From prev lecture: Coupon collector problem, Markov and Chebyshev's inequality.

Recalling Chebyshev's inequality:

Let X be a r.v with expectation $\mathbb{E}[X]$ and standard deviation σ_X. Then for any $t \in \mathbb{R}_{\geq 0}$

$$\Pr[|X - \mathbb{E}[X]| \geq t\sigma_X] \leq \frac{1}{t^2}$$

$\Pr[X \geq 2\mathbb{E}[X]]$

Let $v = t\sigma_X$. Then $t^2 = \frac{v^2}{\sigma_X^2}$ and $\sigma_X^2 = \text{Var}[X]$.

$$\Rightarrow \Pr[|X - \mathbb{E}[X]| \geq v] \leq \frac{\text{Var}[X]}{v^2}.$$

$\Pr[|X - \mathbb{E}[X]| \geq \mathbb{E}[X]]$

$\leq$

We wanted to find out prob that $X \geq 2\mathbb{E}[X]$.
Recall that this was at most $\frac{1}{2}$ from Markov ineq.

By setting $v = \mathbb{E}[X]$, we get that

$$\Pr[|X - \mathbb{E}[X]| \geq \mathbb{E}[X]] \leq \frac{\text{Var}[X]}{(\mathbb{E}[X])^2}.$$

$|X - \mu| \geq t$

$\downarrow$

$X \geq \mu + t$ and $X \leq \mu - t$.

For coupon collector problem, $\text{Var}[X] = O(n^2)$ and $\mathbb{E}[X] = \Theta(n \ln n)$

Thus, we get that $\Pr[|X - \mathbb{E}[X]| \geq \mathbb{E}[X]] \leq O\left(\frac{1}{\ln^2 n}\right)$.

(In comparison, Markov gives only $1/2$.)

Remark: $\mathrm{Var}[X+Y] = \mathrm{Var}[X] + \mathrm{Var}[Y] + 2\,\mathrm{CoVar}[X,Y]$.

$\mathrm{CoVar}[X,Y] = \mathbb{E}\left[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])\right]$.

In coupon collector problem, $X = \sum X_i$ and X_i's were independent r.vs and thus $\mathrm{Var}[X] = \sum_{i=1}^{n} \mathrm{Var}[X_i]$.

Other arguments to obtain tail probabilities:

Probability of not obtaining i^{th} coupon after s steps (trials) for some $s > n\ln n + \Theta(n)$:

$$\left(1-\frac{1}{n}\right)^{s} = \left(1-\frac{1}{n}\right)^{n\cdot\frac{s}{n}} \sim e^{-\frac{s}{n}}.$$

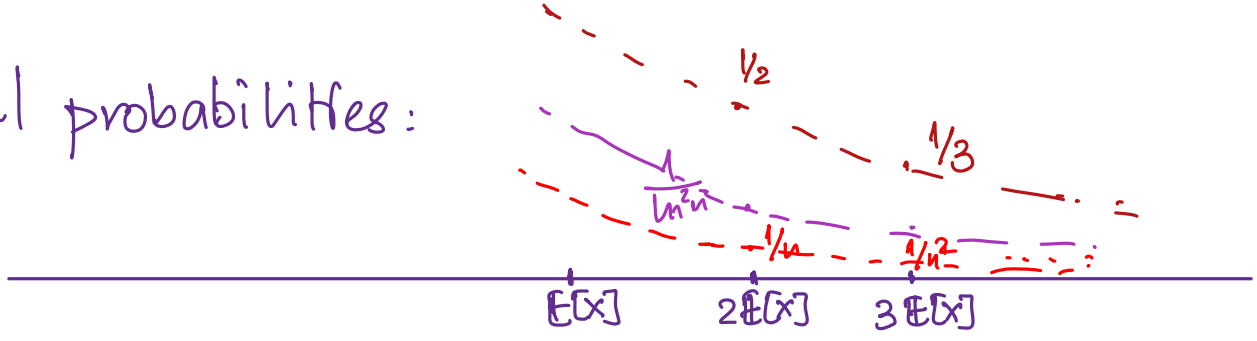
Probability that there exists some coupon that is not yet picked up after s trials, by union bound is $\leq$ sum of probabilities that each coupon is not picked after s trials.

$$\hookrightarrow \approx \sum_{i=1}^{n} e^{-\frac{s}{n}} = \frac{n}{e^{s/n}}.$$

w.p $\geq 1-\frac{1}{n}$; all coupons are collected in $s = 2n\ln n$ trials.

Set $s = 2n\ln n$. Then the tail prob is at most $\frac{1}{n}$.

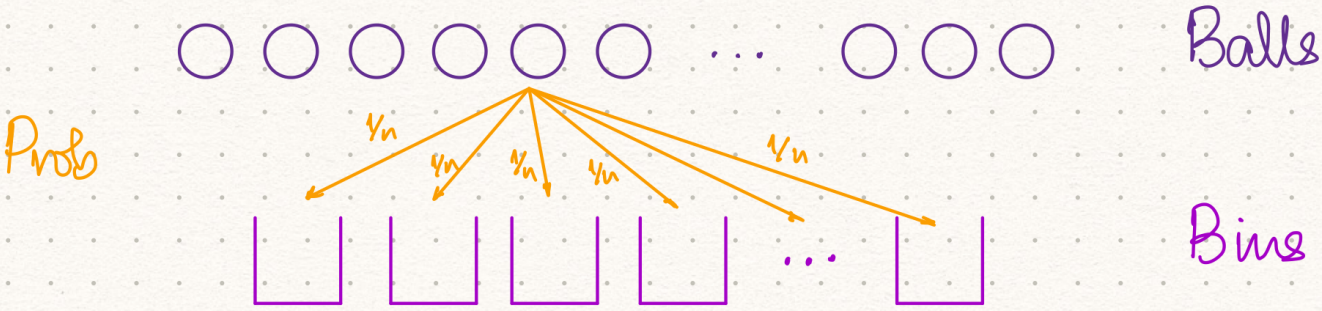
Tail probabilities:



2.2 Balls and Bins.

We are given m "indistinguishable" balls and n bins. Each ball is placed in a bin that is chosen uniformly at random.

Question: How are these balls distributed amongst the bins.



Question 1: What is the expected no. of balls in each bin?

Question 2: What is the prob that no bin gets more than $k-1$ balls?

Let E_j be the event that bin j has more than k balls.

$$\Pr[E_j] = \sum_{\ell=k}^{m} \Pr[\text{bin } j \text{ gets exactly } \ell \text{ balls}]$$

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

$$\left(\frac{a}{b}\right)^b \le \binom{a}{b} \le \left(\frac{ea}{b}\right)^b$$

$$\le \sum_{\ell=k}^{m} \binom{m}{\ell} \cdot \left(\frac{1}{n}\right)^{\ell} \le \sum_{\ell=k}^{m} \left(\frac{em}{\ell}\right)^{\ell} \cdot \left(\frac{1}{n}\right)^{\ell} = \sum_{\ell=k}^{m} \left(\frac{em}{\ell n}\right)^{\ell}.$$

Say $m=n$. Then $\Pr[E_j] \le \sum_{l=k}^{n} \left(\frac{e}{l}\right)^l \le \sum_{l=k}^{n} \left(\frac{e}{k}\right)^l$

$= \left(\frac{e}{k}\right)^k \left[1 + \left(\frac{e}{k}\right) + \cdots + \left(\frac{e}{k}\right)^{n-k}\right]$

$\le \left(\frac{e}{k}\right)^k \cdot \frac{1}{1-\left(\frac{e}{k}\right)}$

	b_1	b_2
trial 1	B_1	B_2 ←
trial 2	B_1	B_1
trial 3	B_2	B_2
trial 4	B_2	B_1 ←

2 bins

Let $k = \left\lceil \frac{3 \ln n}{\ln \ln n} \right\rceil$. Then $\Pr[E_j] \le \frac{1}{n^2}$.

$$\left(\frac{e \ln\ln(n)}{3\ln(n)}\right)^{\frac{3\ln n}{\ln\ln n}} \cdot \frac{3\ln n}{3\ln n - \ln\ln(n)}$$

$$\le \left(\frac{\ln\ln(n)}{\ln(n)}\right)^{\frac{3\ln(n)}{\ln\ln(n)}} \approx e^{(\ln\ln\ln n - \ln\ln n)\frac{3\ln n}{\ln\ln n}}$$

$$= e^{-\left(1 - \frac{\ln\ln\ln(n)}{\ln\ln(n)}\right)\cdot 3\ln(n)}$$

$$\le \frac{1}{n^2}.$$

For sufficiently large n, $1 - \frac{\ln\ln\ln(n)}{\ln\ln(n)} \ge \frac{2}{3}$.

$\mathbb{E}[X_j] = \Pr[X_j = 1] = \Pr[E_j] \le \frac{1}{n^2}$.

X_j be a 0-1 r.v s.t

$$X_j = \begin{cases} 1 & \text{if bin } j \text{ contains more than } k \text{ balls.} \\ 0 & \text{o/w.} \end{cases}$$

Prob that no bin contains more than $k-1$ balls

$$= 1 - \Pr[\exists \text{ a bin with at least } k \text{ balls}]$$

$$= 1 - \Pr\left[\sum_{i=1}^{n} x_i \geq 1\right]$$

$$\geq 1 - \left(\frac{\mathbb{E}\left[\sum_{i=1}^{n} x_i\right]}{1}\right)$$

← Application of Markov ineq.

$$= 1 - \sum_{i=1}^{n} \mathbb{E}[x_i]$$

$$\geq 1 - n \cdot \frac{1}{n^2}$$

$$= 1 - \frac{1}{n}.$$

Theorem: With probability of at least $1 - \frac{1}{n}$, no bin has more than $\frac{3\ln(n)}{\ln\ln(n)}$ balls in it.

(Max load is k w.p $1 - o(1)$).

Question: Does there exist a bin with $\frac{\ln n}{3 \ln\ln n}$ balls?

Recall that by Chebyshev's inequality,

$$\Pr[|x - \mathbb{E}[x]| \geq t] \leq \frac{\mathrm{Var}[x]}{t^2}.$$

Remark: Prob that no bin contains more than $k-1$ balls is at most $\Pr[|x - \mathbb{E}[x]| \geq \mathbb{E}[x]]$ and by Chebyshev's this qty is at most $\frac{\mathrm{Var}[x]}{(\mathbb{E}[x])^2}$ where $X = \sum X_i$.

Thus, we need an upper bound on $\mathrm{Var}[x]$ and lower bound on $\mathbb{E}[x]$.

$\mathbb{E}[x_i]$ = prob that bin i contains at least k balls

$\geq$ prob that bin i contains exactly k balls.

$$= \binom{n}{k}\left(\frac{1}{n}\right)^k\left(1-\frac{1}{n}\right)^k$$

Set $k = \frac{\ln(n)}{3\ln\ln(n)}$.

$$\geq \left(\frac{n}{k}\right)^k \cdot \frac{1}{n^k} \cdot e^{-\frac{k}{n}}.$$

$$\geq \frac{1}{e}\cdot\left(\frac{1}{k}\right)^k.$$

$$\geq \frac{1}{e}\left(\frac{3\ln\ln(n)}{\ln n}\right)^{\frac{\ln n}{3\ln\ln n}}$$

$$= \frac{1}{e}\cdot e^{\frac{\ln n}{3\ln\ln n}\left(\ln(3\ln\ln(n)) - \ln(\ln n)\right)}$$

$$= \frac{1}{e}\cdot e^{-\frac{\ln n}{3}\left(\frac{\ln\ln n - \ln 3 - \ln\ln\ln(n)}{\ln\ln n}\right)}$$

$$\approx \frac{1}{e}\cdot\frac{1}{n^{1/3}}$$

$$\mathbb{E}[X] = \sum_{i=1}^{n}\mathbb{E}[X_i] \geq \frac{n^{2/3}}{e}.$$

Claim: X_i and X_j for $i \neq j$ are negatively correlated.

Thus, $\mathrm{Cov}(X_i, X_j) \leq 0$.

Thus, $\mathrm{Var}[X] = \mathrm{Var}\left[\sum_{i=1}^{n} X_i\right] \leq \sum_{i=1}^{n}\mathrm{Var}[X_i]$.

$\mathrm{Var}[X_i] = \mathbb{E}[(X_i - \mathbb{E}[X_i])^2] \leq 1$ as $X_i \in \{0,1\}$ } Fact

Thus, $\mathrm{Var}[X] \leq n$.

Thus, $\Pr[X=0]$

$\leq \Pr[|X - \mathbb{E}[X]| \geq \mathbb{E}[X]]$

$\leq \dfrac{\mathrm{Var}[X]}{(\mathbb{E}[X])^2}$

$\leq \dfrac{n}{\left(\frac{n^{4/3}}{e^2}\right)} = \dfrac{e^2}{n^{1/3}}$

X = no. of bins with at least k balls.

$\Pr[X=0]$ is a bad event when you want to count $\exists$ of bin w/ $\geq k$ balls.

Theorem: With a prob of at least $1 - \frac{e^2}{n^{1/3}}$ $\exists$ a bin with $\frac{\ln n}{3 \ln \ln n}$ balls.

(Max load is at least k w.p $1 - o(1)$).

Question: Markov vs Chebyshev (when to use them?)
(Can you get $1 - \frac{e^2}{n^{1/3}}$ kind of a bound from Markov?)

Let $Y = n - X$ be a r.v.

$\Pr[X=0] = \Pr[Y=n]$

$= \Pr[Y \geq n]$

$\leq \dfrac{\mathbb{E}[Y]}{n}$

$= \dfrac{\mathbb{E}[n-X]}{n}$

$\geq \dfrac{n - \frac{n^{2/3}}{e}}{n}$

$= 1 - \dfrac{1}{e \cdot n^{1/3}}$.

Success prob $= 1 - \left(1 - \frac{1}{e \cdot n^{1/3}}\right) \sim \frac{1}{e \cdot n^{1/3}}$

$\left(1 - \frac{e^2}{n^{1/3}}\right)$

- Markov states that a non-negative r.v takes values much larger than its expectation with small prob.
- Chebyshev states that a r.v with finite and bounded variance is concentrated around its expectation. Smaller the var $\Rightarrow$ better concentration.