Concentration around the Mean

Josep Díaz Maria J. Serna Conrado Martínez
U. Politècnica de Catalunya

RA-MIRI 2023–2024

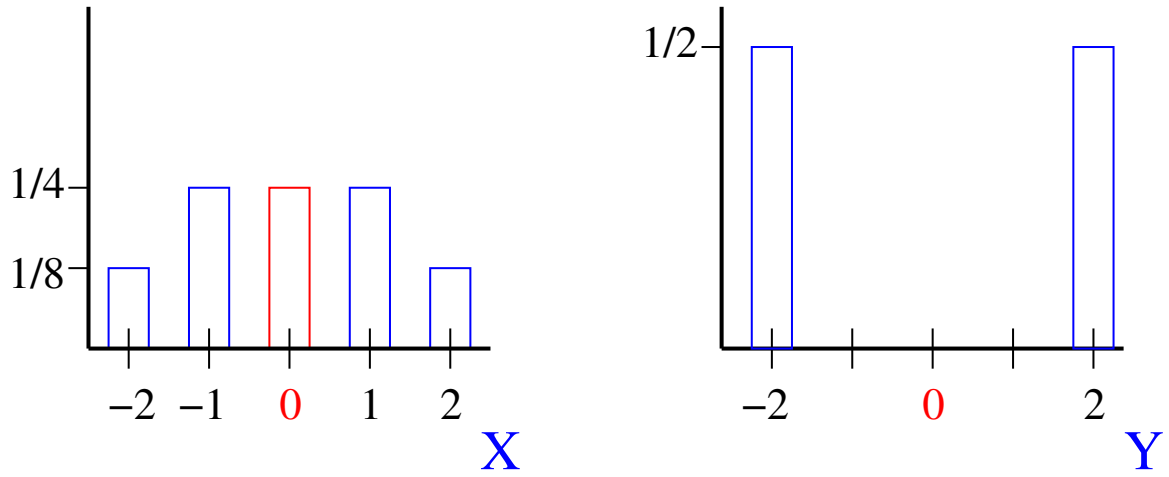
Deviations from the mean

The expected value of a random variable is a nice single number to summarize the random variable, but it leaves out most of the important properties of the r.v.

Consider r.v X with $X(\Omega) = \{-2, -1, 0, 1, 2\}$ with $\mathbb{P}[X = -2] = \frac{1}{8}, \mathbb{P}[X = -1] = \frac{1}{4}, \mathbb{P}[X = 0] = \frac{1}{4}, \mathbb{P}[X = 1] = \frac{1}{4}, \mathbb{P}[X = 2] = \frac{1}{8}$.

and consider r.v. Y with $Y(\Omega') = \{-2, 2\}$ and PMF: $\mathbb{P}[Y = -2] = \frac{1}{2}, \mathbb{P}[Y = 2] = \frac{1}{2}$.

Note that $\mathbb{E}[X] = \mathbb{E}[Y] = 0$, but p_X is totally different from p_Y.



Deviations from the mean

- Consider the deterministic Quicksort algorithm on n-size inputs. Let $T(n)$ be a r.v. counting the number of steps of Quicksort on a specific input with size n
- Its worst case complexity is $\mathcal{O}(n^2)$, but its average complexity is $\mathcal{O}(n \log n)$.
- It does not give information about the behavior of the algorithm on a particular input.
- Given an algorithm, for any input x of size $|x| = n$, how close is $T(x)$ to $\mathbb{E}[T(n)]$.

Deviation of a r.v. and concentration

- For ex.: If $\mathbb{E}[\mathrm{T}(n)] = 10$, then 10 is an average running time on "most inputs" to the algorithm. We want to assure, that for most inputs, $\mathrm{T}(n)$ is concentrated around 10.
- That is, to make sure that the probability of having instances for which $|\mathbb{E}[\mathrm{T}(n)] - \mathrm{T}(n)|$ is large, is very small.
- Intuitively, it seems clear from the definition of $\mathbb{E}[]$, if for the above running time, we get an instance e for which $\mathrm{T}(e) = 10^9$, and $\mathbb{E}[\mathrm{T}(n)] = 10$, the probability of selecting that specific e is going to be quite small, so that its contribution to the average, $10^9\,\mathbb{P}\left[\mathrm{T}(n) = 10^9\right]$, is small.

Markov's inequality

Andrey Markov (1856–1922)

Lemma

If $X \geqslant 0$ *is a r.v, for any constant* $a > 0$,

$$\mathbb{P}[X \geqslant a] \leqslant \frac{\mathbb{E}[X]}{a}.$$

Markov's inequality

Proof

Given the r.v. $X \geqslant 0$ define the indicator r.v.

$$Y = \begin{cases} 1 & \text{if } X \geqslant a \text{ true} \\ 0 & \text{otherwise} \end{cases}$$

Notice if $Y = 1$ then $Y \leqslant X/a$, and if $Y = 0$ also $Y \leqslant X/a$, so $\mathbb{E}[Y] = \mathbb{P}[Y = 1] = \mathbb{P}[X \geqslant a]$ and $\mathbb{E}[Y] = \mathbb{P}[Y = 1] \leqslant \mathbb{E}\left[\frac{X}{a}\right] = \frac{\mathbb{E}[X]}{a}$. □

Markov's inequality

Alternative expression for Markov. Taking $a = b \cdot \mathbb{E}[X]$:

Corollary

If $X \geqslant 0$ *is a r.v, for any constant* $b > 0$,

$$\mathbb{P}[X \geqslant b \cdot \mathbb{E}[X]] \leqslant \frac{1}{b}.$$

Markov's inequality

Consider the randomized hiring algorithm. We computed that the expected number of pre-selected students is $\mathbb{E}[X] = \ln n$. We also know there are instances for which $X = n$.

We would like to show that the probability of selecting a "bad instance" is very small.

Using Markov's inequality, for any constant b, $\mathbb{P}[X \geqslant b \ln n] \leqslant 1/b$. (for ex. $b = 100$)

The problem with Markov is that it does not bound away the probability of *bad cases* as a function of the input size.

“With High Probability”

In the randomized algorithms, we aim to obtain results that hold with high probability, in particular, that the probability that the complexity of the algorithm for any input is “near” the expected value tends to 1 as the size n grows.
An event is said to occur with high probability (whp) if its probability is $\geqslant 1 - \frac{1}{f(n)}$, for some function $f(n) = \Omega(n^c)$ with $c > 0$, so that the probaility goes to 1 as $n \to \infty$.

The parameter n is usually the size of the inputs or the size of the combinatorial structure.

Variance

Given a r.v. X, its variance measures the spread of its distribution.

Given X, with $\mu = \mathbb{E}[X]$, the variance of X is

$$\mathbb{V}[X] = \mathbb{E}\left[(X-\mu)^2\right]$$

Usually it is more easy to use the expression:

$$\mathbb{V}[X] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2$$

Proof

$$\begin{aligned}\mathbb{V}[X] &= \mathbb{E}\left[(X-\mu)^2\right] = \mathbb{E}\left[X^2 - 2\mu\,\mathbb{E}[X] + \mu^2\right] \\ &= \mathbb{E}\left[X^2\right] - 2\mu\underbrace{\mathbb{E}[X]}_{\mu} + \mu^2 = \mathbb{E}\left[X^2\right] - \mu^2\end{aligned}$$

Further properties of the variance

- Since $(X - \mu)^2 \geqslant 0$ for any outcome $\omega \in \Omega$, we must have $\mathbb{V}[X] \geqslant 0$. Alternatively, since $f(x) = x^2$ is convex, by Jensen's inequality, $\mathbb{E}\left[X^2\right] \geqslant \mathbb{E}[X]^2$.
- $\mathbb{V}[X] = 0$ iff $X =$ constant
- For any constant c, $\mathbb{V}[cX] = c^2\,\mathbb{V}[X]$.

Computing $\mathbb{V}[X]$

Given a discrete r.v. X on Ω, such that $X(\Omega) = \{x_1, x_2, \ldots, x_n\}$, we first compute $\mu = \mathbb{E}[X] = \sum_{i=1}^{n} x_i \, \mathbb{P}[X = x_i]$. Then, use one of the following methods:

1. Use $\mathbb{V}[X] = \mathbb{E}\left[(X - \mu)^2\right]$: For each x_i compute $(x_i - \mu)^2$, and then $\mathbb{V}[X] = \sum_{i=1}^{n} (x_i - \mu)^2 \, \mathbb{P}[X = x_i]$
2. Use $\mathbb{V}[X] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2$: For each x_i compute x_i^2, then $\mathbb{E}\left[X^2\right] = \sum_{i=1}^{n} x_i^2 \, \mathbb{P}[X = x_i]$.

From now on, we use the probability mass function of X, $p_X : \mathbb{R} \to [0, 1]$, defined as $p_X(x) = \mathbb{P}[X = x]$—note that p_X is actually defined on a finite or infinite denumerable subset of $\mathbb{R}$ as X is a discrete r.v.

Computing $\mathbb{V}[X]$: Examples

Example

Consider r.v. X with $X(\Omega) = \{1, 3, 5\}$ and PMF: $p_X(1) = \frac{1}{4}, p_X(3) = \frac{1}{4}, P_X(5) = \frac{1}{2}$. Then $\mu = 7/2$.

1. $\mathbb{V}[X] = \frac{1}{4}(3 - \frac{7}{2})^2 + \frac{1}{4}(5 - \frac{7}{2})^2 + \frac{1}{2}(1 - \frac{7}{2})^2 = \frac{11}{4}$
2. $X^2(\Omega) = \{1, 9, 25\}$, so $\mathbb{E}[X^2] = \frac{1}{4} + \frac{9}{4} + \frac{25}{2} = 15$
   $\mathbb{V}[X] = 15 - (\frac{7}{2})^2 = \frac{11}{4}$

Example

Consider r.v. Y with $X(\Omega) = \{-2, 2\}$ and PMF: $p_Y(-2) = \frac{1}{2}, p_Y(2) = \frac{1}{2}$.
Therefore, the values $(X - \mu)^2$ are $(-2 - 0)^2$ and $(2 - 0)^2$
$\Rightarrow \mathbb{V}[X] = \frac{1}{2}4 + \frac{1}{2}4 = 4$
Notice in this case $\mathbb{V}[X] = \mathbb{E}[X^2] = 4$

Computing $\mathbb{V}[X]$: Examples

Example

You win 100€ with probability $= \quad 1/10$, otherwise you win 0€. Let X be a r.v. counting your earnings. What is $\mathbb{V}[X]$?

$\mu = 100/10 = 10$. Therefore, $\mathbb{E}[X^2] = \frac{1}{10}(100^2) = 1000$ and $\mu^2 = 100$, hence $\mathbb{V}[X] = 900$.

Variance of $X + Y$

Let $X_1, \ldots, X_n$ be independent r.v., then

$$\mathbb{V}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{V}[X_i]\,.$$

We prove the particular case that if X and Y are independent $\mathbb{V}[X+Y] = \mathbb{V}[X] + \mathbb{V}[Y]$.

Proof

$$\begin{aligned}
\mathbb{V}[X+Y] &= \mathbb{E}\left[(X+Y)^2\right] - (\mathbb{E}[X+Y])^2 \\
&= \mathbb{E}\left[X^2\right] + \mathbb{E}\left[Y^2\right] + 2\,\mathbb{E}[XY] - (\mathbb{E}[X])^2 - (\mathbb{E}[Y])^2 - 2\,\mathbb{E}[X]\,\mathbb{E}[Y] \\
&= \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2 + \mathbb{E}\left[Y^2\right] - (\mathbb{E}[Y])^2 + 2\underbrace{(\mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y])}_{\mathbb{E}[XY]=\mathbb{E}[X]\,\mathbb{E}[Y]} \\
&= \mathbb{V}[X] + \mathbb{V}[Y]
\end{aligned}$$

□

Variance of $X + Y$

In general,

$$\begin{aligned}
\mathbb{V}[X+Y] &= \mathbb{E}\Big[(X+Y-\mathbb{E}[X+Y])^2\Big] = \mathbb{E}\Big[(X+Y)^2\Big] - (\mathbb{E}[X]+\mathbb{E}[Y])^2 \\
&= \mathbb{E}\Big[X^2\Big] + \mathbb{E}\Big[Y^2\Big] + 2\,\mathbb{E}[XY] - \mathbb{E}[X]^2 - \mathbb{E}[Y]^2 - 2\,\mathbb{E}[X]\,\mathbb{E}[Y] \\
&= \mathbb{V}[X] + \mathbb{V}[Y] + 2(\mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]) \\
&= \mathbb{V}[X] + \mathbb{V}[Y] + 2\,\mathrm{Cov}[X,Y]
\end{aligned}$$

$$\mathrm{Cov}[X,Y] = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]$$

is called the covariance of the two r.v. X and Y.

- If X and Y are independent $\mathrm{Cov}[X,Y] = 0$.
- For any r.v. X, $\mathrm{Cov}[X,X] = \mathbb{V}[X]$.

Variance of some basic distributions

1. If $X \in \text{Bin}(p, n)$ then $\mathbb{V}[X] = npq$, where $q = (1 - p)$.
2. If $X \in \text{Poisson}(\lambda)$ then $\mathbb{V}[X] = \lambda$.
3. If $X \in \text{Geom}(p)$ then $\mathbb{V}[X] = \frac{q}{p^2}$.

Proof

(1) Let $X = \sum_{i=1}^{n} X_i$, where X_i is an indicator r.v s.t. $X_i = 1$ with probability p
Then, $\mathbb{V}[X_i] = \mathbb{E}\left[X_i^2\right] - \mathbb{E}[X_i]^2 = (p \cdot 1^2 + q \cdot 0 - p^2 = p(1-p)$.
Since all X_i are independent, $\mathbb{V}[X] = \sum_{i=1}^{n} \mathbb{V}[X_i] = np(1-p)$.

Variance of some basic distributions

Proof (cont'd)

(2)

$$\mathbb{V}[X] = \mathbb{E}\left[X^{\underline{2}}\right] + \mathbb{E}[X] - (\mathbb{E}[X])^2 = \mathbb{E}\left[X^{\underline{2}}\right] + \lambda - \lambda^2.$$

$$\begin{aligned}
\mathbb{E}\left[X^{\underline{2}}\right] &= \sum_{x=0}^{\infty} x^{\underline{2}} \frac{e^{-\lambda}\lambda^x}{x!} \\
&= \sum_{x=2}^{\infty} x \cdot (x-1) \frac{e^{-\lambda}\lambda^x}{x!} \qquad \text{terms } x = 0 \text{ and } x = 1 \text{ are 0} \\
&= \sum_{x=2}^{\infty} \frac{e^{-\lambda}\lambda^x}{(x-2)!} = \lambda^2 e^{-\lambda} \sum_{x=2}^{\infty} \frac{\lambda^{x-2}}{(x-2)!} \\
&= \lambda^2 e^{-\lambda} \left(\frac{\lambda^0}{0!} + \frac{\lambda^1}{1!} + \frac{\lambda^2}{2!} + \dots \right) \\
&= \lambda^2 e^{-\lambda} e^{\lambda} = \lambda^2.
\end{aligned}$$

$x^{\underline{k}} = x \cdot (x-1) \cdots (x-k+1)$

Variance of some basic distributions

Proof (cont'd)

(3)

If $X \in \text{Geom}(p)$ want to compute $\mathbb{V}[X] = \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2 = \mathbb{E}\left[X^2\right] - \frac{1}{p^2}$.

Need to compute $\mathbb{E}\left[X^2\right]$.

$$\mathbb{E}\left[X^2\right] = \sum_{k=1}^{\infty} k^2\,\mathbb{P}[X=k] = \sum_{k=1}^{\infty} k^2 p(1-p)^{k-1} = p\underbrace{\sum_{k=1}^{\infty} k^2(1-p)^{k-1}}_{*}$$

Recall Taylor: $\frac{1}{1-x} = \sum_{k=0}^{\infty} x^k$. Differentiating $\frac{1}{(1-x)^2} = \sum_{k=1}^{\infty} kx^{k-1}$.

Multiplying by x and differentiating $\frac{x+1}{(1-x)^3} = \sum_{k=1}^{\infty} k^2 x^{k-1}$.

Making $x = 1-p$ then $\frac{2-p}{p^3} = \sum_{k=1}^{\infty} k^2(1-p)^{k-1}$.

By $(*)$ $\mathbb{E}\left[X^2\right] = \frac{2-p}{p^2}$

Therefore: $\mathbb{V}[X] = \frac{2-p}{p^2} - \frac{1}{p^2} = \frac{1-p}{p^2}$ □

Standard deviation

Why we did not define $\mathbb{V}[X] = \mathbb{E}[|X - \mu|]$?
This would be a natural measure of the spread of the r.v. X—all deviations from $\mathbb{E}[X]$ must contribute positively. However, the function absolute value $|\cdot|$ is not differentiable everywhere and it is unfriendly for mathematical manipulation. By "squaring" the errors we make sure all them contribute positively and the function $f(x) = x^2$ behaves nicely from the point of view of the analysis.

Standard deviation

But as we defined the variance, we are using squared units!

Recall the example with X a r.v. counting the wins, when you win 100€ with probability $= 1/10$, otherwise you win 0€. We got $\mathbb{V}[X] = 900€^2$.
To convert the numbers back to the same scale, we need to take the square root.

Definition

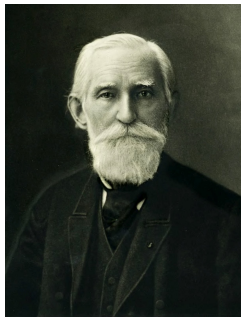
The standard deviation of a r.v. X is defined as

$$\sigma_X = \sqrt{\mathbb{V}[X]}.$$

Example

In our last example, $\sigma_X = \sqrt{900€^2} = 30€$.

Chebyshev's inequality



Pafnuty Chebychev (1821–1894)

If you can compute $\mathbb{V}[X]$ ($\Rightarrow \sigma_X$) then you can get better bounds for concentration of X (positive or negative) around its expected value.

Theorem

Let X be a r.v. with expectation $\mu = \mathbb{E}[X]$ and standard deviation $\sigma = \sigma_X > 0$, then for any $a > 0$

$$\mathbb{P}[|X - \mu| \geqslant a\sigma] \leqslant \frac{1}{a^2}$$

Note that $|X - \mu| \geqslant a\sigma \Leftrightarrow (X \geqslant a\sigma + \mu) \cup (X \leqslant \mu - a\sigma)$.

Chebyshev's inequality

Proof

As the r.v. $|X - \mu| \geqslant 0$, we can apply Markov to it:

$$\begin{aligned}\mathbb{P}[|X - \mu| \geqslant a\sigma] &= \mathbb{P}\left[(X-\mu)^2 \geqslant a^2\sigma^2\right] \qquad \text{(by Markov's ineq.)} \\ &\leqslant \frac{\mathbb{E}\left[(X-\mu)^2\right]}{a^2\sigma^2} = \frac{\mathbb{V}[X]}{a^2\,\mathbb{V}[X]} = \frac{1}{a^2}\end{aligned}$$

□

Chebyshev's inequality

Alternative equivalent statement of Chebyshev's inequality: For all $b > 0$

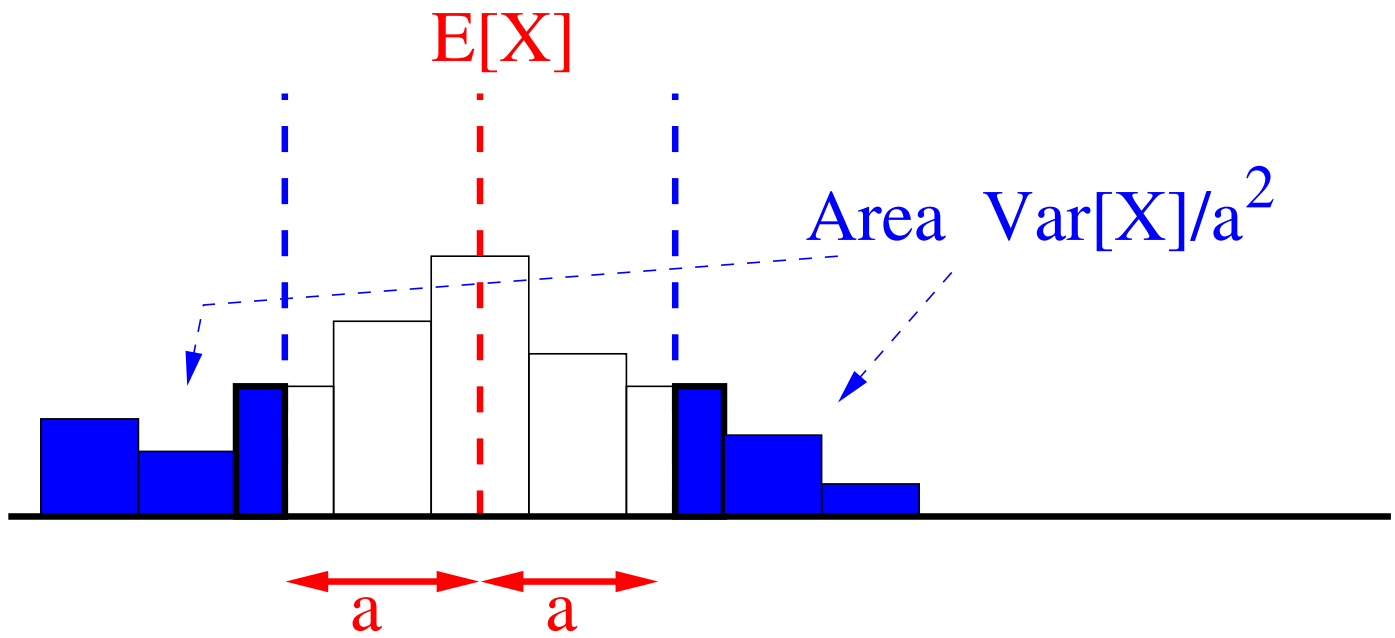
$$\mathbb{P}[|X - \mu| \geqslant b] \leqslant \frac{\mathbb{V}[X]}{b^2}$$

Proof

As before: $\mathbb{P}\left[(X-\mu)^2 \geqslant b^2\right] \leqslant \frac{\mathbb{E}[(X-\mu)^2]}{b^2}$. □

Chebyshev's inequality

$$\mathbb{P}[|X - \mu| \geqslant a] \leqslant \frac{\mathbb{V}[X]}{a^2}.$$



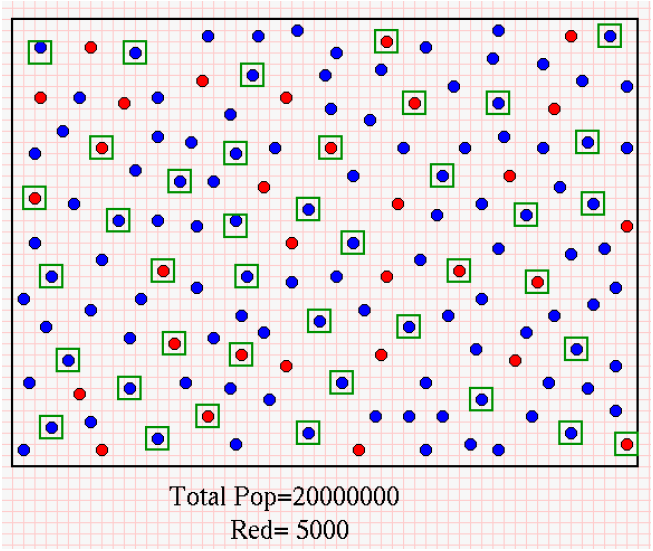
An easy application

Example

Flip n times a fair coin, give an upper bound on the probability of having at least $\frac{3n}{4}$ heads.
Let X be the number of heads. Then $X \sim \mathrm{Bin}(n, 1/2)$, $\mu = \mathbb{E}xpX = n/2$, and $\mathbb{V}[X] = n/4$.
Thus we want to bound $\mathbb{P}\left[X \geqslant \frac{3n}{4}\right]$.

- Markov: $\mathbb{P}\left[X \geqslant \frac{3n}{4}\right] \leqslant \frac{\mu}{3n/4} = 2/3$.
- Chebyshev: We need the value of a s.t.
  $\mathbb{P}\left[X \geqslant \frac{3n}{4}\right] \leqslant \mathbb{P}\left[|X - \frac{n}{2}| \geqslant a\right] \Rightarrow a = \frac{3n}{4} - \frac{n}{2} = \frac{n}{4}$.
  $\mathbb{P}\left[X \geqslant \frac{3n}{4}\right] \leqslant \mathbb{P}\left[|X - \frac{n}{2}| \geqslant \frac{n}{4}\right] \leqslant \frac{\mathbb{V}[X]}{(n/4)^2} = \frac{4}{n}$.

Sampling



- Given a large population Σ, $|\Sigma| = n$, we wish to estimate the proportion p of elements in Σ, with a given property.
- Sampling: Take a random sample S with size $m \ll n$ and compute
  $\hat{p}$ = fraction of elements in S that have the property.
- If n is large and $0 < p < 1$ the $\hat{p}$ is an unbiased estimator of p and sufficiently good, i.e. it is sharply concentrated.
- Many times getting the random sample S is non-trivial.

Finding the median of n elements

From MU 3.4

- Recall that, given a set S with n distinct elements, the median of S is the $\lceil n/2 \rceil$-th smallest element in S.
- We can use Quickselect to find the median with expected time $\mathcal{O}(n)$. Even there is a linear time deterministic algorithm, but in practice is worse than Quickselect.
- We present another randomized algorithm to find the median m in S, which is based in sampling.
- The purpose of this example is to introduce the technique of filtering a large data by sampling an small amount of the data.

Finding the median of n elements

INPUT: An unordered set $S = \{x_1, x_2, \ldots x_n\}$, with $n = 2k + 1$ elements.
OUTPUT: The median, which is the $(k + 1)$-th smallest element in S.
For any element y define the rank$(y) = |\{x \in S | x \leqslant y\}|$.

The idea of the filtering algorithm is to sample with replacement a "small" subset C of elements from S, so we can sort C in time $\mathcal{O}(n)$ (linear with respect to the size of S).
The algorithm outputs **fail** if C turns out to be too large to sort it in time $\mathcal{O}(n)$ or if it doesn't contain the sought median—can be checked in linea time. Otherwise it finds the median of the elements in C sorting it and returns it as the median in S.

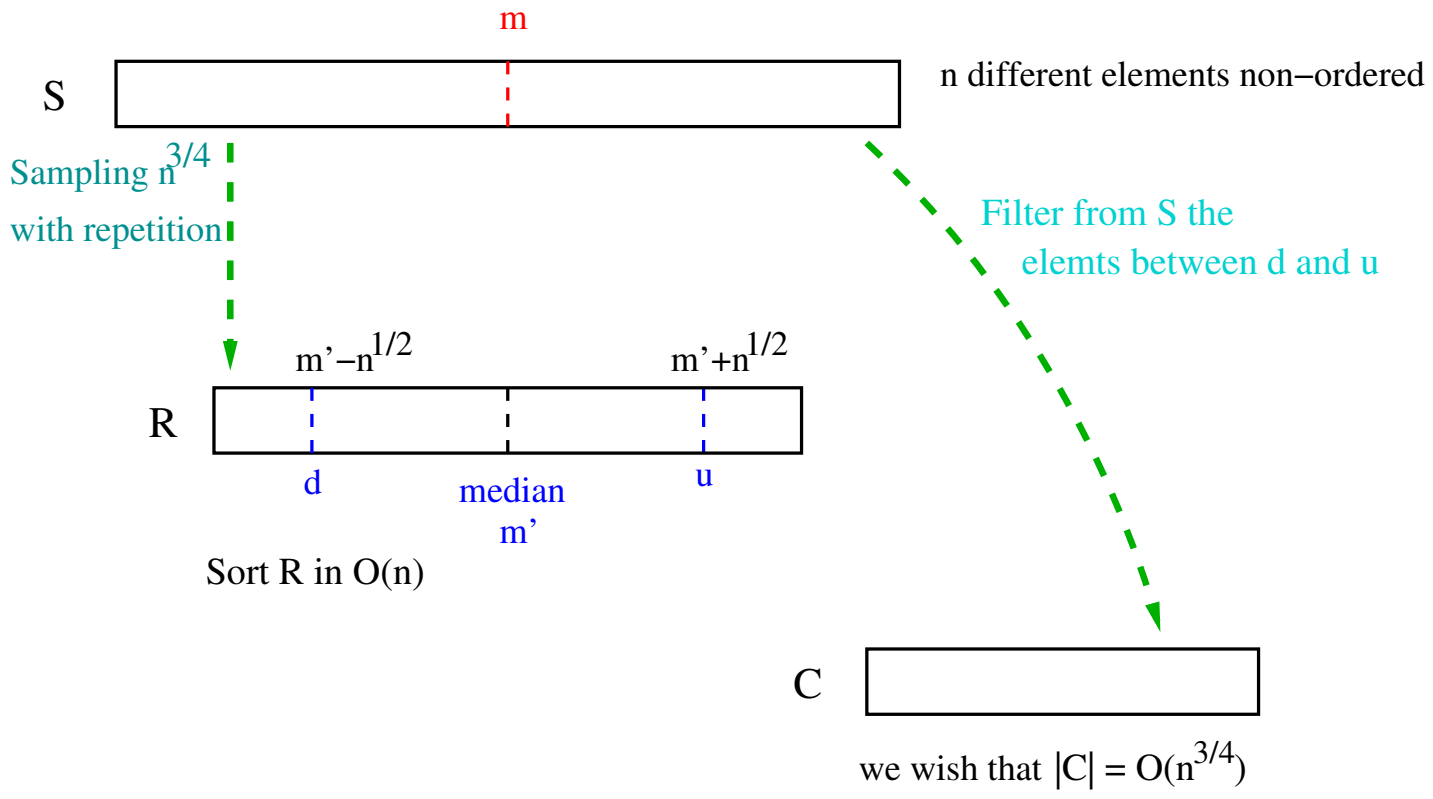
We will prove that whp the algorithm finds the median m of S, in linear time.

Outline of the algorithm

1. Let $\tilde{S}$ be the ordered set S (we do not know $\tilde{S}$). Let $m = \tilde{S}[k+1]$ be its median.
2. Find elements $d, u \in S$ s.t. $d < m < u$ and such that distance between d and u in $\tilde{S}$ is $< n/\log n$, i.e., $\text{rank}(u) - \text{rank}(d) < n/\log n$.
3. To find d and u sample with replacement S to get a multiset R, with $|R| = \mathcal{O}(\lceil n^{3/4} \rceil)$. Notice $\lceil n^{3/4} \rceil < n/\log n$. Find $u, d \in R$ s.t. m will be close to median in S.
4. Filter out the elements $x \in S$, which are $< d$ or $> u$ to form a set $C = \{x \in S | d \leqslant x \leqslant u\}$.
5. Sort elements in C in $\mathcal{O}(n)$ and find its median. This will be the algorithm's output—but if the median of C cannot be the median of S or C is too large the output of the algorithm will be **fail**.

We will prove that the algorithm is correct when it returns an element and that it will return an element (not fail) w.h.p.

Outline of the algorithm



Things that can be wrong:

C too large,

$m \notin C$,

$m \in C$ but no the median in C.

Randomized Median algorithm

1. Sample $\lceil n^{3/4} \rceil$ elements from S, u.a.r., independently, and with replacement. Call R the sample.
2. Sort R in $\mathcal{O}(n)$ steps.
3. Set d the $\lfloor (\frac{1}{2}n^{3/4} - \sqrt{n}) \rfloor$-th smallest element in R
4. Set u the $\lfloor (\frac{1}{2}n^{3/4} + \sqrt{n}) \rfloor$-th smallest element in R
5. Compute $C = \{x \in S | d \leqslant x \leqslant u\}$, $l_d = |\{x \in S | x < d\}|$ and $l_u = |\{x \in S | x > u\}|$ (cost $= \Theta(n)$).
6. If $l_d > \frac{n}{2}$ or $l_u > \frac{n}{2}$ output fail ($m \notin C$)
7. If $|C| \leqslant 4n^{3/4}$, sort C, otherwise output fail.
8. Output the $(\lfloor \frac{n}{2} \rfloor - l_d + 1)$-smallest element in sorted C, that should be the median m.

Complexity and correctness of the Randomized Median algorithm

Theorem

The Randomized Median algorithm terminates in $\mathcal{O}(n)$ *steps. If the algorithm does not output* ***fail****, then it outputs the median* m *of* S*.*

Proof

Asymptotically $n^{3/4} \log(n^{3/4}) = o(n/\log n)$, using Mergesort on R takes $\mathcal{O}(\frac{n}{\log n} \log(\frac{n}{\log n})) = \mathcal{O}(n)$.
The only incorrect answer is that it outputs an item, but $m \notin C$, but if so, it would fail in step 6, as either $l_d > n/2$ or $l_u > n/2$. □

Bounding the probability of failing

Theorem

The Randomized Median algorithm finds the median m with probability $\geqslant 1 - \frac{1}{n^{1/4}}$, i.e., whp.

Proof (Highlights)

In what follows, for simplicity, we will assume that all n elements in S are distinct, that n is odd, and that both $n^{3/4}$ and $\sqrt{n}$ are integers.
Consider the following 3 events:

- $E_1 =$ "$d > m$" $\equiv l_d > n/2$
- $E_2 =$ "$u < m$" $\equiv l_u > n/2$
- $E_3 =$ "$|C| > 4n^{3/4}$"

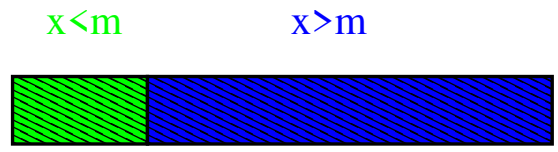
The algorithm outputs **fail** iff one of the three events above occurs.

$$\mathbb{P}[\textbf{fail}] = \mathbb{P}[E_1 \cup E_2 \cup E_3] \leqslant \mathbb{P}[E_1] + \mathbb{P}[E_2] + \mathbb{P}[E_3]$$

Bounding $\mathbb{P}[E_1]$

Proof (cont'd)

Consider R ordered, where R is obtained by sampling $n^{3/4}$ elements from S



Recall: d is the $(\frac{n^{3/4}}{2} - \sqrt{n})$-th element

- $d > m$, when the green block has size $< n^{3/4}/2 - \sqrt{n}$; this happens iff $l_d > n/2$
- Let $Y = |\{x \in R \,|\, x \leqslant m\}|$, then $\mathbb{P}[E_1] = \mathbb{P}\left[Y < n^{3/4}/2 - \sqrt{n}\right]$.
- For $1 \leqslant j \leqslant n^{3/4}$, define $Y_j = 1$ iff the value in the j-th position in R is $\leqslant m$.
- Then $Y = \sum_{j=1}^{n^{3/4}} Y_j$, moreover as the sampling is with replacement, then each Y_j is independent.

As $m =$ median of S ($|S| = n$), then we have $\frac{(n-1)}{2} + 1$ elements in S that are $\leqslant m$.

Bounding $\mathbb{P}[E_1]$

Proof (cont'd)

- $\mathbb{P}\big[Y_j = 1\big] = \frac{(n-2)/2+1}{n} = \frac{1}{2} + \frac{1}{2n}$, as there are $(n-1)/2+1$ elements $\leqslant m$.
- $Y \sim \mathsf{Bin}(n^{3/4}, \frac{1}{2} + \frac{1}{2n})$.
- $\mathbb{E}[Y] = \frac{n^{3/4}}{2} + \frac{1}{2n^{1/4}} \geqslant \frac{1}{2}n^{3/4}$,
- $\mathbb{V}[Y] = n^{3/4}(\frac{1}{2} + \frac{1}{2n})(\frac{1}{2} - \frac{1}{2n}) \leqslant \frac{n^{3/4}}{4}$.

Using Chebyshev's inequality:

$$\begin{aligned}\mathbb{P}[E_1] &= \mathbb{P}\left[Y < \frac{n^{3/4}}{2} - \sqrt{n}\right] \\ &\leqslant \mathbb{P}\big[|Y - \mathbb{E}[Y]| \geqslant \sqrt{n}\big] \leqslant \frac{\mathbb{V}[Y]}{(\sqrt{n})^2} = \frac{1}{4n^{1/4}}\end{aligned}$$

Bounding $\mathbb{P}[E_2]$

Proof (cont'd)

In the same way as for E_1, it holds $\mathbb{P}[E_2] \leqslant \frac{1}{4n^{1/4}}$

Bounding $\mathbb{P}[E_3]$

Proof (cont'd)

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

C is obtained directly from S by filtering, using the values d and u obtained in R.

For C to have $> \quad 4n^{3/4}$ elements, either of the following events must happen:

1. $E_{3,1}$ = "At least $2n^{3/4}$ items in C are $> m$"
2. $E_{3,2}$ = "At least $2n^{3/4}$ items in C are $< m$"

Then

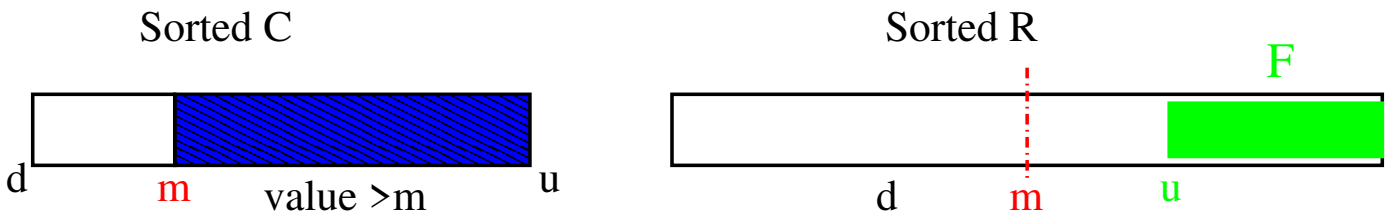
$$\mathbb{P}[E_3] \leqslant \mathbb{P}[E_{3,1} \cup E_{3,2}] \leqslant \mathbb{P}[E_{3,1}] + \mathbb{P}[E_{3,2}] \,.$$

Bounding $\mathbb{P}[E_{3,1}]$

Proof (cont'd)

Event $E_{3,1}$ happens when there are at least $2n^{3/4}$ elements in C which are $> m$.
If so, rank(u) in $\tilde{S}$ is $\geqslant n/2 + 2n^{3/4}$.

Let $F = \{x \in R \mid x > u\}$. Then $|F| \geqslant n^{3/4}/2 - \sqrt{n}$ and any element in F has rank $\geqslant n/2 + 2n^{3/4}$ if $E_{3,1}$ is true.



Bounding $\mathbb{P}[E_{3,1}]$

Proof (cont'd)

- Let X be # of selected items in R that have rank $\geqslant n/2 + 2n^{3/4}$ (i.e., in F)
- Then $\mathbb{P}[E_{3,1}] \leqslant \mathbb{P}\left[X \geqslant n^{3/4}/2 - \sqrt{n}\right]$.
- For $1 \leqslant j \leqslant n^{3/4}$, define $X_j = 1$ iff the j-th item in R is in F. The probability that any element from S is selected is $n^{3/4}/n = n^{-1/4}$.
- Note $X = \sum_{j=1}^{n^{3/4}} X_j$ and $\mathbb{P}\left[X_j = 1\right] = \frac{\frac{n}{2} - 2n^{3/4}}{n} = \frac{1}{2} - \frac{2}{n^{1/4}}$.
- So $\mathbb{E}[X] = \frac{n^{3/4}}{2} - 2\sqrt{n}$ and $\mathbb{V}[X] \leqslant n^{3/4}/4$

$$\mathbb{P}[E_{3,1}] \leqslant \mathbb{P}\left[X \geqslant \frac{n^{3/4}}{2} - n^{1/2}\right] \leqslant \mathbb{P}\left[X \geqslant \mathbb{E}[X] + n^{1/2}\right]$$

$$\leqslant \mathbb{P}\left[|X - \mathbb{E}[X]| \geqslant n^{1/2}\right] \leqslant \frac{\mathbb{V}[X]}{n} < \frac{n^{3/4}/4}{n} = \frac{1}{4n^{1/4}}.$$

Bounding $\mathbb{P}[E_{3,2}]$ and finishing the proof

Proof (cont'd)

In the same way we can compute $\mathbb{P}[E_{3,2}] = \mathcal{O}(\frac{1}{n^{1/4}})$
To end the whole proof, we also proved that

$$\mathbb{P}[E_3] \leqslant \mathbb{P}[E_{3,1}] + \mathbb{P}[E_{3,2}] \leqslant \frac{1}{2n^{1/4}}$$

Hence

$$\mathbb{P}[\text{algorithm fails}] = \mathbb{P}[E_1 \cup E_2 \cup E_3] \leqslant^{\text{Union Bound}} \frac{1}{n^{1/4}}$$

Finally

$$\mathbb{P}[\text{algorithm succeeds}] = 1 - \mathbb{P}[\text{algorithm fails}] \geqslant 1 - \frac{1}{n^{1/4}}$$

□