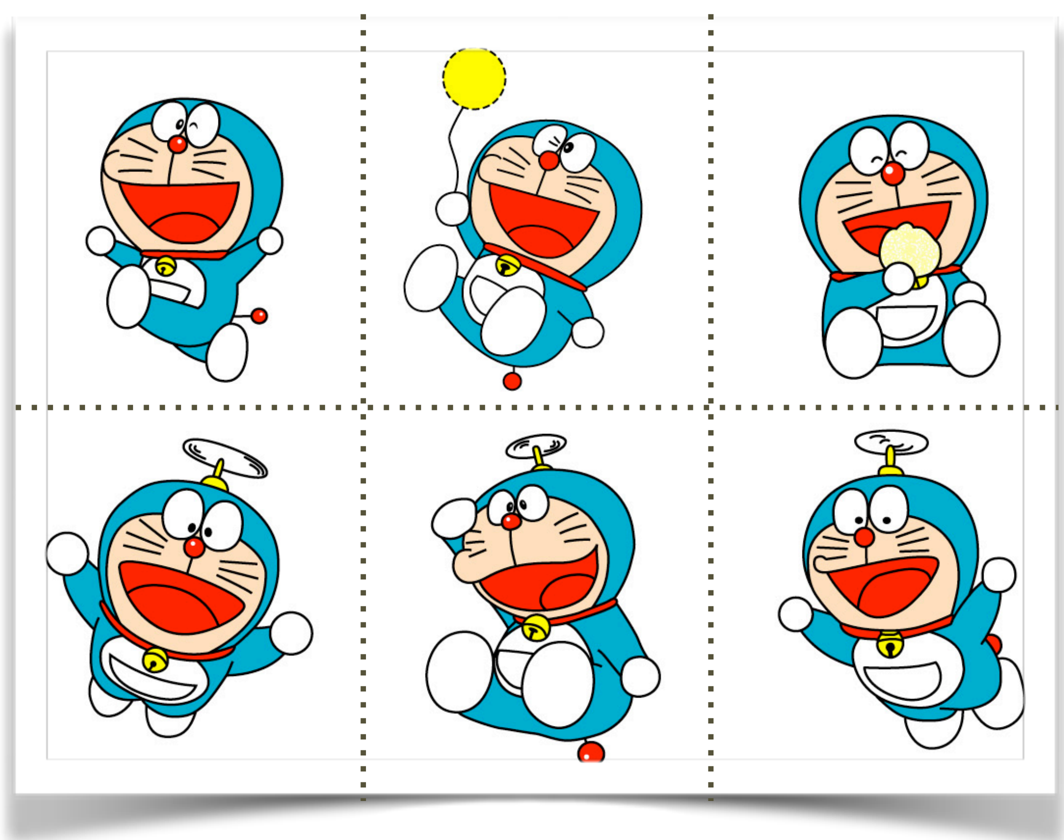
$m = \theta(\sqrt{n})$ for constant ϵ



Coupon Collector

(cover time)

coupons in cookie box



each box comes with a uniformly random coupon

number of boxes bought to collect all n coupons

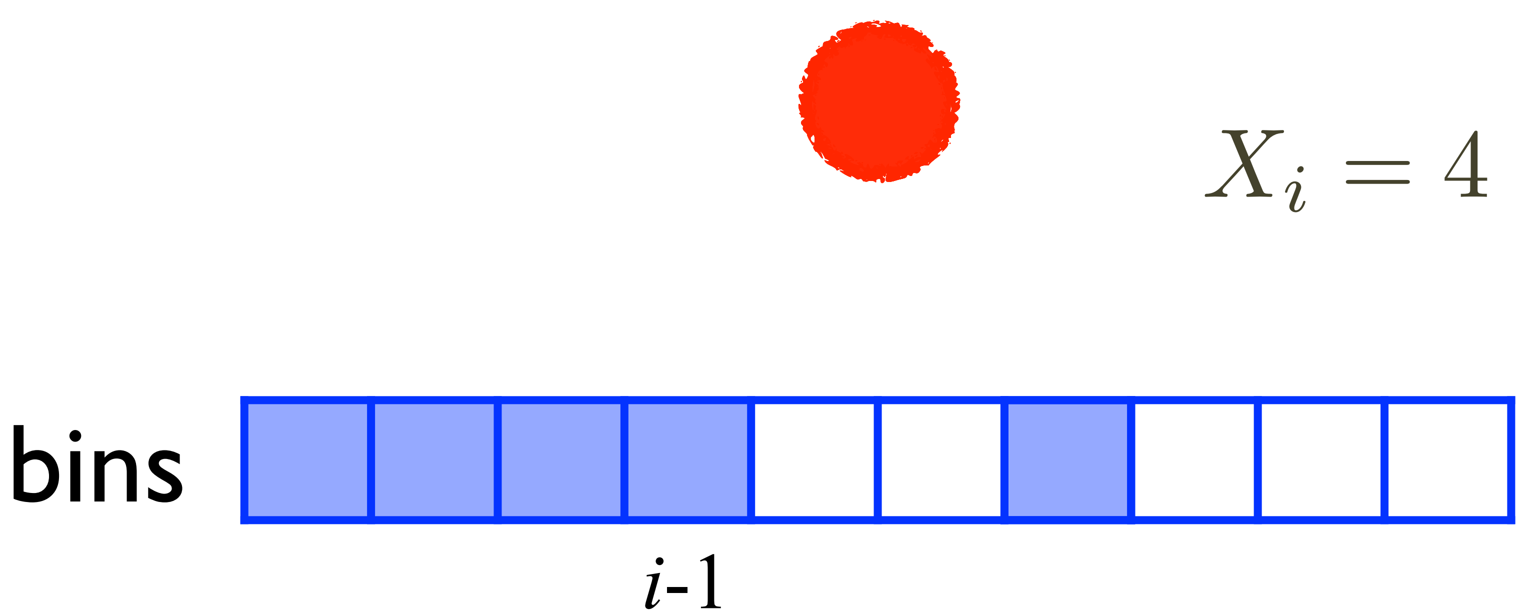
⇕

number of balls thrown to cover all n bins

Coupon Collector

X :	number of balls thrown to make all the n bins nonempty

$$X = \sum_{i=1}^{n} X_i$$



X_i is geometric!

with $p_i = 1 - \frac{i-1}{n}$

$$\mathbf{E}[X_i] = \frac{1}{p_i} = \frac{n}{n-i+1}$$

Coupon Collector

X :	number of balls thrown to make all the n bins nonempty
X_i :	number of balls thrown while there are exactly (i-1) nonempty bins

$$X = \sum_{i=1}^{n} X_i$$

$$\mathbf{E}[X_i] = \frac{1}{p_i} = \frac{n}{n-i+1}$$

$$\mathbf{E}[X] = \sum_{i=1}^{n} \mathbf{E}[X_i]$$

linearity of expectations

$$= \sum_{i=1}^{n} \frac{n}{n-i+1}$$

Expected $n \ln n + O(n)$ balls!

$$= n \sum_{i=1}^{n} \frac{1}{i}$$

$$= nH(n)$$

Harmonic number

Coupon Collector

X : number of balls thrown to make all the n bins nonempty

Theorem: For $c > 0$

$$\Pr[X \geq n \ln n + cn] < e^{-c}$$

Proof: For one bin, it misses all balls with probability

$$\left(1 - \frac{1}{n}\right)^{n \ln n + cn} = \left(1 - \frac{1}{n}\right)^{n(\ln n + c)}$$

$$< e^{-(\ln n + c)}$$

$$= \frac{1}{ne^c}$$

Coupon Collector

X : number of balls thrown to make all the n bins nonempty

Theorem: For $c > 0$
$$\Pr[X \geq n \ln n + cn] < e^{-c}$$

Proof: For one bin, it misses all balls with probability

$$< \frac{1}{ne^c}$$

For all n bins, union bound!

$$\Pr[\exists \text{ a bin misses all balls}] \leq n \cdot \Pr[\text{one bin misses all balls}]$$
$$< n \cdot \frac{1}{ne^c} = e^{-c}$$

Coupon Collector

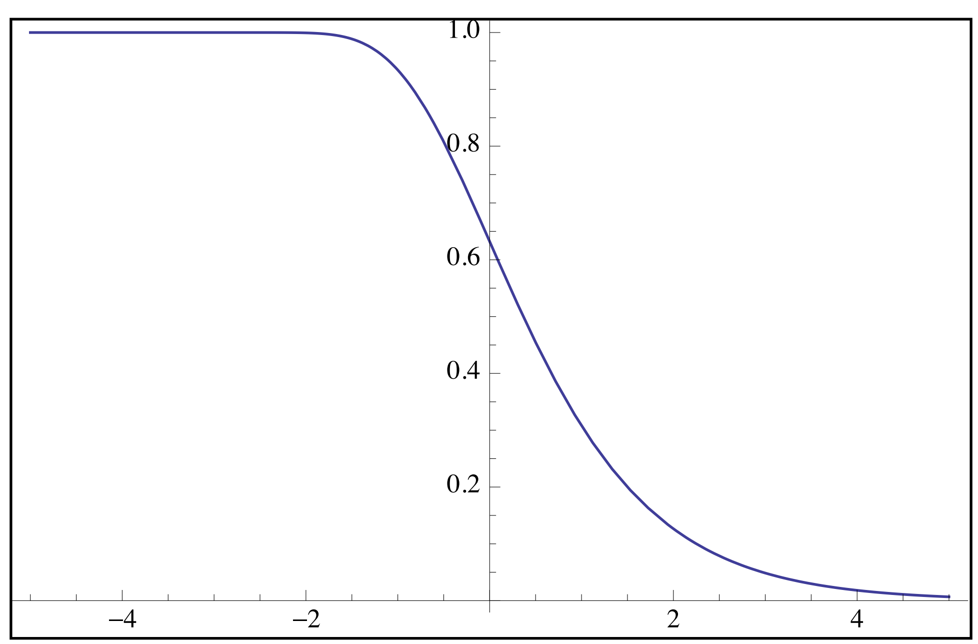
X : number of balls thrown to make all the n bins nonempty

Theorem: For $c > 0$

$$\Pr[X \geq n \ln n + cn] < e^{-c}$$

a sharp threshold:

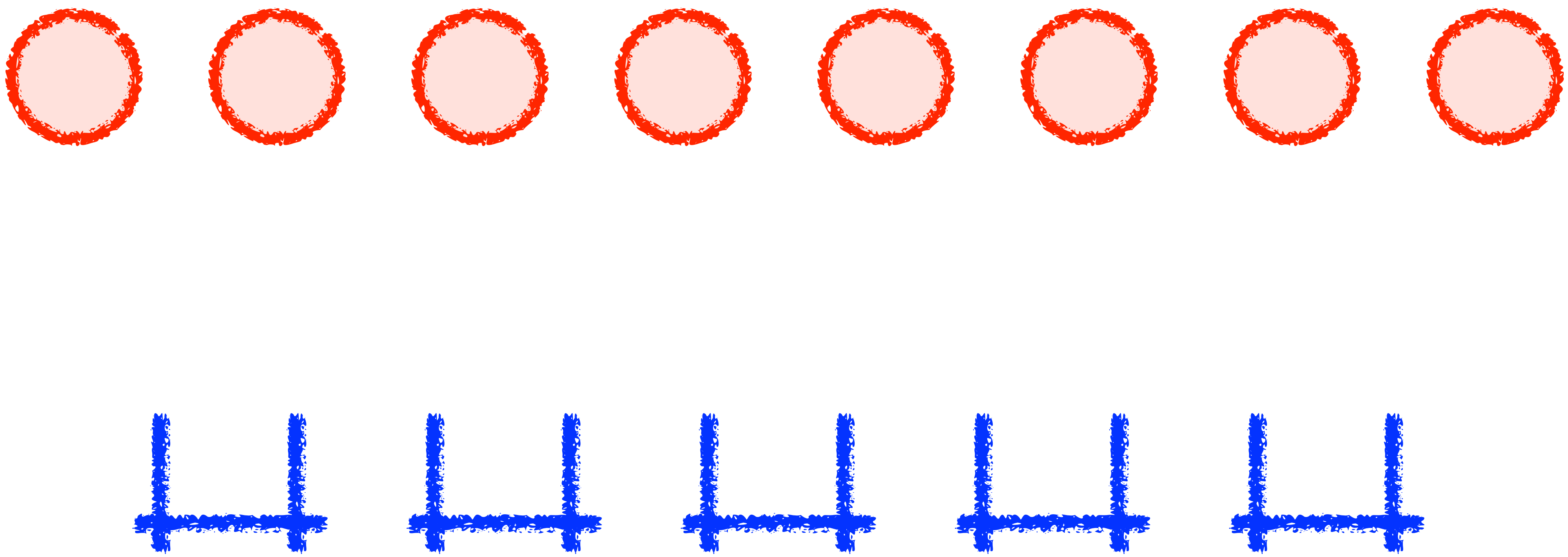
$$\lim_{n\to\infty} \Pr[X \geq n \ln n + cn] = 1 - \mathrm{e}^{-\mathrm{e}^{-c}}$$



Occupancy Problem

(load balancing)

m balls



n bins

$X_1, X_2, \ldots, X_n$

loads of bins

maximum load?

Occupancy Problem

m balls

n bins

$X_1, X_2, \ldots, X_n$

loads of bins

$$\max_{1 \le i \le n} \mathbf{E}[X_i] = ?$$

$$\sum_{i=1}^{n} X_i = m \quad \Rightarrow \quad \sum_{i=1}^{n} \mathbf{E}[X_i] = \mathbf{E}\left[\sum_{i=1}^{n} X_i\right] = m$$

Symmetry! $\Rightarrow$ All $\mathbf{E}[X_i]$ are equal.

$$\max_{1 \le i \le n} \mathbf{E}[X_i] = \frac{m}{n}$$

Occupancy Problem

$$\max_{1 \le i \le n} \mathbf{E}[X_i] = \frac{m}{n}$$

Theorem:

If $m = n$, the max load is $O\left(\frac{\ln n}{\ln \ln n}\right)$ with high probability.

w.h.p.: $\Pr = 1 - O(\frac{1}{n^c})$ or $\Pr = 1 - o(1)$

n balls into n bins:

$$\Pr[\text{ bin-1 has } \geq t \text{ balls }]$$

$$\leq \Pr[\exists \text{ a set } S \text{ of } t \text{ balls s.t. all balls in } S \text{ are in bin-1 }]$$

$$\binom{n}{t} \qquad\qquad \frac{1}{n^t}$$

union bound

$$\leq \sum_{\text{set } S \text{ of } t \text{ balls}} \Pr[\text{all balls in } S \text{ are in bin-1}]$$

$$\leq \frac{1}{n^t}\binom{n}{t} = \frac{n(n-1)(n-2)\cdots(n-t+1)}{t!n^t} \leq \frac{1}{t!} \leq \left(\frac{e}{t}\right)^t$$

Stirling approximation

n balls into n bins:

$$\Pr[\text{ bin-1 has } \geq t \text{ balls }] \leq \left(\frac{\mathrm{e}}{t}\right)^t$$

$$\Pr[\text{ max load } \geq t] = \Pr[\exists \text{ bin-}i \text{ has } \geq t \text{ balls}]$$

$$\leq n \Pr[\text{ bin-1 has } \geq t \text{ balls }]$$ union bound

$$\leq n \left(\frac{\mathrm{e}}{t}\right)^t$$ **choose** $t = \frac{3 \ln n}{\ln \ln n}$

$$= n \left(\frac{e \ln \ln n}{3 \ln n}\right)^{3 \ln n / \ln \ln n} < n \left(\frac{\ln \ln n}{\ln n}\right)^{3 \ln n / \ln \ln n}$$

$$= n e^{3(\ln \ln \ln n - \ln \ln n) \ln n / \ln \ln n}$$

$$\leq n e^{-3 \ln n + 3(\ln \ln \ln n)(\ln n) / \ln \ln n}$$

$$\leq n e^{-2 \ln n} = \frac{1}{n}$$

Occupancy Problem

Theorem: m balls into n bins:

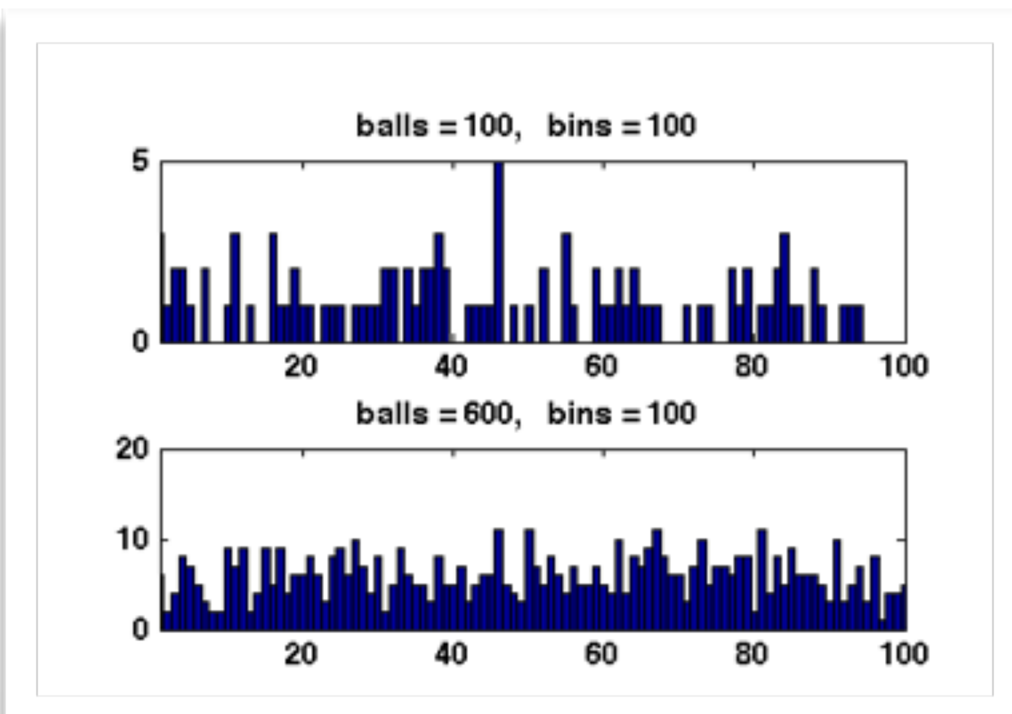
If $m = n$, the max load is $O\left(\frac{\ln n}{\ln \ln n}\right)$ with high probability.

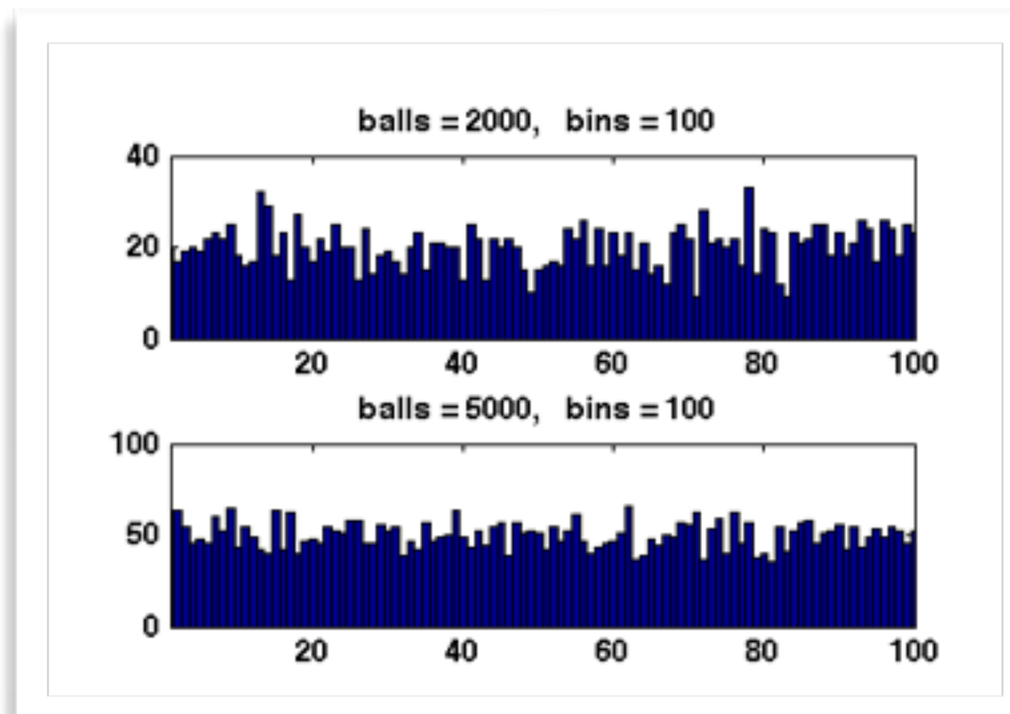
Occupancy Problem

Theorem: m balls into n bins:

If $m = n$, the max load is $O\left(\frac{\ln n}{\ln \ln n}\right)$ with high probability.

When $m = \Omega(n \log n)$, the max load is $O(\frac{m}{n})$ with high probability





Balls-into-bins model

throw *m* balls into *n* bins
uniformly and independently

uniform random function
$$f : [m] \to [n]$$

1-1	birthday problem
on-to	coupon collector
pre-images	occupancy problem

- The threshold for being 1-1 is $m = \Theta(\sqrt{n})$.
- The threshold for being on-to is $m = n \ln n + O(n)$.
- The maximum load is

$$\begin{cases} O(\frac{\ln n}{\ln \ln n}) & \text{for } m = \Theta(n), \\ O(\frac{m}{n}) & \text{for } m = \Omega(n \ln n). \end{cases}$$