ACTIVITY 12, PART A: ANALYSIS OF MEDIAN ALGORITHM

The median m of a list of n numbers has $\lfloor n/2 \rfloor$ of the elements smaller or equal to m. If the elements are distinct then the median is the $\lceil n/2 \rceil^{th}$ element in the sorted list. So you can always find it in $O(n \log n)$ time by sorting, but in fact you can do better. There is a complicated deterministic algorithm which finds it in $O(n)$ time but we can get the same end result in a less complicated way via a randomized algorithm.

To make the analysis simpler, assume n odd and all elements distinct.

Suppose S is a list of n distinct numbers with n odd. The **median** of S is the element $m \in S$ such that $(n-1)/2$ of the elements are less than m and $(n-1)/2$ of the elements are greater than m. Obviously, one way to find m is to sort S and then select the middle element. But that takes $O(n \ln n)$ time and we can find m with high probability in linear time if we use randomness. Our goal is to analyze the following randomized algorithm to find m, which works via **sampling**:

AlgorithmFM. Input: The list S
Step1: Pick $\lceil n^{3/4} \rceil$ elements of S independently and uniformly at random with replacement.
Step2: Sort that collection of elements and let R be the result.
Step3. Let d be the $\lfloor .5n^{3/4} - \sqrt{n} \rfloor^{th}$ smallest element in R.
Step4. Let u be the $\lfloor .5n^{3/4} + \sqrt{n} \rfloor^{th}$ smallest element in the sorted set R.
Step5. By comparing every element in S to d and u. compute the set $C = \{x \in S : d \leq x \leq u\}$ and the numbers $\ell_d = |\{x \in S : x < d\}|$ and $\ell_u = |\{x \in S : x > u\}|$.
Step6. If $\ell_d > n/2$ or $\ell_u > n/2$ then FAIL.
Step7. If $|C| \leq 4n^{3/4}$ then sort C. Otherwise FAIL.
Step8. Output the $(\lfloor n/2 \rfloor - \ell_d + 1)^{st}$ element in the sorted order of C.

Question 1. *Recalling that our goal is an algorithm which runs in linear time, what is the purpose of Step7? Argue why Step7 fulfills this purpose.*

Date: September 23, 2016.

Question 2. *Why are d and u chosen the way they are? (This will be easier to answer after section 2)*

Question 3. *What is the total running time in Big Oh notation? Hint: compute the running times for Step 1, Step 2, and Step 5, and note that* $\log(n^{3/4}) < n^{1/4}$ *asymptotically.*

Question 4. *BONUS: What is the expected number of duplicate elements in R in step 1? Which probability problem is this reminiscent of?*

The algorithm either outputs FAIL or the $(\lfloor n/2 \rfloor - \ell_d + 1)^{st}$ element in the sorted order of C. You may assume $\sqrt{n}$ and $n^{3/4}$ are integers.

Question 5. *Argue that this element is the median of S. It might help to draw a picture of the sorted versions of C and S on a line.*

In order to compute the probability that the algorithm fails, consider the following two events:

$E : m \notin C$
$F : |C| > 4n^{3/4}$

Question 6. *Prove the following claim: the algorithm outputs FAIL if and only if one of these events occurs.*

Observe that E occurs if and only if EITHER:
(1) $u < m$ OR
(2) $d > m$

For each element x chosen to go into R, what is the probability that $x \leq m$?

Let X be the expected number of elements $x \in R$ (henceforward called samples) with $x \leq m$. Write X as a sum of independent Bernoulli random variables.

What is the Variance of X? Why?

Apply Chebyshev's inequality to give an upper bound on

$P(|X - n^{3/4}/2| \geq \sqrt{n})$

Use this to compute a bound on $P(E)$.

In order to bound $P(F)$ we need to show that it is likely that d is greater than the $(n/2 - 2n^{3/4})^{th}$ smallest element in S and that u is less than the $(n/2 + 2n^{3/4})^{th}$ element in S. Let X_d be the number of samples in R which are among the $(n/2 - 2n^{3/4})$ largest elements in S. Write $X_d = \sum_1^{|R|} X_i$ where X_i is 1 if the i^{th} sample is among the $(n/2 - 2n^{3/4})$ largest elements in S.

$$E[X_d] = (.5 - 2n^{-1/4})n^{3/4} \text{ and } Var[X_d] \leq .25n^{1/4}$$

Observe that $|X_d - .5n^{3/4} + 2\sqrt{n}| \geq |X_d - .5n^{3/4}$ so Chebyshev's Inequality says:

$$P(|X - .5n^{3/4}| \geq \sqrt{n}) \leq P(|X_d - E[X]| \geq \sqrt{n}) \leq .25n^{-1/4}$$

The probability p_1 that at least $2n^{3/4}$ elements of C are greater than the median can be computed as the probability that the upper bound R_r is greater than the $(n/2 + 2n^{3/4})^{th}$ element in the sorted version of S. This probability is equal to the probability that $X_d \geq n/2 \cdot n^{-1/4} - \sqrt{n}$. Hence, it's less than $n^{-1/4}/4$ as required. Morally, what's going on is that all ways for the X_d event to occur are contained in the event regarding R_r (which is happening in C).

By symmetry, this same bound applies to the probability p_2 that at least $2n^{3/4}$ elements of C are smaller than the median. Use this to bound

$P(F) \leq p_1 + p_2 \leq .5n^{-1/4}$

Finally, combine this with your bound on $P(E)$ to prove that the probability that the algorithm fails is $\leq n^{-1/4}$.