Seoul
National
University

Bounds

Name: Chong-kwon Kim

SCONE
Lab.

Example – Coupon Collection

- Let X: Time to collect all n types of coupon
- $X = X_1 + X_2 + \cdots + X_n$ (Xi is time to collect i-th coupon types after (i-1) coupon types are collected)

- Xi: Geometric r. v. with pi = (1 – (i-1)/n)

  ➔ $E[X] = n \cdot H_n$

  ➔ $Var[X] = \sum_{i=1}^{n} Var[Xi]$

  $\leq \sum_{i=1}^{n} (n/(n-i+1))^2$

  $\leq \frac{\pi^2 \cdot n^2}{6}$

E[Xi] = (n-i+1)/n

$Var[Xi] = \frac{(1-p_i)}{p_i^2} \leq \frac{1}{p_i^2}$

$\sum_{i=1}^{n} (1/i)^2 = \frac{\pi^2}{6}$

- Find Markov and Chebyshev bounds of $Pr(X \geq 2n \cdot H_n)$

Example – Coin Flips Revisited

- We proved that, for $0 < \delta \leq 1$, $\Pr(X \geq (1+\delta)\mu) \leq e^{-\frac{\mu \cdot \delta^2}{3}}$

X should be sum of Poisson trials

- Also, it can be shown that $\Pr(X \leq (1-\delta)\mu) \leq e^{-\frac{\mu \cdot \delta^2}{3}}$

➔ $\Pr(|X-\mu| \geq \delta \cdot \mu) \leq 2e^{-\frac{\mu \cdot \delta^2}{3}}$

Refer to Theorem 4.5 & Corollary 4.6

- X: # heads in n coin flip
- Find bounds of $\Pr(|X - n/2| \geq n/4)$
  - Markov: $\Pr((X - n/2) \geq n/4)$ =
  - Chebyshev: $\Pr(|X - n/2| \geq n/4)$ =
  - Chernoff: $\Pr(|X - n/2| \geq n/4)$ =

Selection Problem

- Problem: Given an input of N distinct numbers, find i-th largest number
- Median: $\lceil N/2 \rceil$ - th or $\lceil (N+1)/2 \rceil$ - th largest number

- Complexity of find minimum (or maximum) number
  ➔ Θ(N)

- What is the complexity of finding the median?
  - Obviously, we can do in Θ(N lnN)

- Any selection algorithm with Θ(N)?

Randomized Selection

- Similar to randomized QuickSort

- Pick a pivot number randomly
- Partition the input into two subsets,S1 and S2, such that all in S1 are smaller than the pivot and all in S2 are larger than the pivot
- Pick S1 or S2 and repeat the procedure recursively

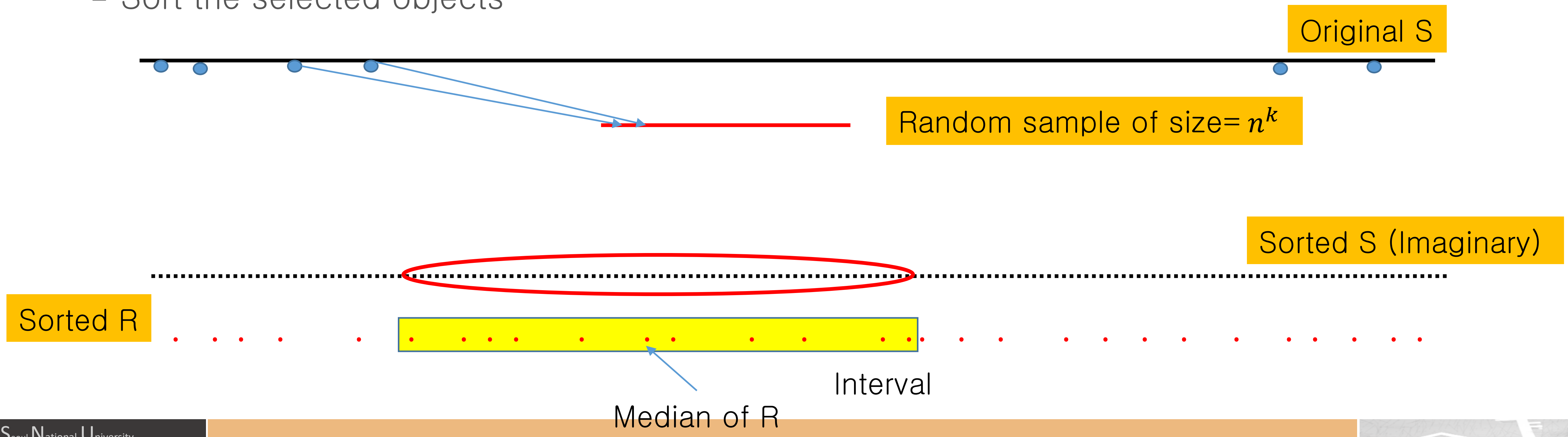
→ Θ(N)?

- Let T(N): # comparison to find the median
  - Then $T(N) \le 1/N \cdot (\sum_{k=1}^{n-1} T(\max(k, N-k)))$
  - T(N) = O(N)

Refer to CLRS

Randomized Median Algorithm

- Sketch of the algorithm
  - Given an original set S (size: n objects)
  - Generate a random sample (say R) of a properly small size, say $\sqrt{n}$, or n^k (k < 1)
    - Sort R (Complexity = O($n^k \cdot \log n^k$)
    - Fix an short interval (say I) that contains the median of R
  - Now, collects the objects that belong to the interval (Complexity??)
  - Sort the selected objects



Example

- S = {17, 7, 14, 6, 1,19, 3, 4, 7, 11, 18, 12, 21, 9, 5, 10, 2, 19, 8, 13, 16}
- Let R1 = {17, 7, 14, 6, 1,19, 3}, R2 = {17, 14, 19, 7, 18, 12, 21}

- Sorted S = {1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21}
- Sorted R1 = {1, $\boxed{3,\ 6,\ 7,\ 14,\ 17,}$ 19}
- Sorted R2 = { 7, $\boxed{12,\ 14,\ 17,\ 18,\ 19,}$ 21}

Randomized Median Algorithm

Input: A set S of n elements

Output: Median (m) of S

1. Construct a multi-set R of $\lceil n^{3/4} \rceil$ elements from S, each element chosen independently and uniformly at random with replacement
2. Sort R
3. Let *d* and *u* be the $\lfloor \frac{1}{2} n^{3/4} - \sqrt{n} \rfloor$ and $\lceil \frac{1}{2} n^{3/4} + \sqrt{n} \rceil$ –th elements, respectively, in sorted R
4. Compare every element in S to *d* and *u*. Construct a set C with elements in [d, u] and count l_d and l_u, the number of elements smaller than d and greater than u, respectively
5. If l_d > n/2 or l_u > n/2 ➔ FAIL
6. If |C| $\leq 4\, n^{3/4}$, then sort C,

   OW FAIL

M is ($\lfloor \frac{n}{2} \rfloor - l_d$+1) –th element in sorted C

Randomized Median Algorithm

- With high probability
  - Probability at least $1-O(1/n^c)$ for some c > 0

    m is between d and u

    |C| is not greater than $4n^{3/4}$

Condition of step 5 SUCCESS

Condition of step 6 SUCCESS

- Easy to prove that
  - If the algorithm does not FAIL, then it finds the median of S

- Need to prove
  1. Randomized median algorithm terminates in linear time O(N)
  2. SUCCESS with high probability

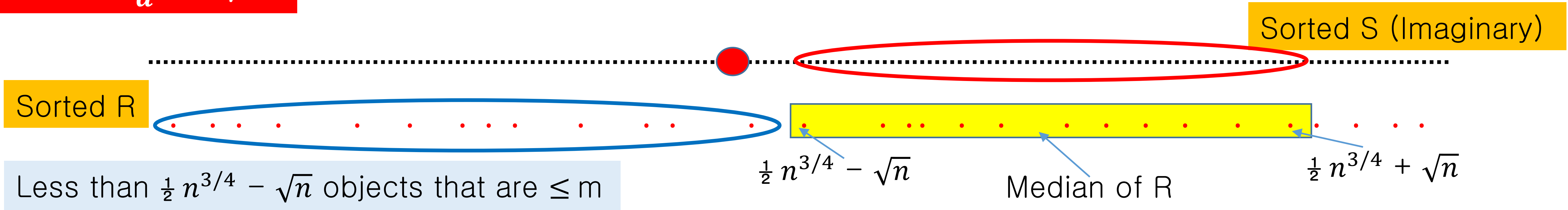
- The algorithm FAILs if any one of following events occur
  - E1: $Y1 = |\{r \in R \mid r \leq m\}| < \frac{1}{2}n^{3/4} - \sqrt{n}$
  - E2: $Y2 = |\{r \in R \mid r \geq m\}| < \frac{1}{2}n^{3/4} - \sqrt{n}$
  - E3: $|C| > 4n^{3/4}$

l_d > n/2 ➔ d is larger than m

➔ Less than $(\frac{1}{2}n^{3/4} - \sqrt{n})$ elements in R are smaller than or equal to m

Randomized Median Algorithm

Case: l_d > n/2



Randomized Median Algorithm

- Lemma: Pr(E1) ≤ (1/4) · $n^{-1/4}$

Probability at least 1−O(1/ n^c) for some c > 0

- Proof
  - Consider random sampling of i-th element and let Xi be a Bernoulli random variable such that
  - Xi = $\begin{cases} 1, & \text{if the sample } \le m \\ 0, & o.w \end{cases}$

    Pr(Xi=1) = $\frac{(n-1)/2+1}{n}$ = ½ + 1/2n
  - Define Binomial random variable Y1 = $\sum_{i=1}^{n^{3/4}} X_i$

    ➔ B(n, p) where n= $n^{3/4}$ and p = ½ + 1/2n

E[Y1}= ?
Var[Y1] = ?

  - Event E1 is equivalent to Y1 = $\sum_{i=1}^{n^{3/4}} X_i < \frac{1}{2}\ n^{3/4} - \sqrt{n}$

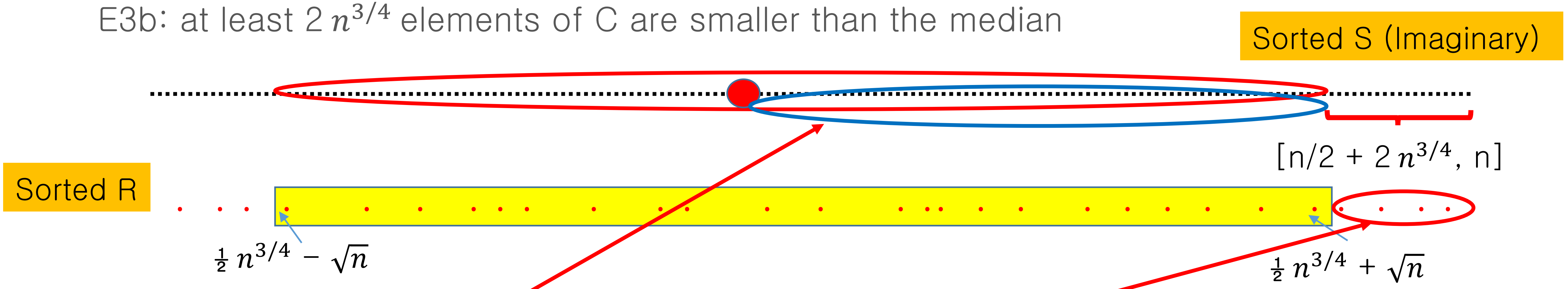
    Pr(Y1) = Pr(Y1 < ½ $n^{3/4} - \sqrt{n}$)

    ≤ Pr(|Y1 − E[Y1]| > $\sqrt{n}$)

    ≤ $\frac{Var[Y1]}{n}$

Randomized Median Algorithm

- Now prove Pr(E3=|C| > 4 $n^{3/4}$) ≤ (1/2) · $n^{-1/4}$
- Proof
  - Note that if E3 occur, then at least one of following two events occurs
    E3a: at least 2 $n^{3/4}$ elements of C are greater than the median
    E3b: at least 2 $n^{3/4}$ elements of C are smaller than the median



  - Focus on E3a
  - There are at least 2 $n^{3/4}$ elements in C that are greater than the median
    ➔ order of u in S is at least n/2 + 2 $n^{3/4}$
    ➔ R has at least $(1/2) \cdot n^{3/4} - \sqrt{n}$ elements in [n/2 + 2 $n^{3/4}$, n]

Randomized Median A)lgorithm

SCONE Lab.

- Again define Bernoulli r. v. Xi such that
- Xi = $\begin{cases} 1, & if\ the\ sample\ is\ in\ [\mathrm{n}/2 + 2\, n^{3/4}, \mathrm{n}] \\ 0, & o.w \end{cases}$

- Let Y3a = $\sum_{i=1}^{n^{3/4}} X_i$
- Pr(E3a) = Pr(Y3a ≥ $(1/2) \cdot n^{3/4} - \sqrt{n}$

  ≤ Pr(|Y3a – E[Y3a]| ≥ $\sqrt{n}$)

  ≤ $\frac{Var[X]}{n} < \frac{1}{4}\, n^{-1/4}$

$$\Pr(X_i = 1) = \frac{n - \frac{n}{2} - 2n^{3/4}}{\mathrm{n}} = \frac{1}{2} - 2n^{-1/4}$$

$$\mathrm{E[Y3a]} = \frac{1}{2} n^{3/4} - 2\sqrt{n}$$

$$Var[Y3a] = n^{3/4}\left(\frac{1}{2} - 2n^{-1/4}\right)\left(\frac{1}{2} + 2n^{-1/4}\right)$$

$$= \frac{1}{4} n^{3/4} - 4n^{-1/4} < \frac{1}{4} n^{3/4}$$

➔ Pr(E1)+Pr(E2)+Pr(E3a)+Pr(E3b) ≤ $n^{-1/4}$

Example – Parameter Estimation

- We are trying to estimate the parameters of a certain distribution
- For example, judge if a coin is fair or biased
- Suspect that Pr(heads) = p
- Perform n coin flips and let X=n· $\tilde{p}$ be # heads
- Definition: 1−γ Confidence Interval (CI) for a parameter p is an interval [$\tilde{p}$−δ, $\tilde{p}$ +δ] such that

Minimize both

$$\Pr(p \in [\tilde{p} - \delta, \tilde{p} + \delta]) \geq 1 - \gamma.$$

"전국 19세 이상 성인 남녀 1000명을 대상으로 한
설문조사 결과 X, Y 정당 지지율은 각각 40%, 30% 이다.
이번 조사는 신뢰수준 95%, 오차는 ±3.1%포인트다."

Trade-off between n, δ, and γ

Example – Parameter Estimation

- $X = n \cdot \tilde{p}$ is a binomial distribution with n and p

- $p \notin [\tilde{p}-\delta, \tilde{p}+\delta]$ ↔ either

  $p < \tilde{p}-\delta \Rightarrow n\tilde{p} > n(p+\delta) = \mathbf{E}[X](1+\delta/p)$;

  $p > \tilde{p}+\delta \Rightarrow n\tilde{p} < n(p-\delta) = \mathbf{E}[X](1-\delta/p)$.

- From Chernoff bound,

$$\Pr(p \notin [\tilde{p}-\delta, \tilde{p}+\delta]) = \Pr\left(X < np\left(1-\frac{\delta}{p}\right)\right) + \Pr\left(X > np\left(1+\frac{\delta}{p}\right)\right)$$

$$< e^{-np(\delta/p)^2/2} + e^{-np(\delta/p)^2/3}$$

$$= e^{-n\delta^2/2p} + e^{-n\delta^2/3p}.$$

Tighter Bounds for Special Cases

- Case 1: Each trial assumes value 1 or −1 with equal probability
- Theorem: Let $X_1, X_2, \cdots, X_n$ be independent r.v. such that

  $\Pr(X_i=1) = \Pr(X_i=-1) = \frac{1}{2}$. Let $X = \sum_{i=1}^{n} X_i$
- For any a > 0, $\Pr(X \geq a) \leq e^{-a^2/2n}$
- MGF of Xi:

$$\begin{aligned}\mathbf{E}[e^{tX_i}] &= \frac{1}{2}e^{t} + \frac{1}{2}e^{-t}.\\ &= \sum_{i\geq 0}\frac{t^{2i}}{(2i)!}\\ &\leq \sum_{i\geq 0}\frac{(t^2/2)^i}{i!}\\ &= e^{t^2/2}.\end{aligned}$$

$$e^{t} = 1 + t + \frac{t^2}{2!} + \cdots + \frac{t^i}{i!} + \cdots$$

$$e^{-t} = 1 - t + \frac{t^2}{2!} + \cdots + (-1)^i\frac{t^i}{i!} + \cdots$$

- MGF of X:

$$\mathbf{E}[e^{tX}] = \prod_{i=1}^{n}\mathbf{E}[e^{tX_i}] \leq e^{t^2 n/2}$$

- $\Pr(X \geq a) \leq e^{\frac{t^2 n}{2} - ta} = e^{-a^2/2n}$

Min. at t=a/n

$$\frac{\exp\{(e^t - 1)\cdot\mu\}}{e^{t(1+\delta)\mu}}$$

Tighter Bounds for Special Cases

SCONE Lab.

- Case 2: Bernoulli trials with p = 1/2
- Corollary: Let $Y_1, Y_2, \cdots, Y_n$ be independent r.v. such that

  Pr(Yi=1) = Pr(Yi=0) = $\frac{1}{2}$. Let Y = $\sum_{i=1}^{n} Y_i$

  1. For a > 0, $\Pr(Y \geq \mu + a) \leq e^{-2a^2/n}$
  2. For δ > 0, $\Pr(Y \geq (1+\delta)\mu) \leq e^{-\delta^2\mu}$.

- Proof:
  - Let $Y_i = (X_i + 1)/2$, Y=$\sum Y_i = \frac{X}{2} + n/2$
  - $\mu = E[Y] = \frac{n}{2}$
  - $\Pr(Y \geq \mu + a) = \Pr(X \geq 2a) \leq e^{-4a^2/2n})$