Chebyshev's Inequality

Let $X$ be a random variable.

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

Proof:

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


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

Chebyshev's Inequality.

This is the useful definition when we have independence.

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

Define $X = X_1 + X_2 + ... + X_N$

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

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

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

**Example**

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

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

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

With Markov's inequality we showed: $P(X > \frac{3}{4}N) \leq \frac{N/2}{(3/4)N} = \frac{2}{3}$ positive random variables, sum of

Let's see what Chebyshev's can offer:

$\text{Var}(X_i) = \frac{1}{2} - \frac{1}{4} = \frac{1}{4}$

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

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

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

$P(X > \frac{3}{4}N) = P(X > E[X] + \frac{1}{4}N) = P\left(\frac{X - E[X]}{\frac{N}{4}} > \frac{N/2}{\frac{N}{4}}\right) \leq \frac{\text{Var}(X)}{\left(\frac{N/2}{\frac{N}{4}}\right)^2} = \frac{\frac{N}{4}}{(\frac{N}{4})^2} = \frac{4}{N}$

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

That is much smaller, than we can show with Markov.

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

$\leq P\left(1 - E[X]\right) \leq \frac{\text{Var}(X)}{\left(\frac{N}{4}\right)^2} = \frac{4}{N}$

So, Chebyshev's Inequality is much stronger!!!
Randomized Median Finding using Random Sampling.

Median of N numbers is a number with rank $N/2$.

Quick Select solves this in $O(N)$ expected, (with a worst case $O(N^3)$) (like Median of Medians) $T(n) \leq 10\cdot c\cdot N$

There are deterministic algorithms to solve this in $O(N)$ time. (but they a very complicated)

• Input: Set S of N elements.
• Output: Median

$O(N^{3/4})$
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.]

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

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

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

5) Let $C = \{ x \in S \mid d \leq x \leq u \}$
   $|C| = \{ x \in S \mid x < d \}$
   $|C| = \{ x \in S \mid x > u \}$
   $\Rightarrow$ cardinality of a set

$O(n)$

every other step is sublinear
6) If \( l_d > \frac{N}{2} \) or \( l_u > \frac{N}{2} \) FAILURE (because the Median is not in \( C \)).

7) If \( |C| > 4N^{3/4} \) then FAILURE.

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

---

If algorithm doesn't fail it runs \( \text{Linear time} \).

What are the bad things that make algorithm fail?

Bad events: \( E_1 : l_d > \frac{N}{2} \)

\[ E_2 : l_u > \frac{N}{2} \]

\[ E_3 : |C| > 4N^{3/4} \]

- If \( l_d > \frac{N}{2} \) then Median of \( S \) is less than \( d \):

\[ Y_1 = \{ r \in R \mid r \leq \text{Median of } S \} < \frac{N^{3/4}}{2} - \sqrt{N} \]

\[ P(l_d > \frac{N}{2}) \leq P(Y_1 < \frac{N^{3/4}}{2} - \sqrt{N}) \]

If this is true then this is true.
We want to compute \( P(\bar{Y}_1 < \frac{N^{3/4}}{2} - \sqrt{N}) \). Let's say we want to apply Chebyshev.

Compute \( E[\bar{Y}_1] \) and \( \text{Var}(\bar{Y}_1) = E[\bar{Y}_1^2] - E[\bar{Y}_1]^2 \).

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

\[
\bar{Y}_1 = \frac{1}{N} \sum_{i=1}^{N^{3/4}} X_i
\]

\[
E[\bar{Y}_1] = E\left( \sum_{i=1}^{N^{3/4}} X_i \right) = \sum_{i=1}^{N^{3/4}} E[X_i] = \sum_{i=1}^{N^{3/4}} \frac{1}{2} = \frac{N^{3/4}}{2}
\]

\[
\text{Var}(\bar{Y}_1) = \sum_{i=1}^{N^{3/4}} \text{Var}(X_i) = \sum_{i=1}^{N^{3/4}} \frac{1}{4} = \frac{N^{3/4}}{4}
\]

Chebyshev

\[
P(\bar{Y}_1 < \frac{\sqrt{N}}{2} - \frac{N^{3/4}}{2}) \leq P\left( |\bar{Y}_1 - E[\bar{Y}_1]| > \sqrt{N} \right) \leq \frac{\text{Var}(\bar{Y}_1)}{(\sqrt{N})^2} = \frac{\frac{N^{3/4}}{4}}{N} = \frac{1}{4N^{1/4}}
\]

E2

- If \( u_1 > \frac{N}{2} \) then Median of S is bigger than u:

\[
Y_2 = \{ r \in R \mid r \geq \text{Median of } S \}
\]

In the exact as with E2 case we can show that

\[
P(Y_2 < \frac{N^{3/4}}{2} - \frac{N^{3/4}}{2}) \leq \frac{1}{4N^{1/4}}
\]

(Use exactly the same indicator variables and so on...)
\( E_1: \) at least \( 2N^{3/4} \) elements in \( C \) are \( \geq \) Median.

\( E_2: \) at least \( 2N^{3/4} \) elements in \( C \) are \( \leq \) Median.

\[ P(1C1 > 4N^{3/4}) \leq P(E_1) + P(E_2) \]

\( P(E_1): \)

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 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 \( E_2 \).

What is the size of \( bu \)? 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 \( bu \) we picked \( \frac{N^{3/4}}{2} - \sqrt{N} \) of them and put them in \( R \).

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

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

What is \( P(X_i = 1) \)? Each element is equally likely;

\[ \sum_{i=1}^{N^{3/4}} P(X_i = 1) = N^{3/4} \cdot \left( \frac{1}{2} - \frac{2}{N^{1/4}} \right) = \frac{N^{3/4}}{2} - 2\sqrt{N} = E[X] \]
\[ P(\varepsilon_1) = P(X > \frac{N^{3/4}}{2} - \sqrt{N}) = P(X - \mu > \frac{N^{3/4}}{2} - \sqrt{N} - \frac{N}{2} + 2\sqrt{N}) = \]
\[ = P(X - \mu > \sqrt{N}) \leq P(|X - \mu| > \sqrt{N}) \leq \frac{\text{Var}(X)}{(\sqrt{N})^2} \leq \frac{N^{3/4}}{4N^{1/4}} = \frac{1}{4N^{1/4}} \]

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

Very small probability that \( \varepsilon_1 \) happens

So,
\[ P(1C1 > 4N^{3/4}) \leq P(\varepsilon_1) + P(\varepsilon_2) \leq \frac{1}{4N^{1/4}} + \frac{1}{4N^{1/4}} = \frac{1}{2N^{1/4}} \]

\[ P(l_u > \frac{N}{2}) + P(l_d > \frac{N}{2}) + P(1C1 > 4N^{3/4}) \leq \frac{1}{4N^{1/4}} + \frac{1}{4N^{1/4}} + \frac{1}{2N^{1/4}} \leq \frac{1}{N^{1/4}} \]

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

\[ \text{Var}(X_i) = \text{E}(X_i^2) - \text{E}(X_i)^2 = \text{E}(X_i) - \text{E}(X_i)^2 = \frac{1}{2} - \frac{2}{N^{1/4}} - (\frac{1}{2} - \frac{2}{N^{1/4}})^2 = \]
\[ = \frac{1}{2} - \frac{2}{N^{1/4}} - \frac{1}{4} + \frac{2}{2N^{1/4}} - \frac{1}{N^{1/2}} = \frac{1}{4} - \frac{1}{N^{1/4}} \]

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