Sampling-based Median
Finding the median of $n$ things

• You may have seen an $O(n)$ time algorithm in CS161.
  • It was pretty complicated.

• Today: a simpler randomized algorithm (with optimal constant in the big-Oh runtime)!
Array $S$ of $n$ distinct numbers:

<table>
<thead>
<tr>
<th>9</th>
<th>5</th>
<th>34</th>
<th>1</th>
<th>2</th>
<th>33</th>
<th>12</th>
<th>4</th>
<th>15</th>
<th>3</th>
<th>6</th>
<th>8</th>
<th>10</th>
<th>18</th>
<th>0</th>
</tr>
</thead>
</table>

$n = 15$

Choose a set $R$ of size $n^{3/4}$ by drawing that many things uniformly at random, independently.

Sort $R$:

| 5 | 12 | 15 | 5 | 10 | 3 | 33 |

Find all the things in $S$ between $a$ and $b$ (time $O(n)$), to form a list $T$:

| 9 | 5 | 12 | 15 | 6 | 8 | 10 |

If $|T| < 4n^{3/4}$, sort $T$: (otherwise output FAIL)

| 5 | 6 | 8 | 9 | 10 | 12 | 15 |

- We can see in time $O(n)$ that there are 5 things in $S$ less than $a$, and 3 things in $S$ larger than $b$.

- The median is the 8′th smallest thing in $S$, which is the $8 - 5 = 3′rd$ smallest thing in $T$.

- Return 8

If this calculation shows that the median is not in $T$, output FAIL.
Group work...
Solutions to group work

2. Suppose that:
   • With probability at least 0.9, the median of $S$ is in $T$.
   • With probability at least 0.9, $|T| < 4t$.
   • Then the algorithm returns the correct answer with probability at least 0.8.
Choose a set $R$ of size $n^{3/4}$ by drawing that many things uniformly at random, independently.

Sort $R$:

Find all the things in $S$ between $a$ and $b$ (time $O(n)$), to form a list $T$:

If $|T| < 4n^{3/4}$, sort $T$: (otherwise output FAIL)

- We can see in time $O(n)$ that there are 5 things in $S$ less than $a$, and 3 things in $S$ larger than $b$.
- The median is the 8’th smallest thing in $S$, which is the $8 - 5 = 3$’rd smallest thing in $T$.
- Return 8. If this calculation shows that the median is not in $T$, output FAIL.
Solutions to group work

2. Suppose that:
   • With probability at least 0.9, the median of $S$ is in $T$.
   • With probability at least 0.9, $|T| < 4t$.
   • Then the algorithm returns the correct answer with probability 0.8.

   • If both events happen, then the algorithm never returns FAIL.
   • If it doesn’t return FAIL, then it returns the right answer by construction.
Solutions to group work

3. The running time is $O(n)$ operations.....actually $(3/2)n + o(n)$
Choose a set $R$ of size $n^{3/4}$ by drawing that many things uniformly at random, independently.

Sort $R$:

$O\left(\frac{3}{n^4}\log\left(\frac{3}{n^4}\right)\right) = o(n)$ operations.

Find all the things in $S$ between $a$ and $b$ (time $(3/2)n$), to form a list:

$O(n)$

If $|T| < 4n^{3/4}$, sort $T$: (otherwise output FAIL)

$O\left(\frac{3}{n^4}\log\left(\frac{3}{n^4}\right)\right) = o(n)$ operations.

- We can see in time $O(n)$ that there are 5 things in $S$ less than $a$, and 3 things in $S$ larger than $b$.

- The median is the 8’th smallest thing in $S$, which is the $8 - 5 = 3’rd$ smallest thing in $T$.

- Return 8

If this calculation shows that the median is not in $T$, output FAIL.
Solutions to group work

• Question 4: want to show that $\text{median}(S) \in T$ w.h.p.
Solutions to group work

4a. Consider two events:

\[ |\{r_i \in R : r_i < m\}| < \frac{t}{2} + \sqrt{n} \]

\[ |\{r_i \in R : r_i > m\}| < \frac{t}{2} + \sqrt{n} \]

Sorted version of S:

\[ \begin{array}{cccccccc}
0 & 1 & 2 & 3 & 4 & 5 & 6 & 8 \\
\end{array} \]

\[ m \]

\[ \Rightarrow b \geq m \]

Sorted version of R:

\[ \begin{array}{cccccccc}
3 & 5 & 5 & 10 & 12 & 15 & 33 \\
\end{array} \]

\[ a \quad \sqrt{n} \quad \sqrt{n} \quad b \]

\[ \text{median}(R) \]

\[ \frac{t}{2} + \sqrt{n} \text{'th smallest thing in } R. \]
Solutions to group work

4a. Consider two events:

\[ |\{r_i \in R : r_i < m\}| < \frac{t}{2} + \sqrt{n} \]
\[ |\{r_i \in R : r_i > m\}| < \frac{t}{2} + \sqrt{n} \]

\[ \Rightarrow b \geq m \]
\[ \Rightarrow a \leq m \]

• Then \( a \leq m \leq b \), aka \( m \in T \)
Solutions to group work

4b. Let $X = |\{ r_i \in R : r_i < m \}|$

- Then $X = \sum_i X_i$ where $X_i = 1$ iff $r_i < m$ and 0 otherwise, for $i = 1, \ldots, t$
- $\mathbf{E}[X_i] = \Pr[r_i < m] \leq \frac{1}{2}$
- $\text{Var}[X_i] \leq \frac{1}{4}$

- $\Pr\left[ \sum_i X_i \geq \frac{t}{2} + \sqrt{n} \right] \leq \Pr[\sum_i (X_i - \mathbf{E}X_i) \geq \sqrt{n}] \leq \frac{t/4}{n} = \frac{1}{4n^{1/4}} = o(1)$
Solutions to group work

4c. Consider two events:

\[ |\{ r_i \in R : r_i < m \}| < \frac{t}{2} + \sqrt{n} \quad |\{ r_i \in R : r_i > m \}| < \frac{t}{2} + \sqrt{n} \]

\[ \Rightarrow b \geq m \quad \Rightarrow a \leq m \]

• Then \( a \leq m \leq b \), aka \( m \in T \)

Both have probability at least \( 1 - O(n^{-1/4}) \)

\[ \Pr[m \in T] \geq 1 - O(n^{-1/4}) \]
Solutions to group work

• Question 5: want to show that $|T| < 4t$ w.h.p.
Solutions to group work

• 5(a). Say that $a$ is not one of the $\frac{n}{2} - 2t$ smallest elements of $S$

• Say that $b$ is not one of the $\frac{n}{2} - 2t$ largest elements of $S$

• Then $|T| < 4t$
Solutions to group work

• 5(b) Let $Y_i = 1$ iff $r_i$ is in the $\frac{n}{2} - 2t$ smallest elements of $S$, 0 else

\[ \sum Y_i = \text{number of things in } R \text{ down here.} \]

\[ \sum Y_i \geq \frac{t}{2} - \sqrt{n} \iff a \text{ is among the } \frac{n}{2} - 2t \text{ smallest elements of } S \]
Solutions to group work

• 5(b) Let $Y_i = 1$ iff $r_i$ is in the $\frac{n}{2} - 2t$ smallest elements of $S$, 0 else.
  
  • $EY_i = \frac{1}{2} - \frac{2t}{n} = \frac{1}{2} - \frac{2}{n^{1/4}}$

• $\Pr \left[ \sum_i Y_i \geq \frac{t}{2} - \sqrt{n} \right] \leq \Pr \left[ \sum_i (Y_i - EY_i) \geq \frac{2t}{n^{1/4}} - \sqrt{n} \right]$
  
  • $= \Pr \left[ \sum_i (Y_i - EY_i) \geq \sqrt{n} \right]$
  
  • $\leq \frac{\text{Var}[\sum_i Y_i]}{n}$
  
  • $\leq \frac{t}{4n} = \frac{1}{4n^{1/4}} = o(1)$

$\text{Var}[\sum_i Y_i] = \sum_i \text{Var}[Y_i] \leq \frac{t}{4}$. 
Solutions to group work

• 5(c). Say that $a$ is not one of the $\frac{n}{2} - 2t$ smallest elements of $S$

• Say that $b$ is not one of the $\frac{n}{2} - 2t$ largest elements of $S$

• Then $|T| < 4t$

$\Rightarrow \Pr[|T| < 4t] \geq 1 - O(n^{-1/4})$
All together:

• Question 2: To show that this algorithm works whp, it’s enough to show that:
  • whp, $\text{median}(S) \in T$
  • whp, $|T| < 4t$

• Question 4: whp, $\text{median}(S) \in T$
• Question 5: whp, $|T| < 4t$

• (And Question 2: it runs in time $O(n)$).