Chebyshev's Inequality

Let X be a random variable

For any $a > 0$, $\boxed{P(|X - \mathbb{E}[X]| \geq a) \leq \frac{Var[X]}{a^2} = \frac{\mathbb{E}[X^2] - \mathbb{E}[X]^2}{a^2}}$

probability that X deviates from the expected value of X

Proof:

$$P(|X - \mathbb{E}[X]| \geq a) \leq P((X - \mathbb{E}[X])^2 \geq a^2) \underset{\text{Markov}}{\leq} \frac{\mathbb{E}[(X - \mathbb{E}[X])^2]}{a^2}$$

$$\mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2 - 2 \cdot X \cdot \mathbb{E}[X] + \mathbb{E}[X]^2] \overset{\text{linearity of expectation}}{=} \mathbb{E}[X^2] - 2 \cdot \mathbb{E}[X] \cdot \mathbb{E}[X] + \mathbb{E}[\mathbb{E}[X]^2]$$

$$= \mathbb{E}[X^2] - 2\,\mathbb{E}[X]^2 + \mathbb{E}[X]^2 =$$

$$= \underbrace{\mathbb{E}[X^2] - \mathbb{E}[X]^2}_{\text{Variance of } X}$$

$$\mathbb{E}[\underbrace{2}_{\text{number}} \cdot X \cdot \underbrace{\mathbb{E}[X]}_{\text{number}}] = 2 \cdot \mathbb{E}[X] \cdot \mathbb{E}[X]$$

$$\mathbb{E}[\underbrace{\mathbb{E}[X]^2}_{\text{number}}] = \mathbb{E}[X]^2$$

$\boxed{\mathbb{E}[3x] = 3\mathbb{E}[x]}$

$$P(|X - \mathbb{E}[X]| \geq a) \leq \frac{\mathbb{E}[X^2] - \mathbb{E}[X]^2}{a^2}$$

$$P(|X - E[X]| \geq a) \leq \frac{E[X^2] - E[X]^2}{a^2} = \frac{Var(X)}{a^2}$$

Chebyshev's Inequality.

This is ~~the~~ very useful definition when we have independence.

$X_1, X_2, X_3, \dots, X_N$ are iid RV - independent identically distributed Random Variables.

Define $X = X_1 + X_2 + \dots + X_N$

$$\boxed{Var(X) = Var\left(\sum_{i=1}^{N} X_i\right) = \sum_{i=1}^{N} Var(X_i)}$$ when you have **independence**!

Sort of like linearity of expectation, except now you need independence.

- Is Chebyshev's inequality is a stronger bound than Markov's?
- If you remember how we proved Chebyshev's inequality - we just applied Markov.

ex. Coin flips. Flip a fair coin N times.

X = # of Heads $= \sum_{i=1}^{N} X_i$, $\quad X_i = \begin{cases} 1, & \text{if } i^{th} \text{ flip is Heads} \\ 0, & \text{otherwise} \end{cases}$

With linearity of expectation we showed that $E[X] = N/2$

With Markov's inequality we showed: $P(\underbrace{X}_{} \geq \frac{3}{4}N) \leq \frac{N/2}{(3/4)N} = \frac{2}{3}$

positive random variables, sum of

Let's see what Chebyshev's can offer:

$Var(X_i) = E[X_i^2] - E[X_i]^2 = \frac{1}{2} - \frac{1}{4} = \frac{1}{4}$

because X_i can be 1 or 0, so X_i^2 can be $1^2 = 1$ or $0^2 = 0$.

$E[X_i^2] = 1 \cdot \frac{1}{2} + 0 \cdot \frac{1}{2} = \frac{1}{2}$

$Var(X) = \sum_{i=1}^{N} Var(X_i) = \frac{N}{4}$ since each coin flip is independent.

If A then B then: $P_r(A) \leq P_r(B)$

$P(X \geq \frac{3}{4}N) = P(X \geq \underbrace{\underbrace{E[X]}_{= N/2} + \frac{1}{4}N}_{\frac{3}{4}N}) = P(X - E[X] \geq \frac{N}{4}) \leq$

if we take absolute value of left-hand side then the probability gets bigger.

Apply Chebyshev's Inequality

$\leq P(|X - E[X]| \geq \frac{N}{4}) \leq \frac{Var(X)}{(\frac{N}{4})^2} = \frac{\frac{N}{4}}{(\frac{N}{4})^2} = \frac{4}{N}$ ← That is much smaller, than we can show with Markov.

So, Chebyshev's Inequality is much stronger!!!

Randomized Median Finding using Random Sampling.

Median of N numbers is a number with rank $N/2$. half the numbers are smaller than the Median, and half bigger.

Quick Select ← randomized solves this in $O(N)$ expected. (with a worst case $O(N^2)$)

There are deterministic $\underbrace{\text{algorithms}}_{}$ (like Median of Medians) ← $T(n) \le 10 \cdot c \cdot N$ to solve this in $O(N)$ time. (but they a very complicated)

- **Input:** Set S of N elements.
- **Output:** Median

There is nothing special about $N^{3/4}$, we just want things to cancel nicely. We want N in some power smaller than 1.

$O(N^{3/4})$ 1) Pick $N^{3/4}$ elements at random from S with replacement. Call this set R. (I pick any element with probability $1/N$), but don't remove it from S

[We want R to sort of look like S, to represent S.] but also to be small.

$o(N)$ 2) Sort R in $O(N^{3/4} \cdot \log N)$ time which is $o(N)$ $\boxed{\log N^c = c \cdot \log N}$ const

$O(1)$ 3) Let d = element of rank $\frac{N^{3/4}}{2} - \sqrt{N}$ in R.

$O(1)$ 4) Let u = element of rank $\frac{N^{3/4}}{2} + \sqrt{N}$ in R.

We want true median lie in between. We want u and d to be close together – to have C small, But also far apart enough so the will contain the median **with high probability**

$O(n)$ 5) Let

$$C = \{ x \in S \mid d \le x \le u \}$$
$$\ell_d = |\{ x \in S \mid x < d \}|$$
$$\ell_u = |\{ x \in S \mid x > u \}|$$

← cardinality of a set

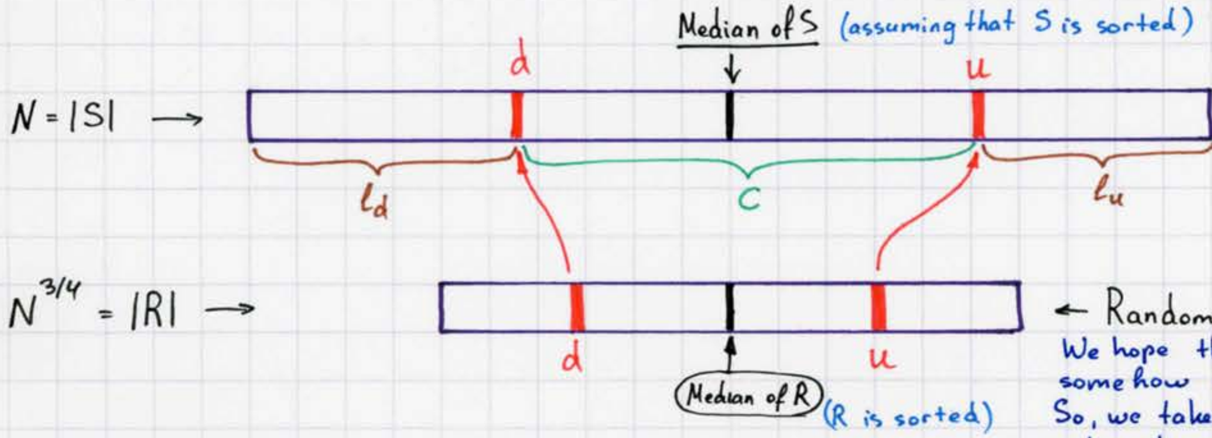
every othes step is sublinear

6) If $\ell_d > \frac{N}{2}$ or $\ell_u > \frac{N}{2}$ FAIL (Because the Median is not in C.)

7) If $|C| > 4N^{3/4}$ then FAIL

8) Sort C and output element of rank : $\frac{N}{2} - \ell_d + 1$

Compare all the items of S to d and u and discard them. Sort elements that are left: C and output the median, because we know how many elements before d and after u.



← Random sample of S

We hope that median of R is somehow close to median of S. So, we take an interval around m. of R and we hope tha it will give an interval around median of S.

If algorithm doesn't fail it runs in linear time.

- What are the bad things that make algorithm fail?

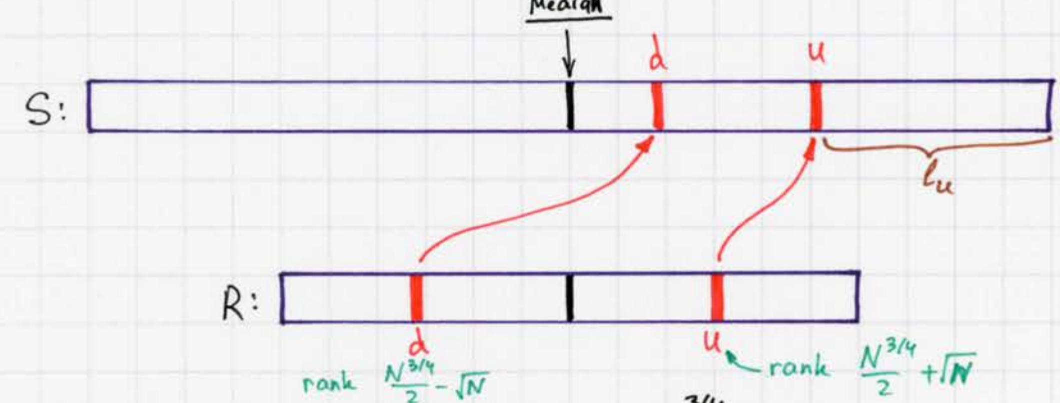
Bad events : E_1 : $\ell_d > N/2$

E_2 : $\ell_u > N/2$

E_3 : $|C| > 4N^{3/4}$

E_1

- If $\ell_d > \frac{N}{2}$ then Median of S is less than d:



How much distance is expected in S between two consequtive elements in R? → $N^{1/4}$

We give a really big window of $2 \cdot \sqrt{N}$

The window in S is expected to be $2 \cdot N^{1/2} \cdot N^{1/4} = 2N^{3/4}$

$$y_1 = |\{ r \in R \mid r \le \text{Median of } S \}| < \frac{N^{3/4}}{2} - \sqrt{N}$$

$$\underbrace{P\left(\ell_d > \frac{N}{2}\right)} \le \underbrace{P\left(y_1 < \frac{N^{3/4}}{2} - \sqrt{N}\right)}$$

If this is true than this is true.

Random Sampling

We want to compute $P(y_1 < \frac{N^{3/4}}{2} - \sqrt{N})$. Let's say we want to apply Chebyshev.

What is the operation I used to build R?
We define an indicator random variable that looks at each of those selections.

Compute $\mathbb{E}[y_1]$ and, $Var(y_1) = \mathbb{E}[y_1^2] - \mathbb{E}[y_1]^2$.

$$X_i = \begin{cases} 1, & \text{if } i^{th} \text{ sample placed in } R \text{ is } \leq \text{Median} \\ 0, & \text{otherwise} \end{cases}$$

[or 1, if i^{th} sample falls in y_1]

$$y_1 = \sum_{i=1}^{N^{3/4}} X_i$$

$\mathbb{E}[X_i] = \frac{1}{2}$ (half elements is smaller than Med. and half (bigger)

$$\mathbb{E}[y_1] = \mathbb{E}\left(\sum_{i=1}^{N^{3/4}} X_i\right) = \sum_{i=1}^{N^{3/4}} \mathbb{E}[X_i] = \sum_{i=1}^{N^{3/4}} \frac{1}{2} = \frac{N^{3/4}}{2}$$

$$Var(X_i) = \mathbb{E}[X_i^2] - \mathbb{E}[X_i]^2 = \frac{1}{2} - \frac{1}{4} = \frac{1}{4}$$

For indicator R.V.s $\mathbb{E}[X_i^2] = \mathbb{E}[X_i]$

That's why we did random sampling with replacement.

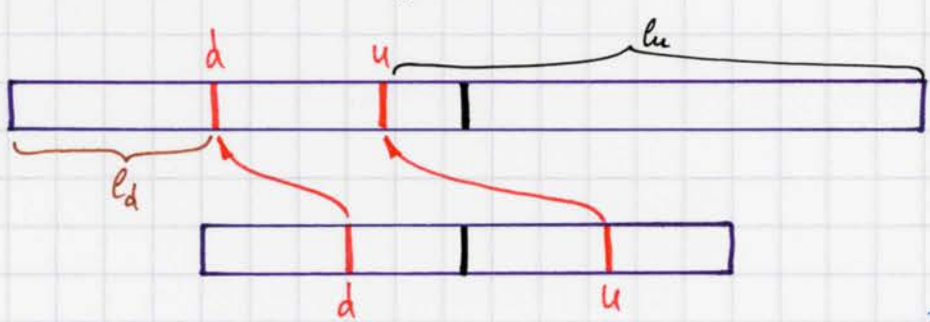
because all variables are independent by construction

$$Var(y_1) = \sum_{i=1}^{N^{3/4}} Var(X_i) = \sum_{i=1}^{N^{3/4}} \frac{1}{4} = \frac{N^{3/4}}{4}$$

$$P\Big(y_1 < \underbrace{\frac{N^{3/4}}{2}}_{=\,\mathbb{E}[y_1]} - \sqrt{N}\Big) \leq P\left(|y_1 - \mathbb{E}[y_1]| > \sqrt{N}\right) \overset{\text{Chebyshev}}{\leq} \frac{Var(y_1)}{(\sqrt{N})^2} \leq \frac{N^{3/4} \cdot \frac{1}{4}}{N} = \underbrace{\frac{1}{4N^{1/4}}}_{\text{small}}$$

E_2

- If $l_u > \frac{N}{2}$ then Median of S is bigger than u:



$$y_2 = |\{r \in R \mid r \geq \text{Median of } S\}|.$$

In the exact way (as with E_1 case) we can show that $P\Big(y_2 < \overbrace{\frac{N^{3/4}}{2} - \sqrt{N}}^{N^{3/4} - \left(\frac{N^{3/4}}{2} + \sqrt{N}\right)}\Big) \leq \frac{1}{4N^{1/4}}$

(Use exactly the same indicator variables and so on...).

E_3: $|C| > 4N^{3/4}$ ← is a bad event

We break this event into two events:

$\mathcal{E}_1$: at least $2N^{3/4}$ elements in C are $\geq$ Median.

$\mathcal{E}_2$: at least $2N^{3/4}$ elements in C are $\leq$ Median.

$$P(|C| > 4N^{3/4}) \leq P(\mathcal{E}_1) + P(\mathcal{E}_2)$$

$P(\mathcal{E}_1)$: What is a rank of u in R? By definition, it is $\frac{N^{3/4}}{2} + \sqrt{N}$ (that's how we picked u).

How many elements in R are bigger than or equal to u? $N^{3/4} - \left(\frac{N^{3/4}}{2} + \sqrt{N}\right) = \frac{N^{3/4}}{2} - \sqrt{N}$.

What is the rank of u in S? – At least $\frac{N}{2} + 2N^{3/4}$ because we are in case $\mathcal{E}_1$.

What is the size of ℓ_u? – at most $N - \left(\frac{N}{2} + 2N^{3/4}\right) = \frac{N}{2} - 2N^{3/4}$

Among $\frac{N}{2} - 2N^{3/4}$ elements in ℓ_u we picked $\frac{N^{3/4}}{2} - \sqrt{N}$ of them and put them in R.

$$X_i = \begin{cases} 1 & \text{if } i^{th} \text{ sample is among the } \frac{N}{2} - 2N^{3/4} \text{ elements.} \\ 0 & \text{otherwise.} \end{cases}$$

$X = |\{r \in R \mid r > u\}|$

$$X = \sum_{i=1}^{N^{3/4}} X_i \quad ; \qquad \mathbb{E}[X] = \mathbb{E}\left[\sum_{i=1}^{N^{3/4}} X_i\right] = \sum_{i=1}^{N^{3/4}} \mathbb{E}[X_i] = \sum_{i=1}^{N^{3/4}} P(X_i = 1) =$$

What is $P(X_i = 1)$? Each element is equally likely; $\dfrac{\overbrace{\frac{N}{2} - 2N^{3/4}}^{\text{Choices for } X_i \text{ to be } 1}}{\underbrace{N}_{\text{All choices}}} = \frac{1}{2} - \frac{2}{N^{1/4}}$

$$\sum_{i=1}^{N^{3/4}} P(X_i = 1) = N^{3/4} \cdot \left(\frac{1}{2} - \frac{2}{N^{1/4}}\right) = \boxed{\frac{N^{3/4}}{2} - 2\sqrt{N} = \mathbb{E}[X]}$$

$$P(\mathcal{E}_1) = P\left(X \geq \frac{N^{3/4}}{2} - \sqrt{N}\right) = P\left(X - \mathbb{E}[X] \geq \frac{N^{3/4}}{2} - \sqrt{N} \overbrace{- \frac{N^{3/4}}{2} + 2\sqrt{N}}^{-E[X]}\right) =$$

$$= P\left(X - \mathbb{E}[X] \geq \sqrt{N}\right) \leq P\left(|X - \mathbb{E}[X]| \geq \sqrt{N}\right) \overset{\text{Chebyshev}}{\leq} \frac{Var(X)}{(\sqrt{N})^2} \leq \frac{N^{3/4}}{4N} = \frac{1}{4N^{1/4}}$$

$$\boxed{\text{HW: } Var(X) = \sum_{i=1}^{N^{3/4}} Var(X_i) \leq \frac{N^{3/4}}{4}}$$

Very small probability that $\mathcal{E}_1$ happens

So, $P(|C| > 4N^{3/4}) \leq P(\mathcal{E}_1) + P(\mathcal{E}_2) \leq \frac{1}{4N^{1/4}} + \frac{1}{4N^{1/4}} = \frac{1}{2N^{1/4}}$

$$P\left(l_u > \frac{N}{2}\right) + P\left(l_d > \frac{N}{2}\right) + P\left(|C| > 4N^{3/4}\right) \leq \frac{1}{4N^{1/4}} + \frac{1}{4N^{1/4}} + \frac{1}{2N^{1/4}} \leq \frac{1}{N^{1/4}}$$

42:00

$$P_r(X_i = 1) = \frac{1}{2} - \frac{2}{N^{1/4}}$$

$$Var(X_i) = \mathbb{E}(X_i^2) - \mathbb{E}(X_i)^2 = \mathbb{E}(X_i) - \mathbb{E}(X_i)^2 = \frac{1}{2} - \frac{2}{N^{1/4}} - \left(\frac{1}{2} - \frac{2}{N^{1/4}}\right)^2 =$$

$$= \frac{1}{2} - \frac{2}{N^{1/4}} - \frac{1}{4} + \frac{2}{2} \cdot \frac{2}{N^{1/4}} - \frac{4}{N^{1/2}} = \boxed{\frac{1}{4} - \frac{4}{N^{1/2}}}$$

$$Var(X) = Var\left(\sum_{i=1}^{N^{3/4}} X_i\right) = \sum_{i=1}^{N^{3/4}} Var(X_i) = N^{3/4} \cdot \left(\frac{1}{4} - \frac{4}{N^{1/2}}\right) = \frac{N^{3/4}}{4} - 4N^{1/4} \leq \frac{N^{3/4}}{4}$$