

**Randomized Algorithms**  
**Prof. Benny George Kenkireth**  
**Department of Computer Science & Engineering**  
**Indian Institute of Technology, Guwahati**

**Lecture – 09**  
**Median Algorithm**

Hello. So, today we will see the algorithm for finding the Median. In more details, we discussed the algorithm in the previous class. We will fill in the details and do the analysis of the algorithm today.

(Refer Slide Time: 00:47)

Randomized Median Find

Input :  $S = \{a_1, a_2, \dots, a_n\}$   
Output : Median of  $S$

Algorithm.

- ① Randomly choose  $n^{3/4}$  elements of  $S$  to form the multiset  $R$ .
- ② Sort  $R$  and determine 'd' & 'u'

$d \triangleq \left(\frac{1}{2} n^{3/4} - \sqrt{n}\right)^{\text{th}}$  element of  $R$ .  
 $u \triangleq \left(\frac{1}{2} n^{3/4} + \sqrt{n}\right)^{\text{th}}$  element of  $R$ .

So, in this problem the input is a set  $S$  of numbers which we call as  $a_1, a_2, \dots, a_n$ . We may assume that  $n$  is an odd number. So, that the median is the  $(n+1)/2$ th element and the output will be the median of  $S$  ok. So, the our algorithm the basic steps are as follows: first step is randomly choose  $n^{3/4}$  elements of  $S$  to form the multi set, let us say  $R$  ok. We could have just chosen and raised  $3/4$  elements without replacement, but here, we do it with replacement because that makes the analysis simpler. The second step is sort  $R$  and determine two numbers  $d$  and  $u$ . So,  $d$  is the  $(\frac{1}{2} n^{3/4} - \sqrt{n})$ th element of  $R$  and  $u$  is the  $(\frac{1}{2} n^{3/4} + \sqrt{n})$ th element of  $R$  ok.

So, we have randomly picked the set which we call as R, we sort the elements of R and if this was our collection R, the middle element let us say that is the median of R that is the  $n$  by  $n$  raised the  $n$  raised to 3 by 4 elements. So, this the  $n$  raised to 3 by 4 by two'th earth element. We go left side by a distance root  $n$  and we go right to a distance root  $n$  and pick the elements at these locations that will be our  $d$  and  $u$ . This is our set R. Now based on this  $d$  and  $u$ , we are going to split the entire array S into 3 parts. So, that will be our third step.

(Refer Slide Time: 04:07)

Handwritten notes on a blue grid background detailing the steps of a selection algorithm:

- ① Randomly choose  $n^{3/4}$  elements of  $S$  to form the multiset  $R$ .
- ② Sort  $R$  and determine ' $d$ ' & ' $u$ '  
 $d \triangleq \left(\frac{1}{2} n^{3/4} - n\right)^n$  element of  $R$ .  
 $u \triangleq \left(\frac{1}{2} n^{3/4} + n\right)^n$  " " " $R$ .  
 $O(n \log^{3/4} n)$
- ③ Split  $S$  into  $A$ ,  $B$  &  $C$   
 $A \triangleq$  elements of  $S$  which are smaller than  $d$ .  
 $B \triangleq$  " " " which are b/w  $d$  &  $u$ .  
 $C \triangleq$  " " " larger than  $u$ .  
 $O(n)$
- ④ If  $|A| > n/2$  or  $|C| > n/2$  or  $|B| > 4 n^{3/4}$  Quit.  
 $I_d = |A|$   
 $\left(\frac{n - I_d}{2}\right)^n$  at  $d$
- ⑤ Determine the median by examining  $B$ .

Diagram illustrating the partitioning of  $S$  into  $A$ ,  $B$ , and  $C$  based on  $d$  and  $u$ . The diagram shows a horizontal line representing  $S$  with points  $d$  and  $u$  marked. The region to the left of  $d$  is labeled  $A$ , the region between  $d$  and  $u$  is labeled  $B$ , and the region to the right of  $u$  is labeled  $C$ . The size of  $B$  is indicated as  $|B|$ .

Split  $S$  into  $A$ ,  $B$  and  $C$  where  $A$  is the elements of  $S$  which are smaller than  $d$  and  $C$  is a set of elements which are larger than  $u$  and  $B$  is the in between elements; elements of  $S$  which are between  $d$  and  $u$  ok.

So,  $C$  is the set of elements is that larger than  $u$ . So, we can think of those as let us say if this was our entire set  $S$ , these side would contain all the elements smaller than  $d$  and this so these the region would consist of all the elements which are greater than  $u$ . So, this is our point  $d$  and this is our point  $u$  and these are the in between elements that would be  $B$  ok. Now once this  $d$  and  $u$  are determined, what we will do is, we will run some checks to ensure that the median is in  $B$  and  $B$  is small enough. So, if size of  $A$  is greater than  $n$  by 2 or size of  $C$  is greater than  $n$  by 2 or size of  $B$  is greater than 4 times  $n$  raised to 3 by 4 ok. This is some the size of  $B$  is some parameter that we have chosen. So, that the algorithm works nicely and the analysis goes through smoothly ok.

So, if any of these condition happens, then we will say we will just quit ok; we will not will not output a median will just say that we failed. What we have to do later on is that we will have when we analyze the algorithm, we will show that the probability of such a thing happening is smart. So, if this is not the case; that means,  $A$  is not greater than  $n$  by 2 size of  $A$  is not greater than  $n$  by 2 and  $B$  is not greater than sorry  $C$  is not greater than  $n$  by 2 and  $B$  is smaller than  $4n$  raised to 3 by 4, this is the case, then what we will do is; we will find the median.

So, in this case the median will clearly lie in  $B$  because  $A$  contains less than  $n$  by 2 elements. So, if you look at  $A$  if you look at  $S$  in the sorted order the median will be right at the middle the elements which are smaller than  $d$  that will be less than  $n$  by 2, there are less than  $n$  by 2 elements of that kind and the elements which are larger than  $u$  they are also less than  $n$  by 2 ok.

So, this is less than  $n$  by 2, this is less than  $n$  by 2 ok. So, median lies in small region ok. So, then so based on the size of  $A$ , we can determine the median. How do we do that? So, suppose  $l$  is the size. So,  $l$  is equal to size of  $A$ , then if you find the  $n$  by 2 minus  $l$  d'th element of  $B$  that will be our median. So, already the elements in  $A$  are known to be smaller than median and they are all smaller than  $d$ . So,  $d$ 's position is  $l$ .

Now what we are interested in is the  $n$  by second element the middle element in the entire array. So, that can be determined by going  $n$  by 2 minus  $l$ . So, the element which is at that position in the array  $B$ . I will just write this as determine the median by examining  $B$ ; that means, go look in  $B$  find the  $n$  by 2 minus  $l$  d'th element that will be the median. Now how do we find the  $n$  by 2 minus  $l$  d'th element in  $B$ ? We could just sort  $B$ ;  $B$  is guaranteed to be a small set it contains at most  $4n$  raised to 3 by 4 elements. Therefore, we can completely sort the array and after sorting we will we can determine the element inside I mean the element at any position inside  $B$  ok. So, that is our median algorithm ok.

Now first let us see that this algorithm works in linear time. The first step of choosing  $n$  raised to 3 by 4 elements that can of course, be done in order  $n$  time and since  $R$  consists of only  $n$  raised to 3 by 4 elements, we can just use an  $n \log n$  sorting algorithm and since the total number of elements is bounded by is exactly equal to  $n$  raised to 3 by 4. This step will take only  $n$  raised to 3 by 4  $\log n$  raised to 3 by 4. So, that is going to be the time

taken by step 2 that is going to be less than order of  $n$ . And then the splitting of once you have determined  $d$  and  $u$  splitting of the entire array into 3 parts will take no more than linear time ok. So, each element after examining, we can determine for any element inside us whether it falls in A or B or C.

So, this can be done in order of  $n$  time and the size of this sets also can be determined in linear time. And once we have done that since B is guaranteed to be a small set, we can just sort the entire collection B and find the median of the complete set. So, the algorithm clearly runs in linear time ok.

(Refer Slide Time: 11:27)

Analysis:

① The algorithm runs in linear time  $O(n)$ .

What is the probability that this algorithm fails to produce an answer?

① $ A  > n/2$	$E_1$
② $ C  > n/2$	$E_2$
③ $ B  > 4n^{3/4}$	$E_3$

Claim  $\rightarrow P(E_1 \cup E_2 \cup E_3) < \frac{1}{n^{1/4}}$

$n = 10000$   
 90% probability of success  
 $\frac{1}{10} \times \frac{1}{10} \times \frac{1}{10} = \underline{\underline{.1\% \text{ error}}}$

So, first the algorithm runs in linear time, it takes no more than  $O(n)$  steps. We need to compute the error probability. So, what is the probability that this algorithm fails to produce an answer ok? So, this algorithm can fail if 3 conditions happen, if any of these 3 conditions happen. The first condition size of A greater than  $n$  by 2.

So, this we will call it as event 1. The second condition was size of B greater than  $n$  by 2 that is going to be event 2 and the third condition is sorry the second condition was size of C greater than  $n$  by 2 and the third condition is size of B is greater than  $4n$  raise to 3 by 4 and this we will call as event  $E_3$ . And what we are interested in is the probability of  $E_1 \cup E_2 \cup E_3$ . We will show that this is less than say something like  $1/n^{1/4}$ . So, this is what we will show in our analysis. If you can show this that

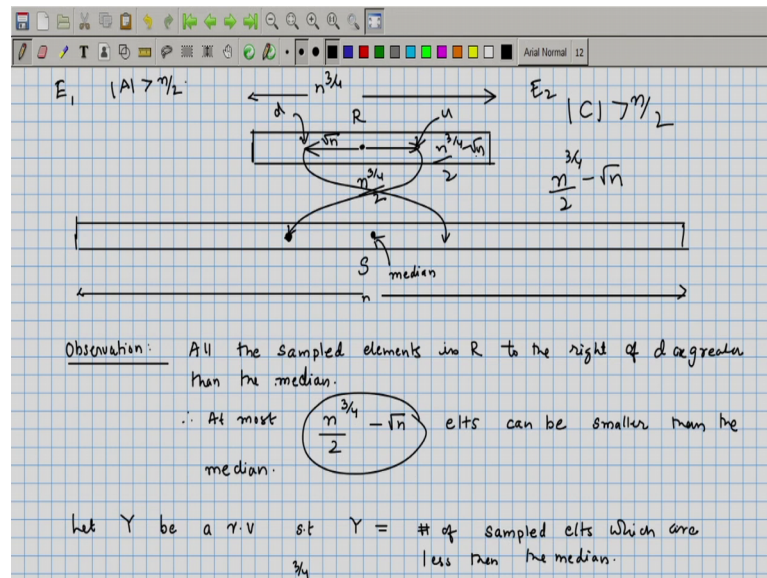


would mean that if  $n$  is say something like if you have to sort 10,000 numbers, there is at least a 90 percent guarantee that the algorithm will work correctly ok.

So, for  $n$  equals 1000 this gives 90 percent probability of success. Now observe that if you are given an element and you have to find the median and you have to check whether it does a median that can of course, be done in linear time ok. In any case, our algorithm always gives an answer which is a correct answer or it just fails and does not give any answer. So, we could just run the algorithm multiple times; let us say we have run the algorithm 3 times and if it gave an answer the median that it gave in all the 3 different trials will be same ok. There is of course, is small 10 percent chance that the algorithm will fail to produce an answer, but that happens with only 1 by 10 probability. So, if we do it 3 times, the probability of failure becomes no more than 1 by 10 into 1 by 10 into 1 by 10 ok.

So, that is going to be a 0.1 percent error ok. So, just by we had a linear time algorithm we just ran it 3 times and ensured that we have 99.9 percent accuracy ok. You can repeat it more times and get significantly higher. I mean whatever is the probability of success that you are interested in, you can get that by choosing the number of times you have to run the algorithm. Another way is to keep on running the algorithm till it finds an answer. The first time, it finds an answer; first time it runs without quitting that answer can be declared as the median. Now in this algorithm there is a small chance that the algorithm will continue running for a long time. We can analyze what is the expected running time of this modified algorithm and so that the modified algorithm also will run in linear time ok. All this is if we prove this particular claim that probability of error is less than 1 by  $n$  raise to 1 by 4. So, how do we? Let us see how do we prove that.

(Refer Slide Time: 16:39)



So, first let us look at this particular event  $E_1$  namely size of  $A$  greater than  $n$  by 2. The event  $E_2$  is  $A$  symmetric event so, the probability that we compute in case one can be directly used for the case 2. So, let us just draw the running of the algorithm in a different manner. So, this is our set  $R$ , this is a set  $R$  that we had randomly sampled and let us say this was our original set  $S$ . So,  $S$  had  $n$  elements and  $R$  had  $n$  raised to 3 by 4 elements.

So, let us imagine this  $S$  to be a I mean let us imagine that  $S$  has a sorted collection and draw the entire thing. So, from this collection we are drawing the set  $R$  without with replacement ok. So, let us say this is the middle position that is the  $n$  raised to 3 by 4 by 2'th position and if you go  $\sqrt{n}$  towards this direction, you will get your  $d$ . This is the position  $d$  and if you go  $\sqrt{n}$  towards this direction what you will get is your  $u$ . This  $d$  and  $u$  are elements of the original set where do they lie. If we had imagined  $S$  as a sorted collection; we want to figure out where does this  $d$  lie.

So, this was a middle portion of  $S$ ; if the this element here denotes the median if your  $d$ . so,  $d$  is a random element in the sense  $d$  is determined on the basis of the elements that we have sampled that is a random sample. So, once the random sample is fixed, your  $d$  has a particular value. So, now, we want to look at that particular random value  $d$  and figure out where would it be in the entire set  $S$ . If  $A$  is greater than  $n$  by 2, then this  $d$  is essentially falling inside the right side of this diagram ok. So, we want to determine what is the probability of that happening ok. So, if  $d$  is here, then what can we tell about all

these elements to the right of  $d$ . So, I will write it is an observation. All the sampled elements in  $R$  to the right of  $d$  are greater than the median ok. Therefore, at most  $n$  raised to  $3$  by  $4$  by  $2$  minus  $\sqrt{n}$  elements can be smaller than the median ok.

So, whenever the event  $E_1$  happens, the number of sampled elements which are smaller than the median is going to be strictly less than this quantity over here ok. I will repeat that whenever the event even occurs the number of elements that we had randomly sampled which are going to be smaller than the median is going to be at most  $n$  raised to  $3$  by  $4$  by  $2$  minus  $\sqrt{n}$  ok. We will show that that happens with very small probability ok. So, let us set things up.

(Refer Slide Time: 21:35)

Observation: All the sampled elements in  $R$  to the right of  $d$  are greater than the median.

$\therefore$  At most  $\frac{n^{3/4}}{2} - \sqrt{n}$  elts can be smaller than the median.

Let  $Y$  be a r.v. s.t.  $Y = \#$  of sampled elts which are less than the median.

$E[Y] \approx \frac{n^{3/4}}{2}$

When  $E_1$  occurs,  $Y < \frac{n^{3/4}}{2} - \sqrt{n}$

$P(E_1) < P\left(Y < \frac{n^{3/4}}{2} - \sqrt{n}\right)$

Let  $Y$  be a random variable with counts. So, let  $Y$  be a random variable such that  $Y$  is equal to number of sampled elements which are less than the median ok. We know that the expected value of  $Y$  should be approximately  $n$  by  $2$  sorry  $n$  raised to  $3$  by  $4$  by  $2$ . I am writing approximately because I do not want to bother about even and odd ok.

So, when you randomly sample an element the probability that the randomly sample element is less than the median is half, greater than the median is also half then approximately so there is a half plus  $1$  by  $n$  depending upon whether you include the median or not, but we can say that it is approximately  $1$  by  $2$ . So, the expected value of  $Y$  is going to be  $n$  raised to  $3$  by  $4$  by  $2$ . Now when the event  $E_1$  occurs what happens is the random variable  $Y$  has value much smaller than the than its expected value. It

deviates from the expected value ok. For the sample that we have chosen, the value of  $Y$  happens to be smaller than the expected value of  $Y$  by an amount  $\sqrt{n}$ .

So, we will write this down. When  $E_1$  occurs when the event  $E_1$  occurs  $Y$  is less than  $n$  raised to  $3/4$  by  $2$  minus  $\sqrt{n}$ . So, probability of  $E_1$  it is going to be less than the probability that  $Y$  is less than  $n$  raised to  $3/4$  by  $2$  minus  $\sqrt{n}$  and this probability we will determine.

So, this event  $Y$  less than  $n$  raised to  $3/4$  by  $2$  minus  $\sqrt{n}$  the probability of that event we will determine ok. So, how do we find that? So, you will use Chebyshev's inequality.

(Refer Slide Time: 24:31)

$$P(|X - E[X]| > a) \leq \frac{\text{Var}(X)}{a^2} \quad (\text{Chebyshev's Inequality})$$

$$P\left(Y > \frac{n^{3/4}}{2} - \sqrt{n}\right) < P(|Y - E[Y]| > \sqrt{n})$$

$$\leq \frac{\text{Var}(Y)}{n} = \frac{n^{3/4}}{4 \cdot n} = \frac{1}{4n^{1/4}}$$

$$Y = Y_1 + Y_2 + \dots + Y_k \quad (k = n^{3/4})$$

$$Y_i = 1 \text{ if the } i^{\text{th}} \text{ sample} < \text{median} \quad \frac{1}{2}$$

$$E[Y_i] = \frac{1}{2} \quad \text{Var}(Y) = \frac{k}{4} = \frac{n^{3/4}}{4}$$

$$\text{Var}[Y_i] = \frac{1}{4}$$

So, Chebyshev's equality says that probability that a random variable deviates from its mean. So,  $X$  minus expectation of  $X$  greater than  $A$  is going to be less than variance of  $X$  divided by a square ok. So, we need to look at this. So, we need to look at this particular event  $Y$  greater than  $n$  raised to  $3/4$  by  $2$  minus  $\sqrt{n}$  ok. So, this we can. So, probability of this is surely less than the probability that  $Y$  minus expectation of  $Y$  greater than  $\sqrt{n}$  ok. Whenever  $Y$  minus expectation of  $Y$  is greater than  $\sqrt{n}$  when whenever  $Y$  is greater than  $n$  raised  $3/4$  by  $2$  minus  $\sqrt{n}$ ,  $Y$  minus expectation of  $Y$  is going to be greater than  $\sqrt{n}$  ok.

So, this probability is going to be less than or equal to variance of  $Y$  divided by a square is  $\sqrt{n}$ . So, that will be  $n$ . So, now, we need to determine what is the variance of  $Y$ . So,

what again was our  $Y$ ?  $Y$  is the number of sampled elements which are less than median. So,  $Y$  we can write it as  $Y_1$  plus  $Y_2$  plus  $Y_k$  where  $k$  is the number of elements that we have sampled ok. So,  $k$  equals  $n$  raised to  $3/4$  ok. And these samples are independent samples because here is where the assumption that we had sampled with replacements come handy. So, when you sample with replacement all these  $Y_1, Y_2, Y_k$  are all independent of each other and probability that. So, what are these  $Y_i$ ?  $Y_i$ 's are indicator random variable  $Y_i$  equals 1 if the  $i$ 'th sample less than median.

Now, the probability of this happening will be half will be  $1/2$  by  $n$ . So, we will just assume it to be half that just I mean maybe there is a  $1/2$  by  $n$  term depending upon whether you picked median or not, but half is a reasonably good approximation. I mean it will I mean you can replace it with  $1/2$ , I mean you can correct for the middle element, but we will just do the calculation with half ok. So, this happens with probability  $Y_i$  equals 1 with probability  $1/2$  ok. So, expectation of  $Y_i$  equals half and variance of  $Y_i$  equals  $1/4$  ok. And in this case when you add up the random variables their variance is just gets added up. So, variance of  $Y$  is equal to number of samples that is  $k$  divided by 4.

So, basically it will be  $1/4$  plus  $1/4$  that will be up to  $k$  terms. So, that this  $k$  by 4 that is equal to  $n$  raised to  $3/4$  by 4 ok. So, we can substitute that here. So, variance of  $Y$  is going to be  $n$  raised to  $3/4$  divided by 4 divided by  $n$  that is going to be  $1/4$   $n$  raised to  $1/4$ . So, this is a probability of  $E_1$  where  $E_1$  was the event that the element  $d$  that we had found out by sampling is going to be greater than I mean it is going to lie towards the right of the median ok.

(Refer Slide Time: 29:47)

$$P(E_1) < \frac{n^{-1/4}}{L} \quad E_3 = |B| > 4n^{3/4}$$

$$P(E_3)$$

B

$4n^{3/4}$

$E_{31}$ :  $2n^{3/4}$  elements of B are greater than the median

$E_{32}$ :  $2n^{3/4}$  elements of B are less than the median

$$P(E_3) = P(E_{31} \cup E_{32}) < P(E_{31}) + P(E_{32}) = 2P(E_{31}) < \frac{n^{-1/4}}{2}$$

So, probability of that bad event happening we have bounded it by  $1/4n$ . Similarly  $E_2$  will also happen with so probability of  $E_2$  will also be less than  $1/4n$ . So, in that the analysis will be slightly different. So, if you take  $u$ . Now  $u$  has to fall in this side to the left of  $S$  that is when size  $C$  is going to be greater than  $n/2$ , but when that happens, the number of elements which are greater than the median is at most  $n/2$ . That is if  $u$  falls here on this on the left side of the median, then all the elements to the left of  $u$  in  $R$  are going to be less than the median. So, only these  $n/2$  elements stand a chance of being greater than median. So, the maximum number of elements which are greater than the median in the sampling of  $R$  is going to be bounded by  $n/2$  so then, the analysis is essentially same and we will get the probability of  $E_2$  to be less than  $1/4n$ .

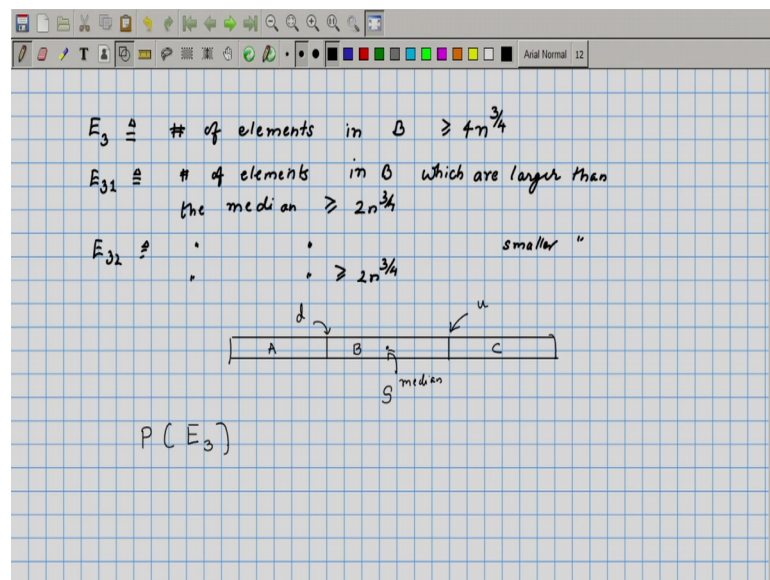
Now we need to look at this event of probability of  $E_3$  ok. So,  $E_3$  was this event. So,  $E_3$  is the event that size of  $B$  is greater than  $4n^{3/4}$  ok. So, once we had split the entire set into 3 parts, the middle element being greater than  $n^{3/4}$ . What is the probability of that happening? We will show that that probability is also small ok. We will look at this set  $B$  and if  $4n^{3/4}$  elements are there in  $B$ , we can say that either half of those elements are going to be greater than the median or half of those elements are going to be less than the median ok.

So, you will split it into 2 events. So, let us say  $E_{31}$  is the event that  $2n^{3/4}$  elements of  $B$  are greater than the median and the next event is  $2n^{3/4}$  elements of  $B$  are less than the median.



elements of B are less than the median ok. So, what is the probability of; so we need to determine. So, whenever B has  $4n$  raised to  $3$  by  $4$  either  $2n$  raised to  $3$  by  $4$  elements in that will be greater than the median or  $2n$  raised to  $3$  by  $4$  elements of B will be less than the median. So, probability of  $E_1, E_{31}$  union  $E_{32}$  is surely less than probability of  $E_{31}$  plus probability of  $E_{32}$  and by symmetry we can say that both these probability will be same. So, this is equal to twice the probability of  $P E_{31}$ . So, we will determine the probability of  $E_{31}$  and show that that is also significantly small. Therefore, the probability of  $E_3$  that is going to be less than some quantity. We will show that this is going to be less than let us say  $n$  by  $n$  raised to minus  $1$  by  $4$  by  $2$ .

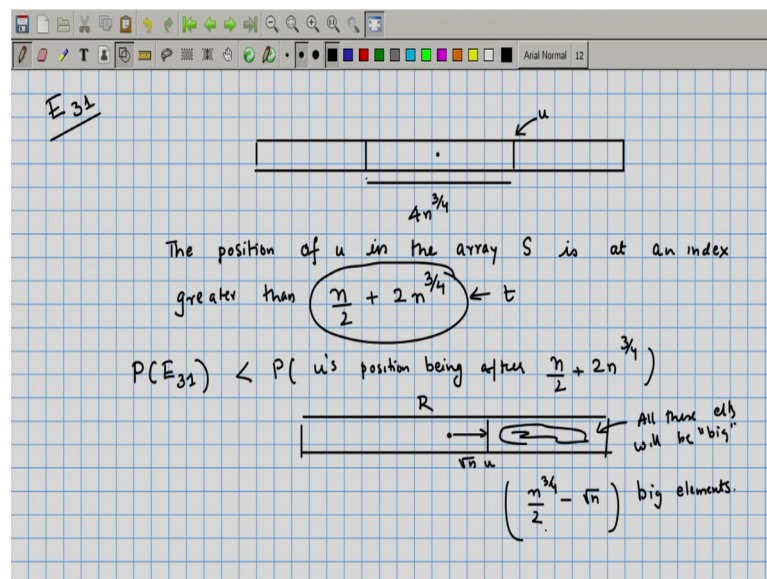
(Refer Slide Time: 34:07)



So, now we look more closely at the third event which is that the elements which are in between  $d$  and  $u$  is larger than  $4n$  by  $3$  by  $4$  times  $n$  raised to  $3$  by  $4$ . So, this was our diagram. If we looked at the set  $S$ , this is its median. Now these were the set  $A$  and these were the set  $C$ , this was  $A$  set  $A B$  and this is  $C$ .  $A$  is all those elements which is smaller than  $d$  and  $C$  is all those elements which are larger than  $u$ ,  $B$  is the in between elements. We want our algorithm will work correctly only when  $B$  is small; small for us is  $4$  times  $n$  raised to  $3$  by  $4$ . We want to analyze what is the probability of this not happening that is what is the probability that number of elements which are in between  $d$  and  $u$  is going to be greater than  $4n$  by  $3$ .

Now when this happens and if the middle third has larger than  $4n^{3/4}$  elements, we can just simply say that at least I mean either half of the elements are larger than median or half of the elements are smaller than the median ok. So, we will look at these 2 sub events and compute their probability and that probability will be a bound on the probability of  $E_3$  ok. So, we are interested in this probability that  $E_3$  occurs ok. So,  $E_3$  is the event that number of elements in  $B$  which are larger than the median is greater than  $2n^{3/4}$  ok. So, how do we compute this ok?

(Refer Slide Time: 36:25)



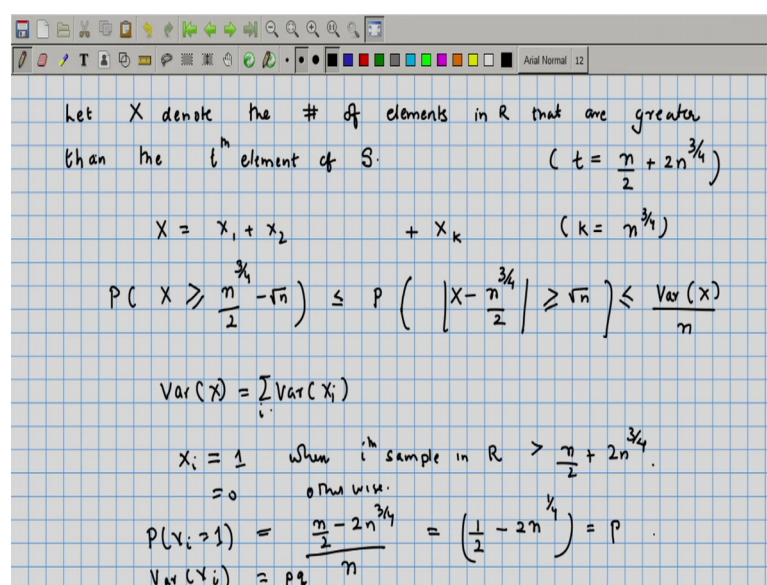
So, if the middle third contains more than  $4n^{3/4}$  elements. The first case was when half of those elements are greater than the median. If half of them are greater than median, then we can say that the position  $u$  comes after all these  $2n^{3/4}$  elements. So, the position of  $u$  in the array  $S$  in the input array is at an index greater than  $n/2 + 2n^{3/4}$  ok. Because  $u$  is appearing after all the elements in  $B$  and in  $B$ , we are assuming that there are at least  $2n^{3/4}$  elements which are larger than the median. If there are  $2n^{3/4}$  elements larger than the median, the position of  $u$  is after all these larger elements and therefore, it is at an index which is greater than  $n/2 + 2n^{3/4}$ . This is the case  $E_{31}$  ok.

We need to compute the probability of this happening. So, probability of  $E_{31}$  is surely less than the probability of  $u$  use position being after  $n/2 + 2n^{3/4}$

4. Now when this happens that is use position is after  $n$  by  $2$  raised to  $n$  by  $2$  plus  $2$  times  $n$  raised to  $3$  by  $4$ . Let us look at the elements that we had randomly sampled. In the elements that we had randomly sampled, this is the position of  $u$  that is look at the middle of  $R$  from there go right by a length  $\sqrt{n}$ . Let that position whatever element was there that was our  $u$ . Since  $u$  itself is greater than the  $n$  by  $2$  plus  $2$   $n$  raise to  $3$  by  $4$  fourth element, all the elements to the right of  $u$  will also be greater than these elements. So, all these elements ok; these will all be will be big ok. Big here means greater than the if we call this as  $t$  greater than the  $t$ 'th element of  $A$  and we have  $n$  raise to  $3$  by  $4$  by  $2$  minus  $\sqrt{n}$  big elements ok.

So, in other words when the event  $E_{31}$  happens, we are guaranteed that our sample contains lot of here a lot of means  $n$  by  $n$  raise to  $3$  by  $4$  by  $2$  minus  $\sqrt{n}$ . So, our sample contain lot of large numbers; large numbers being the numbers which are greater than the number at the  $t$ 'th position where  $t$  is this particular number. We will show that the probability of that is small that is the probability of having large number of large elements when we randomly sample that is going to be small. So, let us compute that probability. We will set up indicator random variables.

(Refer Slide Time: 41:03)



let  $X$  denote the # of elements in  $R$  that are greater than the  $t^{\text{th}}$  element of  $S$ .  $(t = \frac{n}{2} + 2n^{\frac{3}{4}})$

$$X = X_1 + X_2 + \dots + X_k \quad (k = n^{\frac{3}{4}})$$

$$P\left(X \geq \frac{n}{2} - \sqrt{n}\right) \leq P\left(\left|X - \frac{n}{2}\right| \geq \sqrt{n}\right) \leq \frac{\text{Var}(X)}{n}$$

$$\text{Var}(X) = \sum_i \text{Var}(X_i)$$

$X_i = 1$  when  $i^{\text{th}}$  sample in  $R > \frac{n}{2} + 2n^{\frac{3}{4}}$   
 $= 0$  otherwise.

$$P(X_i = 1) = \frac{\frac{n}{2} - 2n^{\frac{3}{4}}}{n} = \left(\frac{1}{2} - 2n^{\frac{1}{4}}\right) = p$$

$$\text{Var}(X_i) = p(1-p)$$

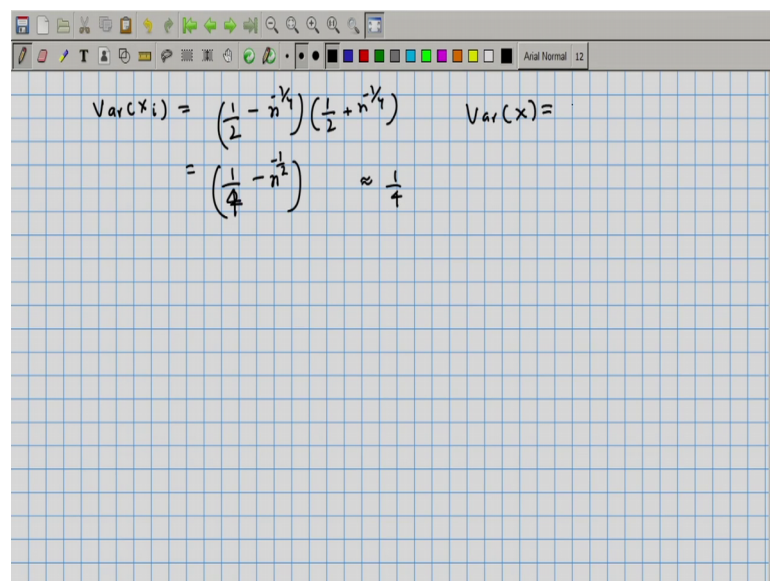
So, let us say  $X$  denote the number of elements in  $R$  that are greater than the  $t$ 'th element of  $S$  ok. So, this is a random variable  $X$ .  $X$  we can write as  $X$  is equal to  $X_1$  plus  $X_2$  plus  $X_k$  that  $k$  is number of samples which was  $n$  raise to  $3$  by  $4$  ok. And we want to

compute the probability that  $X$  is greater than  $ok$ . So, here we have at least these many large elements  $ok$ . So, we will compute the probability that  $n$  raised to  $3$  by  $4$  minus  $2$  minus root  $n$  big elements are present. So, this is by Markov by; so, will apply Chebyshev's inequality here. So, this is less than the probability that  $X$  minus  $n$  raised to  $3$  by  $4$  by  $2$  mod of this is greater than root  $n$  or equal to root  $n$   $ok$ . So, by Chebyshev's inequality, this is going to be less than variance of  $X$  divided by root  $n$  square that is  $n$   $ok$ .

So, we need to compute variance of this random variable that is going to be the; so, variance of  $X$  is going to be variance of  $X_1$  plus  $X_2$   $ok$ . So, that is sum over  $i$  variance of  $X_i$  and the variance of each of these  $X_i$ 's will be same and what is that going to be  $ok$ . So,  $X_i$  is going to be  $1$ . So, what are these indicator random variable? This is  $X_i$  is equal to one when  $i$ 'th sample in  $r$  is greater than  $n$  by  $2$  plus  $2n$  raise to  $3$  by  $4$  and this is  $0$  otherwise  $ok$ . So, probability that  $X_i$  equals  $1$  is equal to there are a total of  $n$  samples out of that the numbers which are greater than this is going to be  $n$  by  $2$  minus  $2n$  raised to  $3$  by  $4$  if you add this sum is  $n$ . So, this divided by  $n$  is the probability that  $X_i$  equals  $1$ . So, that is going to be equal to half minus  $2n$  raised to one by  $4$   $ok$ .

So, therefore so this if we think of it as  $P$  variance of  $X_i$  is going to be  $pq$  for a geometric for a random variable, for a random variable which takes value  $0$  and  $1$  with probability  $P$ . Its variance is going to be  $P$  times  $1$  minus  $P$   $ok$ .

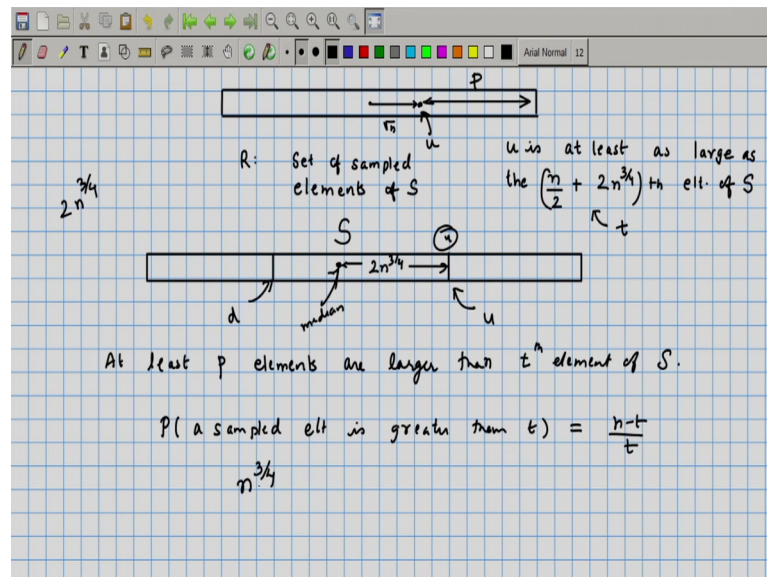
(Refer Slide Time: 45:13)



$$\begin{aligned} \text{Var}(X_i) &= \left(\frac{1}{2} - \frac{1}{n^{1/4}}\right) \left(\frac{1}{2} + \frac{1}{n^{1/4}}\right) & \text{Var}(X) = \dots \\ &= \left(\frac{1}{4} - \frac{1}{n^{1/2}}\right) & \approx \frac{1}{4} \end{aligned}$$

So, that is going to be equal to so, variance of  $X_i$  is going to be half minus  $n$  raised to 1 by 4 into half plus  $n$  raised to 1 by 4 that is equal to half one-fourth minus; so, this is  $n$  raised to minus 1 by 4. So,  $n$  raised to minus half ok. So, you can say that this is approximately one-fourth. So, total variance of  $X$  is that will not (Refer Time: 45:47).

(Refer Slide Time: 45:49)



So, when this set  $R$  has more than  $4n^{3/4}$  elements, we look at this sub event that half of them are greater than the median. If half of them are greater than the median, then the element  $u$  from this sampled elements is going to be at least these are these are greater than the  $2n^{3/4}$  sampled elements. So,  $u$  is going to be at least as large as the  $n/2 + 2n^{3/4}$  element of  $S$  ok.

So, we will call this as  $t$ . So,  $u$  is going to be at least as large as the  $t^{th}$  element of  $S$  ok. Once again this is happening because we assume that  $2n^{3/4}$  elements are larger than the median. So, in our sample these many elements are larger than the median. If these many elements are larger than the sorry, we assume that in our sample when we split the set  $S$  on the basis of the sampled elements, the middle portion contains large number of large elements when large number of elements which are in between  $d$  and  $u$ . So, in there are lots of elements between  $d$  and  $u$  and if we assume that half of those elements are greater than  $u$ , I mean are greater than the median then since  $u$  comes after all these elements,  $u$  is going to be greater than  $n/2 + 2n^{3/4}$  element of  $S$ .



So, this was our  $d$  in our set  $S$ , this was our  $d$  and this was our  $u$  and if there are a large number of elements that is  $2n$  raised to  $3/4$  elements which are greater than the median, then  $u$  is clearly greater than the  $t$ 'th element of  $S$ . Median is then by  $2$  the element at  $n$  by  $2$  position and  $2$  times  $3n$  by  $4$  elements are already greater in this middle portion. So,  $u$  comes certainly after this position. Therefore, if we look our samples we can claim that at least these many elements are greater than the number at the  $t$ 'th position. So, at least  $t$  elements so let us call this as  $p$  at least  $p$  elements are larger than the  $t$ 'th element of  $S$ . If you pick one element what is the probability that it is greater than the  $t$ 'th element?

Well that is going to be so, probability that a sample element is greater than  $t$  that is going to be equal to  $n$  minus  $t$  by  $n$  ok. So, now, we have picked  $n$  raised to  $3/4$  samples [noise.] We need to estimate what is the probability that  $p$  of those samples are greater than  $t$  ok. So, we will compute that.

(Refer Slide Time: 50:35)

$$\begin{aligned}
 X &\triangleq \# \text{ of elements in } R \text{ which are greater than } t. \\
 X &= X_1 + X_2 + \dots + X_k \quad (k = n^{3/4}) \\
 X_i &= \begin{cases} 1 & \text{if the } i^{\text{th}} \text{ sample is greater than } t. \\ 0 & \text{otherwise} \end{cases} \\
 P(X_i=1) &= \frac{n-t}{n} = \frac{n - (\frac{n}{2} + 2n^{3/4})}{n} = \left( \frac{\frac{n}{2} - 2n^{3/4}}{n} \right) = \frac{1}{2} - \frac{2}{n^{1/4}} \\
 E[X_i] &= \frac{1}{2} - \frac{2}{n^{1/4}} = p \\
 E[X] &= n^{3/4} \times \left( \frac{1}{2} - \frac{2}{n^{1/4}} \right) = \left( \frac{\frac{n^{3/4}}{2} - 2\sqrt{n}}{1} \right) \\
 \text{Var}[X] &= \sum_i \text{Var}(X_i) = n^{3/4} \times \left( \left( \frac{1}{2} \right)^2 - \frac{4}{\sqrt{n}} \right) = \left( \frac{\frac{n^{3/4}}{4} - 4 \times n^{1/4}}{1} \right)
 \end{aligned}$$

So, let us say  $X$  denotes the random variable. So,  $X$  is the number of elements in  $R$  which are greater than  $t$ . So,  $X$  we will write as  $X_1$  plus  $X_2$  plus  $X_k$  where  $k$  is going to be equal to  $n$  raised to  $3/4$ . And  $X_1$  is equal to or any  $X_i$  is equal to  $1$  if the  $i$ 'th sample is greater than  $t$  ok. And probability that  $X_i$  is equal and this is equal to  $0$  otherwise probability that  $X_i$  is equal to  $1$  it is going to be  $n$  minus  $t$  by  $n$  that is going to be equal to  $n$  minus  $n$  by  $2$  plus  $2n$  raised to  $3/4$  that is the value of  $t$  divided by  $n$  that is equal



to  $n^{3/4} - 2\sqrt{n}$ . Therefore, expectation of  $X_i$  will be  $\frac{1}{2} - \frac{2}{n^{1/4}}$ .

So, that is going to be  $\frac{1}{2} - \frac{2}{n^{1/4}}$ . So, expectation of  $X_i$  is  $\frac{1}{2} - \frac{2}{n^{1/4}}$ . So, expectation of  $X$  will be just the sum of these. So, that will be  $n^{3/4} \left( \frac{1}{2} - \frac{2}{n^{1/4}} \right)$ . It is going to be  $n^{3/4} - 2\sqrt{n}$ . So, variance of  $X$  is going to be the sum of the variances these are independent random variables.

So, that will be sum over  $i$  variance of  $X_i$  that is going to be again  $n^{3/4}$  times the individual variance that is going to be  $\frac{1}{4}$ , then the overall then the variance of this is going to be  $\frac{n^{3/4}}{4}$ . So, that is going to be. So, variance is going to be  $\frac{n^{3/4}}{4}$ .

(Refer Slide Time: 54:45)

$$E[X_i] = \frac{1}{2} - \frac{2}{n^{1/4}} = p$$

$$E[X] = n^{3/4} \left( \frac{1}{2} - \frac{2}{n^{1/4}} \right) = \left( \frac{n^{3/4}}{2} - 2\sqrt{n} \right)$$

$$Var[X] = \sum_i Var(X_i) = n^{3/4} \left( \left( \frac{1}{2} \right)^2 - \frac{4}{\sqrt{n}} \right) = \frac{n^{3/4}}{4}$$

$$P\left(X > \frac{n^{3/4}}{2} - \sqrt{n}\right) = P\left(X - \left(\frac{n^{3/4}}{2} - 2\sqrt{n}\right) > +2\sqrt{n} - \sqrt{n}\right)$$

$$= P\left(X - E[X] > \sqrt{n}\right) \leq \frac{Var(X)}{\frac{n^{3/4}}{4}}$$

$$\leq \frac{\frac{n^{3/4}}{4}}{\frac{n^{3/4}}{4}} = 1$$

Now so, we need to look at this particular probability; probability that  $X$  is greater than  $n^{3/4} - 2\sqrt{n}$ . We can write this as probability that  $X - n^{3/4} + 2\sqrt{n}$ .

So, this is going to be expectation of  $X$  plus so, greater than plus  $2\sqrt{n} - \sqrt{n}$ . Just rearrange the expression so, that is the probability that  $X - E[X] > \sqrt{n}$ .

than root  $n$ . So, this probability by Chebyshev bounds, we can say that is going to be less than variance of  $X$  divided by a square  $a$  is  $n$  here. So, that is going to be less than  $n$  raised to 3 by 4 by 4 minus 4  $n$  raised to 1 by 4 by  $n$  ok. So, that is clearly less than  $n$  raised to 3 by 4 by 4  $n$  ok.

(Refer Slide Time: 56:43)

$$\begin{aligned}
 P(E_{31}) &< \frac{n^{3/4}}{4n} = \frac{1}{4n^{1/4}} \\
 P(E_{32}) &< \frac{1}{4n^{1/4}} \\
 P(E_3) &< \frac{1}{2n^{1/4}} \\
 P(E_1) &\leq \frac{1}{4n^{1/4}} \\
 P(E_2) &\leq \frac{1}{4n^{1/4}} \\
 P(E) &\leq \frac{1}{n^{1/4}} \\
 \hline
 &\text{1 - } \frac{1}{n^4}
 \end{aligned}$$

So, we know that probability of the bad event that is  $E_{31}$  is less than 4  $n$  raised to 3 by 4 by  $n$ . So, this is this is equal to 4 by  $n$  raised to 1 by 4. Similarly, we can show that probability of  $E_{32}$  is also less than 4 by  $n$  raised to 1 by 4. So, this is less than 4 by  $n$  raised to 1 by 4. So, together is in the denominator. So, you can.

So, this is going to be less than 1 1 by 4  $n$  raise to 1 by 4. So, total probability so, probability of  $E_3$  is going to be less than half  $n$  raised to 1 by 4. So, probability of  $E_1$  and probability of  $E_2$ , they are also less than 1 by 4 to the power  $n$  raised to 1 by 4. This is  $n$  raised by 4  $n$  raised 1 by 4. So, together probability of bad event is going to be less than 1 by  $n$  raised to 1 by 4 ok. So, the probability of success is 1 minus 1 by  $n$  raised to 1 by 4 ok. So, that is a reasonable successes probability.