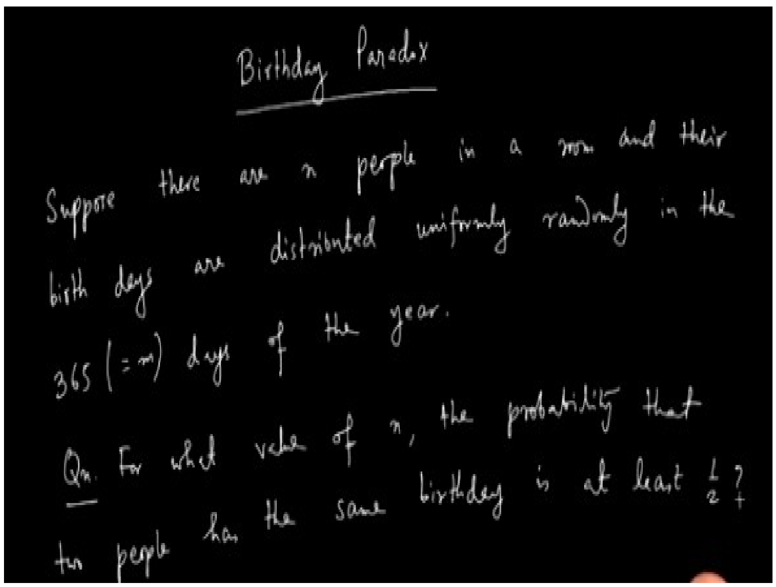
Selected Topics in Algorithm
Prof. Palash Dey
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur

Module No # 04
Lecture No # 17
Balls and Bins

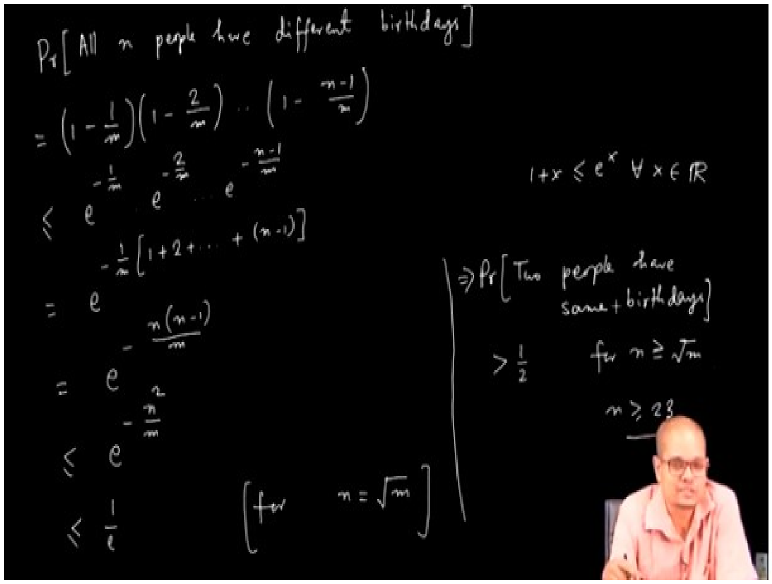
Welcome in the last lecture we have seen some application of concentration bounds in coupon collector problems. We continue that line of discussion here so our first problem today is what is called popularly known as birthday paradox.

(Refer Slide Time: 00:43)



Suppose there are n people in a room and their birthdays are distributed uniformly, randomly in the 365 let us call these m days of the year. Question for what value of n the probability that 2 people has the same birthday is at least half of course if n = 365 or more than 365 then with probability 1 there must exist at least 2 people with the same birthday. But the paradox is that you know we will see that for much smaller value of n the probability will be at least will be more than half that 2 people have the same birthday and that is why the paradox. So again so we will assume that m is the number of days m is 365.

(Refer Slide Time: 04:03)



So if there are 2 people who have the same birthday there let us call this collision this particular event collision probability of c. And what is the probability that you know there is no collision so let us all in people have different birthdays no collision. So what is this probability? So the first people can have birthday any day the second people for the second person the probability that its birth is not with the first is this the probability that the third person does not have birthday with the first 2 is this and so on $1-\frac{n-1}{m}$.

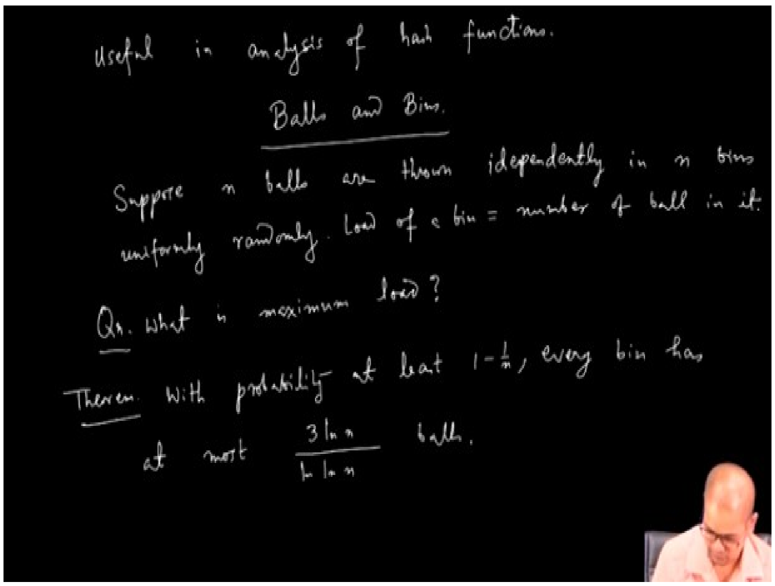
Now we apply the inequality that $1+x \le e^x$ for all x in real number both positive and negative this is then $e^{\frac{-1}{m}} e^{\frac{-2}{m}} \ldots e^{\frac{-(n-1)}{m}}$ so this is $e^{\frac{-1}{m}(1+2+\ldots+(n-1))}$. This is $e^{\frac{-n(n-1)}{m}}$ this is less than equal to $e^{\frac{-n^2}{m}}$. So if this will be less than equal to if we pick $n=\sqrt{m}$ then this is $\frac{1}{e}$ which is already less than half.

So for $n=\sqrt{m}$ the probability that all people have different vertices is less than equal to $\frac{1}{e}$. That is the probability that 2 people have different bodies this is greater than half for in is $n=\sqrt{m}$ so for if we put m = 365, for m = 365 then n is greater than equal to 23. So if there are 23 people in a room the probability that at least 2 people have same but they have at least 2 people have same birth you know that this probability is more than 50%.

So this particular analysis turns out to be very useful in say entries of hash functions you know. What is the how big the size of the universe or whatever how big the size of the range or codomain of the hash function should be if I want to hash n numbers? So if the hash n numbers and if the size of the hash table is more than n^2 or hash function the size of the range of the hash function is more than n^2.

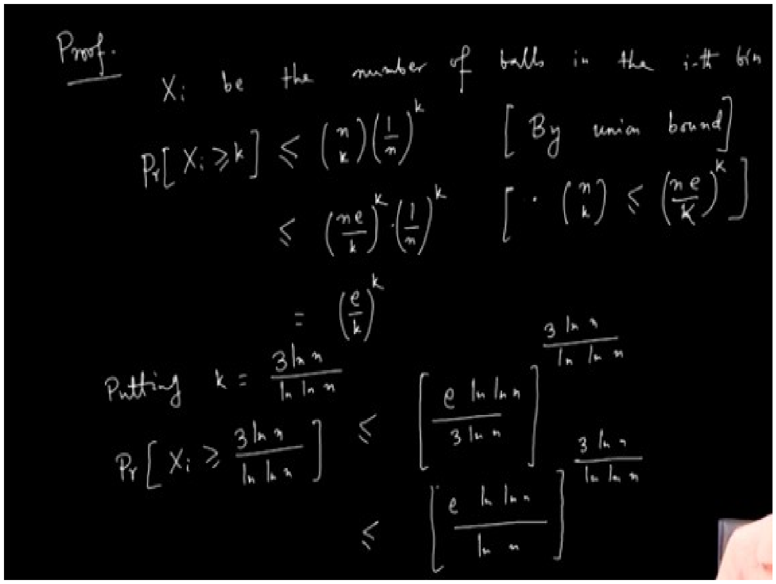
So $2n^2$ then the probability that there is a collision is less than half but if it is less than n square the size is not quadratically more than the number of elements we are hashing then this probability of collision becomes more than 50%.

(Refer Slide Time: 09:21).



So these are useful in analysis of hash functions now again motivated by hash function our next experiment is balls and bins. So suppose in balls are thrown independently in n bins uniformly randomly and load of a bin is number of balls thrown in it. So question is what is maximum load? So for that we prove this theorem with probability at least $1-\frac{1}{n}$ every bin has at most $\frac{3\ln n}{\ln\ln n}$ balls.

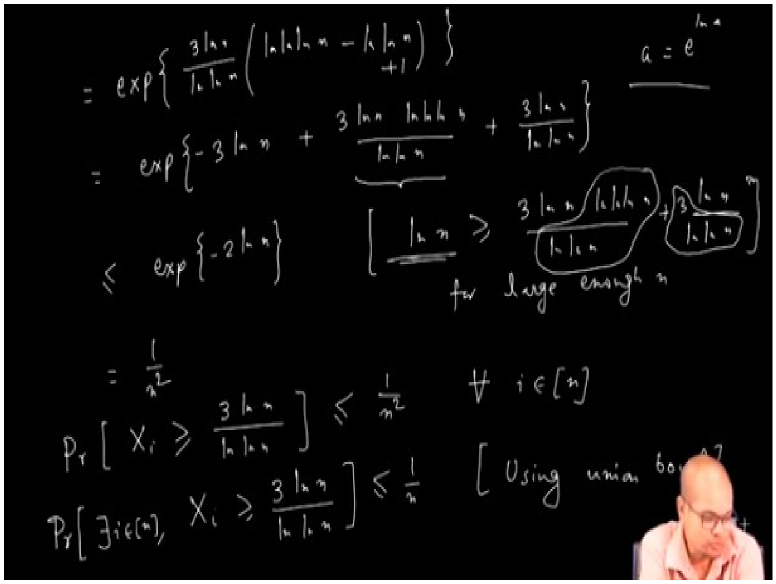
(Refer Slide Time: 12:33)



Proof so let X_i be the number of balls in the i-th bin so probability that X_i is greater than equal to k is by union bound less than equal to ${}^nC_k\left(\frac{1}{n}\right)^k$ this is by union bound this is inches case less than equal to $\left(\frac{ne}{k}\right)^k\left(\frac{1}{n}\right)^k$ which is like $\left(\frac{e}{k}\right)^k$. This is since ${}^nC_k \leq \left(\frac{ne}{k}\right)^k$ now put $k=\frac{3\ln n}{\ln\ln n}$.

So putting $k=\frac{3\ln n}{\ln\ln n}$ probability that X_i is greater than equal to $\frac{3\ln n}{\ln\ln n}$ is less than equal to $\left(\frac{e\ln\ln n}{3\ln n}\right)^{\frac{3\ln n}{\ln\ln n}}$. So you can get rid of the 3 from the denominator and can write this is less than equal to $\left(\frac{e\ln\ln n}{\ln n}\right)^{\frac{3\ln n}{\ln\ln n}}$.

(Refer Slide Time: 16:04)



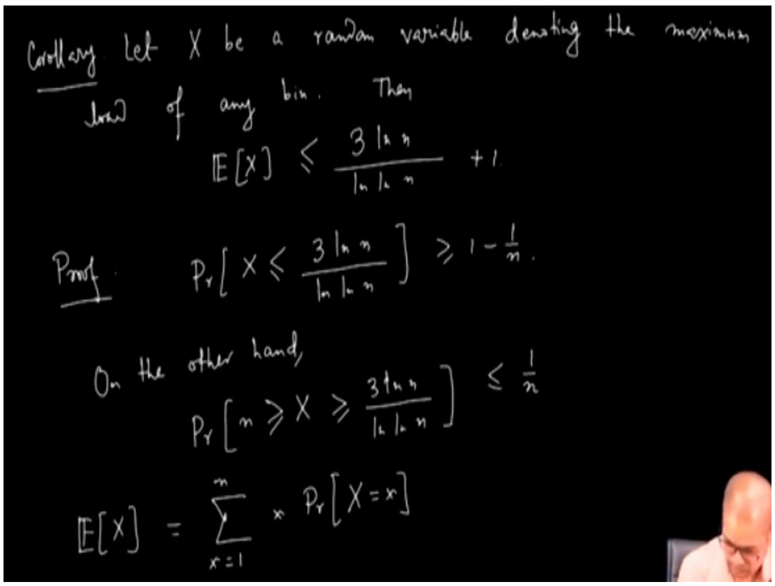
Now let us write everything as a power of e so let me write this way exponential you know this exponent was there $\frac{3\ln n}{\ln\ln n}$ and any number say e can be written as e to the power and ln n a so using this what we write is that this inside e $\frac{\ln n}{\ln\ln n}$ as e to the power ln n of everything this is $\ln\ln\ln n - \ln\ln n$. So this is exponent of $-3\ln n$ let me write this second term first $\frac{+3\ln n\ln\ln\ln n}{\ln\ln n}$. And now this term $\frac{\ln n\ln\ln\ln n}{\ln\ln n}$ this increases slowly then ln n.

So we can write it that this is exponent of -2 ln n here will be another that e was there this so this e is e to the power $\frac{\ln n}{\ln\ln n}$ so that term also will be there this +1 so here it will be $\frac{+3\ln n}{n}$. So why we can write it because ln n is more than equal to $\frac{3\ln n\ln\ln\ln n}{\ln\ln n}+\frac{3\ln n}{\ln\ln n}$ this is for large enough n so this bound in the theorem also holds is for large enough n and so this is nothing but one over n square.

So this particular thing follows from you know this function is ln n whereas if you see here the coefficient of ln n is decreasing in n. So this function ln n times something which decreases in n grows smaller than ln n and here also the coefficient of ln n $\frac{3}{\ln\ln n}$ is decreasing in n so both functions grows smaller than slower than ln n. Hence for large enough in ln n will be more than any constant not even 3 any constant times sum of these 2 functions.

So now what we have is probability that X_i is greater than equal to $\frac{3\ln n}{\ln\ln n}$ is less than equal to 1 over n square this holds for all i in. So now we apply union bound by union bound probability that there exists an i in n such that X_i is greater than equal to $\frac{3\ln n}{\ln\ln n}$ there are n such terms this is 1 by n using union bond which proves a theorem. The probability that there exists at least; 1 bin whose load is more than or at least 3. A $\frac{3\ln n}{\ln\ln n}$ these at most $\frac{1}{n}$ so the probability that all bins has at most $\frac{3\ln n}{\ln\ln n}$ balls is at most $1-\frac{1}{n}$ or at least $1-\frac{1}{n}$.

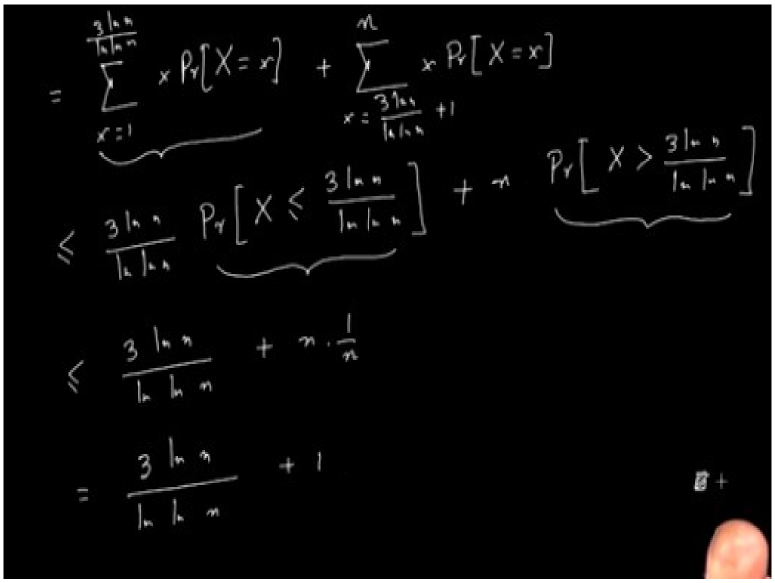
(Refer Slide Time: 23:02)



So using this we can bound the expected max load so let me write it as a corollary let X be a random variable denoting the maximum load of any bin then expectation of X is less than equal to $\frac{3\ln n}{\ln\ln n}$ + 1. So let us prove it proof so what you already have is probability that X maximum load is less than equal to $\frac{3\ln n}{\ln\ln n}$ this is greater than equal to $1-\frac{1}{n}$. And what could be the maximum load?

Of course on the other hand probability X is greater than equal to $\frac{3\ln n}{\ln\ln n}$ and this is less than equal to n because it can maximum be n this is then less than equal to $\frac{1}{n}$. Now let us compute the expectation of X what is it of x is by definition summation x is 1 to, n probability that X = x.

(Refer Slide Time: 25:48)



So you can break this sum into 2 parts X probability X = x this is from x = 1 to $\frac{3\ln n}{\ln\ln n}$ +x times probability X = x = 3 ln in by ln ln n +1 up to n. So now we would do some upper bound for the first term X can be upper bounded by $\frac{3\ln n}{\ln\ln n}$ and then this is the probability that X is less than equal to 3 ln n by ln ln n plus in this part for the second part X can be upper bounded by n and this is the probability that X is greater than $\frac{3\ln n}{\ln\ln n}$.

Now this probability is at most 1 this is again less than equal to $\frac{3\ln n}{\ln\ln n}$ + n times this probability is at most 1 -1 this probability is at most $\frac{1}{n}$. So this is $\frac{3\ln n}{\ln\ln n}$ +1 which concludes the proof so we will stop here today.