Scribe Notes From CSE203 – Class on Thursday, 1/15 Scribe: Aaron Arpi, aarpi@eng.ucsd.edu

Finding the median

Say that there is a set of n numbers S = {a_1, a_2,…, a_n}, and that the numbers are not necessarily ordered. How can we find the median in O(n) time?

The median can be defined as the number $a_i \in$ S for which the following conditions hold:

$$|\{x \leq a_i : x \ \in S\}| \geq floor\left(\frac{n}{2}\right)$$

$$|\{x \leq a_i : x \ \in S\}| \geq floor\left(\frac{n}{2}\right)$$

A simple deterministic algorithm involves an O(n log n) sort operation, followed by selection of the middle element in the sorted list. But we would prefer O(n) rather than O(n log n).

There is a complex O(n) deterministic algorithm that involves:

1)partitioning S into multiple rows having 5 numbers each.

2)sorting each row

3)selecting the middle number in each sorted row.

4) sorting the rows so as to order the rows in order of their middle numbers.

5)eliminating portions of these rows.

6) repeating steps 1-5 until you have arrived at the median

But we will evaluate a randomized algorithm that will either find the median correctly in O(n) time, or output “fail” in O(n) time. We will need to evaluate the probability that it will output fail.

The algorithm works as follows:

1) form a set R by making $n^{3/4}$ random selections of numbers from S, and do this selection with replacement. Obviously, the cardinality of R will be $n^{3/4}$.
2) Sort the set R. This takes O(n) time. Note that $n^{3/4}$ln(3/4) = O(n) because every ln(n)=O(n^x) for every positive x. Even the smallest polynomials grow faster than the natural log.
3) Say that b=floor(.5$n^{3/4}$ – $n^{1/2}$) and c=ceiling(.5$n^{3/4}$ + $n^{1/2}$). Identify the number that is the b^{th} smallest in ordered set R, and call this number d. Identify the number that is the c^{th} smallest in ordered set, and call this number u.
4) Define three subsets of S. Call the first subset l_d and define this subset to hold those numbers of S that are less than or equal to d. Call the second subset C, and define this

subset to hold those numbers of S that are between d and u. Call the third subset l_u. Define l_u to hold those numbers of S that are greater than or equal to u.

5) Based on the values of d and u, determine $f=|l_d|$ and $g=|l_u|$. If either f or g is greater than n/2, output "FAIL".
6) If no failure at 6), then determine |C|. If |C| is greater than $4n^{3/4}$, output "FAIL". Otherwise, sort set C. Let i=floor(.5n-f). Find the ith smallest number in C. This number is the median.

There are 3 events that will cause a "FAIL" to be returned, otherwise the algorithm returns the correct result. The 3 events are:

a) The set R is selected such that $|C| > 4n^{3/4}$. Under this condition the sorting done at step 6) could not be done in linear time. Instead the sorting is skipped, the algorithm terminates in FAIL so that the algorithm does not exceed linear time.
b) l_d>n/2. In this case, the median lies in l_d and not in C.
c) l_u>n/2. In this case, the median lies in l_u and not in C.

Determine the probability of event b) occurring first:

Define S_2 s.t. $S_2 \subset$ S, $S_2 = \{a_i > m : a_i \in \text{S}\}$

Event b) occurs iff, of the $n^{3/4}$ numbers in R randomly selected from S, at least $(.5n^{3/4}+n^{1/2})$ of them were selected from S_2.

Let X_i be a binary random variable that is 1 if the ith number selected into R is in S_2, and is 0 otherwise. $P(X_i =1) = ½$ for all i.

Define a random variable $Y = \sum_i X_i$

Each X_i is a Bernoulli Trial, so Y is a binomial random variable,

The binomial distribution of random variable Y has parameters $n_B=n^{3/4}$, and p=1/2. ***Note I am using n_B (the number of Bournoulli trials, normally designated only as n) to differentiate from the use of n to represent the cardinality of S.

The expectation of a binomial random variable is $n_B p$. So E(Y) = $.5n^{3/4}$

The variance of a binomial random variable is $n_B p(1\text{-}p)$. So $V(Y)=n^{3/4}(.5)(1\text{-}.5)=(1/4)n^{3/4}$

Chebyshev's inequality states that P(Y-E(Y)>=a) = VAR(Y)/a^2. We are interested in those cases when $Y>(.5n^{3/4}+n^{1/2})$ – in other words, when $Y\text{-}E(Y)>=n^{1/2}$.

So, by applying Chebyshev's inequality $P(Y\text{-}E(Y)>n^{1/2}) = (.25n^{3/4})/[(n^{1/2})^2]=.25n^{-1/4}$. Thus, the probability that the algorithm fails due to occurrence of event b) is $.25n^{-1/4}$.

Event c) is clearly disjoint from event b). By considering event c), it is easy to see that the probability of this event occurring is exactly the same as P(event b), and this can be shown using the same analysis as was applied to analyze P(event b) above.

Thus P(b U c) = $.25n^{-1/4}+.25n^{-1/4} = .5n^{-1/4}$.

Analysis of the Probability of Event a)

Event a) occurs when $|C|>4n^{3/4}$. For this to occur, exactly one of the following occurs:

a) $2n^{3/4}$ or more numbers in C are greater than the median of set S;
b) $2n^{3/4}$ or more numbers in C are less than the median of set S.

Let's assume that $2n^{3/4}$ or more numbers in C are greater than the median, and analyze the probability of. This means that at least $.5n^{3/4} – n^{1/2}$ numbers in R are ranked more than $2n^{3/4}$ positions below the median in set S. That is $.5n^{3/4} – n^{1/2}$ numbers in R must be selected from amongst the largest $.5n-2n^{3/4}$ numbers in set S. From this point on, S_{top} will be used to denote the subset of S that includes exactly the largest $.5n-2n^{3/4}$ numbers of S.

Let X be a random variable that is equal to the number of elements in R that are also in S_{top}. Moreover, let X_i be an indicator variable that is equal to 1 if the ith number of R is in S_{top}, and is equal to 0 otherwise.

Because R is formed by selecting numbers from S with replacement, $P(X_i = 1) = (.5n-2n^{3/4})/n = .5-$
$2n^{-1/4}$
for each number i in R.

$$X = \sum_{i=1}^{n^{3/4}} Xi$$

and

$$E(X) = \sum_{i=1}^{n^{3/4}} E(Xi)$$

$E(X_i)=P(X_i=1)*1+P(X_i=0)*0=.5-2n^{-1/4}$ for each i. Thus $E(X) = n^{3/4} (.5-2n^{-1/4})=.5n^{3/4}-2n^{1/2}$

X is a binomial random variable with each X_i being a Bernoulli trial with $p=.5-n^{-1/4}$. There are $n^{3/4}$ such trails so, as before we will say that $n_B=n^{3/4}$. Recall that the variance of a Bernoulli random variable is $n_B p(1-p)$. So $V(X)=n^{3/4}(.5-2n^{-1/4})(.5+2n^{-1/4})$. Now, apply Chebyshev's Theorem to determine the probability that X is greater than $.5n^{3/4} – n^{1/2}$:

P(X-E(X)>=a) = VAR(X)/a

Let X-E(X) = Δx, so we are interested in the probability of $\Delta x > [.5n^{3/4} – n^{1/2}]-[.5n^{3/4}-2n^{1/2}] = n^{1/2}$.

$$P(\Delta x>n^{1/2})= \frac{n^{\frac{3}{4}}\left(.5-2n^{-\frac{1}{4}}\right)\left(.5+2n^{-\frac{1}{4}}\right)}{n} = .5n^{-1/4} = P(\text{event a})$$

Total Failure Probability of the algorithm

By the union bound the probability of failure cannot be greater than the sum P(event a)) + P(b U c)

Thus, the overall probability of failure is less than or equal to $n^{-1/4}$.