

**Probability & Computing**  
**Prof. John Augustine**  
**Department of Computer Science and Engineering**  
**Indian Institute of Technology Madras**

**Module – 03**

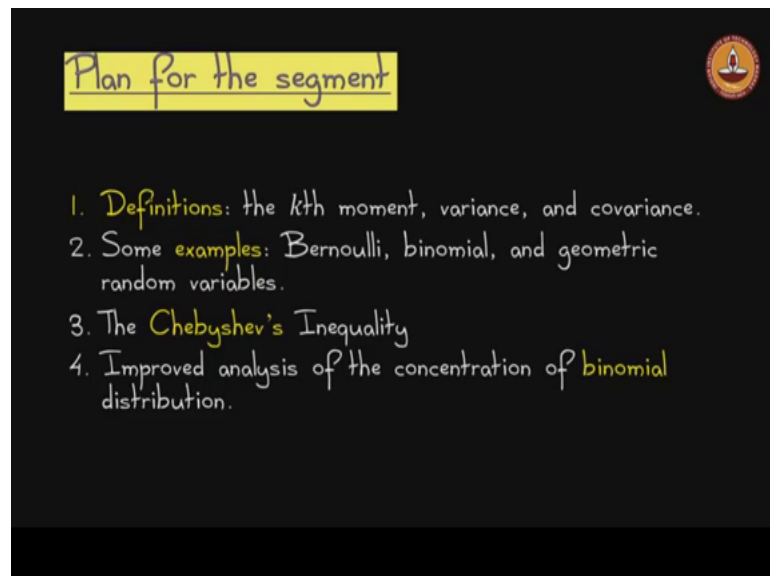
**Tail Bounds I**

**Lecture - 16**

**Tail Bounds I - The Second Moment, Variance & Chebyshev's Inequality**

So, now we are starting the second segment in module three we will see a few definitions and introduce the next tail bound.

(Refer Slide Time: 00:21)



So, the definitions will particularly be that be the  $k$ th moment use that in particular the second moment will be used to define something called the variance and a related notion called the covariance. Well run through some examples to understand the notion of variance and use the variance to get a better bound called the Chebyshev's inequality and we will see that the binomial distribution is captured better by the Chebyshev's inequality that is the goal for this segment.

(Refer Slide Time: 00:58)

Some definitions

- The  $k^{\text{th}}$  moment of a random variable  $X$  is  $E[X^k]$ .
- The first moment is simply  $E[X]$ .
- The second moment along with the first yields the notion of variance:  
$$\text{Var}[X] = E[(X - E[X])^2]$$
$$= E[X^2] - (E[X])^2.$$

Recall Jensen's Inequality

The slide also features a small video inset of a man in a blue shirt speaking in front of a green chalkboard.

The  $k$ th moment is simply this its  $E$  the expectation of  $X$  to the  $k$  and the first moment is something that you have already seen extensively that is nothing, but the expectation you have  $X$  to the 1 is this  $E$  of  $X$  that is expectation ah. The second moment along with the first moment yields the notion of variance and if you recall if you look at the second moment  $X X$  squared; it is a convex function right.

So,  $E$  of  $X$  squared is going to be greater than  $E$  of  $X$  the whole squared that was Jensen's inequality right and that difference if you recall when we looked at Jensen's inequality we claimed that the difference actually measures how much the random variable deviates from the mean and that is essentially the idea that we are going to capture here.

So, variance of  $X$  is nothing, but the expectation of  $X$  minus  $E$  of  $X$  squared this is one definition. The other definition is  $E$  of  $X$  squared minus  $E$  of  $X$  the whole square ok. So,  $E$  of  $X$  squared is nothing, but this second moment  $E$  of  $X$  is nothing, but the first moment putting the two together you get the variance ok; these two definitions are equivalent and a quick homework for you would be the check that they are in fact, equivalent. So, if you just apply the formulas and run through you should be able to get them.

(Refer Slide Time: 02:29)

$$\begin{aligned} \text{Var}[X + Y] &= ? \\ \text{Var}[X + Y] &= E[(X + Y - E[X] - E[Y])^2] \\ &= E[(X - E[X])^2 + E[(Y - E[Y])^2] + 2E[(X - E[X]) \cdot (Y - E[Y])] \\ &= \text{Var}[X] + \text{Var}[Y] + 2\text{Cov}(X, Y). \end{aligned}$$

$E[X \cdot Y] - E[X] \cdot E[Y]$

So, let us now try to understand what is variance of X plus Y where X and Y are two random variables. Remember this is in line with how we approached expectation we tried to understand what is expectation of the sum of two random variables and we want to know if something similar comes up with variance as well.

So, we apply the formula variance of X plus Y is simply the expectation of we want X plus Y. So, X plus Y minus E of X plus Y which when you apply the linearity of expectation is going to be minus E of X minus E of Y; the whole square that is variance of X plus Y. And I have coloured it up in yellows and blues because I am going to regroup them in accordance to their colours it is its expectation of ah.

So, here I am going to take the square I am going to consider X minus E of X is one term and Y minus E of Y as the other term. So, and then when you square it you are going to get all these terms and then apply the linearity of expectation over all those terms; so, I will just skip that. So, essentially what you will get E of expectation of X minus E of X the whole square plus expectation of Y minus E of Y the whole squared times some quantity. So, this you have expectation of X minus E of X the whole square is nothing, but variance of X then you get variance of Y plus 2 times this strange looking object here ok. And again another small homework is if you work this out its actually going to be nothing, but E of X Y minus E of X times E of Y and this up strange object is called the covariance.

This is if you look at this variance of X plus Y; we have shown this is variance of X plus variance of Y plus something ok. We want to know whether that something becomes a 0 and what is the condition under which it becomes 0 because that then relates to the expectation because there E of X equal to E of X E of X plus Y would have been equal to E of X plus E of Y.

(Refer Slide Time: 04:42)

Is  $E[X \cdot Y] = E[X] \cdot E[Y]$ ?

$$E[X \cdot Y] = \sum_x \sum_y xy \Pr((X = x) \cap (Y = y))$$

Now what? Except... ?

When  $X$  and  $Y$  are independent, then,

$$E[X \cdot Y] - E[X] \cdot E[Y]$$

$$= \sum_x \sum_y xy \Pr(X = x) \cdot \Pr(Y = y)$$

$$= \sum_x x \Pr(X = x) \cdot \sum_y y \Pr(Y = y)$$

$$= E[X] \cdot E[Y].$$

So, we and this object will become a 0 when these two terms are equal right. So, can they be made equal what condition will they be made equal. So, let us actually work that out E of X times Y; you apply the formula its summation over X, summation over Y; X Y times the probability that X equal to X and Y equal to Y you apply all the formulas, but then you are stuck here you do not know what to do with probability of X equal to X intersected with Y equal to Y; except when the two random variables are independent. If the two random variables are independent then you can replace it with X equal to X times Y equal to Y in which case it works out to be E of X times E of Y alright.

So, essentially what; that means, is that when the two random variables X and Y are independent, you will get something similar to the linearity of expectation otherwise you will lose the linearity of expectation. But nevertheless we can this quantity this covariance actually has some meaning because if its independent its equal what if it is not independent there is some meaning to it.

(Refer Slide Time: 06:02)

$E[X \cdot Y] - E[X] \cdot E[Y]$

$0 - \frac{1}{4} = -\frac{1}{4}$     
  $0 - \frac{1}{3} = -\frac{1}{3}$     
  $\frac{1}{4} - \frac{2}{3} = -\frac{1}{3}$     
  $\frac{1}{2} - \frac{1}{4} = \frac{1}{4}$

So, the best way to at least get some sense of it is to actually work out some examples. So, let us take two random variables let us say they are outcomes of coin tosses and one if its heads, 0 if its tails or something like that ah. In this case so, what is what; what is this quantity of  $X \cdot Y$  minus  $E[X] \cdot E[Y]$  should be immediate, it is 0 why because these are two independent random variables ok.

So, let us do something silly here let us limit the sample space to one of these. So, let us let us just focus on this one ok. So, this in this case we are we are not considering this outcome. So, the sample space is limited to the three outcomes shown in red ok. So, then now and let us say this is always  $X$  comma  $Y$ ; so, what is the expectation of  $X$  here? Expectation of  $X$  this one is a 0, 0 and 1; so, the expectation of  $X$  is one third right.

So, I am trying to work out this thing expectation of  $Y$ ; what is expectation of  $Y$ ? that is also 1 by 3 and what is the expectation of  $X$  times  $Y$ ? 0 why because at least one of them is always remaining as 0. So, it is; so, it is going to be minus 1 by 9 and you can work out something like that for this other example shown here as well ok. Let us do something similar what happened what about this case? So, let us again pick the one that will ok.

So, here let us lets again work on this one ok. So, what is the expectation of  $X$  here? It is going to be two thirds; what is the expectation of  $Y$ ? It os going to be one third; what is the expectation of  $X$  times  $Y$ ? One third and so, that will turn out to be one third minus 2

over 9 right. So, it is going to be plus 1 over 9 that is what do you think would be the case over here? Well  $E$  of  $X Y$  is in this case is  $E$  of  $X$  times;  $Y$  is 0 minus 4. Here if you work it out its going to be what is that going to be what is  $E$  of  $X$  times  $Y$ ? It is it is going to be half minus  $E$  of  $X$  again minus 1 by 4 I think well 2 plus 1 by 4  $Y$  we going through all of this?

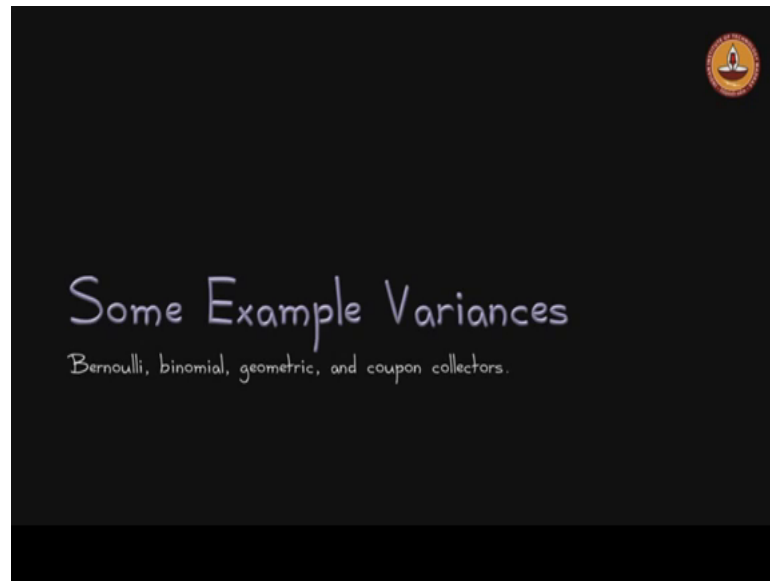
So, it is it is what we would we sort of guess is its ranging from minus 1 by 4 to plus 1 by 4 what is this measure telling you? So, why is this for example, this one by 4 what is the intuition here as to why the covariance is maximized here?

Student: (Refer Time: 10:09).

Yes it is like you have tied the two coins together. So, if one happens to be a tails the other one also the tails. So, they are what is called positively correlated whereas, here you have tied them together, but you have tied them. So, that when one appears tails the other one appears heads always. So, they are negatively correlated when one is high the other one automatically becomes low and so, on ok.

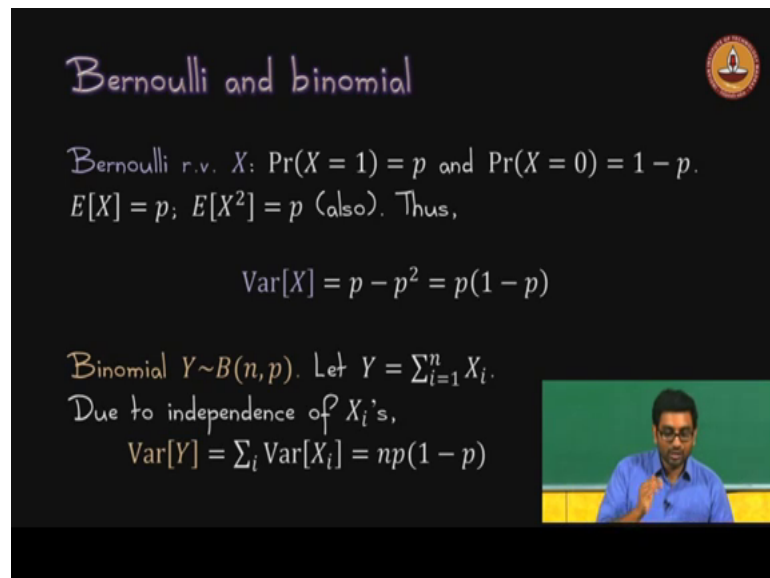
So, this covariance actually is a very meaningful object; it actually measures how connected these two variables are ok. And when it is when they are completely dis or unconnected or the technical term being independent it is equal to a 0 right. So, this is a very useful object, but for our purposes we will stop at this level of intuition because what we want to do is use this the notion of variance to go on to tail bounds, but I want to make sure this notion is also clear in our minds at least at the intuitive level ok.

(Refer Slide Time: 11:10)



So, let us quickly look at some variances. So, let us look at the Bernoulli random variable.

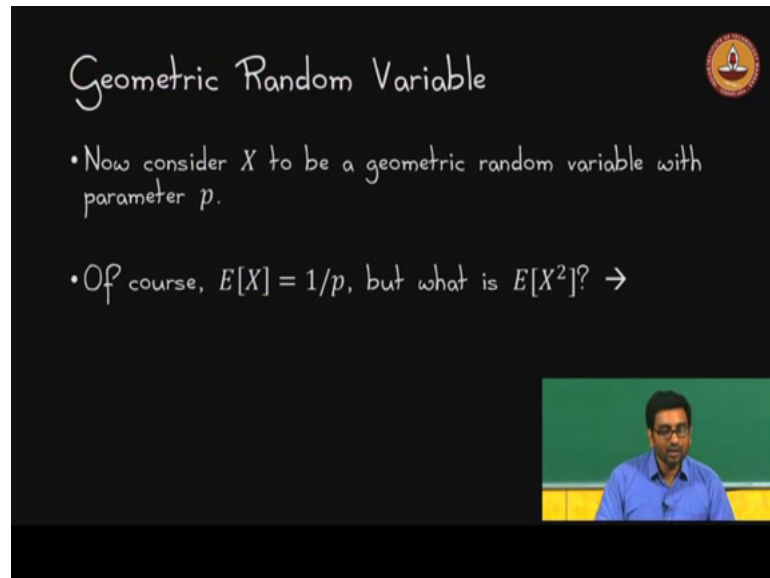
(Refer Slide Time: 11:17)



So,  $X$  equal to 1 with probability  $p$ ; 0 probability  $1 - p$ ; the expectation of  $X$  is  $p$ , expectation of  $X$  squared is also  $p$ . Because when you square 0 1 random variables remains 0 1 random variable right. So, the variance of  $X$  you run it through the formula, you are going to get  $p$  times  $1 - p$ . The binomial random variable it is just the

summation of  $n$  Bernoulli random variables. So, you just do to independence you can just sum them up.

(Refer Slide Time: 11:52)



The slide has a black background with white handwritten text. At the top right is a small circular logo. The text on the slide reads:

### Geometric Random Variable

- Now consider  $X$  to be a geometric random variable with parameter  $p$ .
- Of course,  $E[X] = 1/p$ , but what is  $E[X^2]$ ? →

In the bottom right corner, there is a small video inset showing a man in a blue shirt speaking in front of a green chalkboard.

So, you get  $n$  times  $p$  into  $1$  minus  $p$ . Geometric random variable on the other hand is a little bit more tricky because you do not have this neat summation you do not know how long you are going to do this these iterations right.

Student: The binominal (Refer Time: 12:03) they are independent. So, we.

Sorry did I say dependent?

Student: No.

They are independent that is why the variance of  $Y$  can be written as the sum of the variance of the individual  $X$  is. So, here I am using  $X_i$ 's to denote.

Student: How know like can you get a similar expression for covariance here like ah?

The covariances will be  $0$  right; oh I see you are you are talking about the more general thing where you know it is not just  $X$  plus  $Y$ .

Student: (Refer Time: 12:35).



So, there is that will be a little bit more complicated thing, but its it is something that can be written ok.

But since all of those covariance terms are going to be 0 because of independence. So, let us so, we are kind of sweeping some details under the rug for now, but essentially since all the covariance terms are going to be 0 just you.

Let us now consider the geometric random variable  $X$  is a geometric random variable and it has parameter  $p$ . So, this means you toss a coin with bias  $p$  until you see the first heads. And by now we should know that the expectation of  $X$  is  $1$  over  $p$ ; you can work that out; our question now is what is the variance of this random variable  $X$ ? And towards understanding the variance, we first want to compute  $E$  of  $X$  square and that is what we are going to do now.

(Refer Slide Time: 13:40)

**Geometric Random Variable**

$$\begin{aligned}
 E[X^2] &= \Pr(\text{first flip tails}) \cdot E[X^2 \mid \text{first flip tails}] + \\
 &\quad \Pr(\text{first flip heads}) \cdot E[X^2 \mid \text{first flip heads}] \\
 &= (1-p)E[(X'+1)^2] + p \\
 &\quad \{X' \text{ also geometric with parameter } p\} \\
 &= (1-p)(E[(X')^2] + 2E[X'] + 1) + p \\
 &= E[X^2] + 2E[X] + 1 - pE[X^2] - 2pE[X] - p + p
 \end{aligned}$$

Thus,  $E[X^2] = \frac{2}{p^2} - \frac{1}{p}$  and

$$\text{Var}[X] = E[X^2] - E[X]^2 = \frac{1-p}{p^2}$$

Let us consider what  $E$  of  $X$  square is and we will write it out using the law of total expectation. Now  $X$  squared to compute  $X$  squared you can you break it break the universe into two parts. The first part corresponds to where the first coin flip lands tails and the second part corresponds to the first flip landing heads ok.

So, now our expectation of  $X$  squared is split into these two parts and of course, they have to be weighted by their corresponding probabilities ok. And let us look at this expression let us look at the second one the one where the first flip is heads. If the first

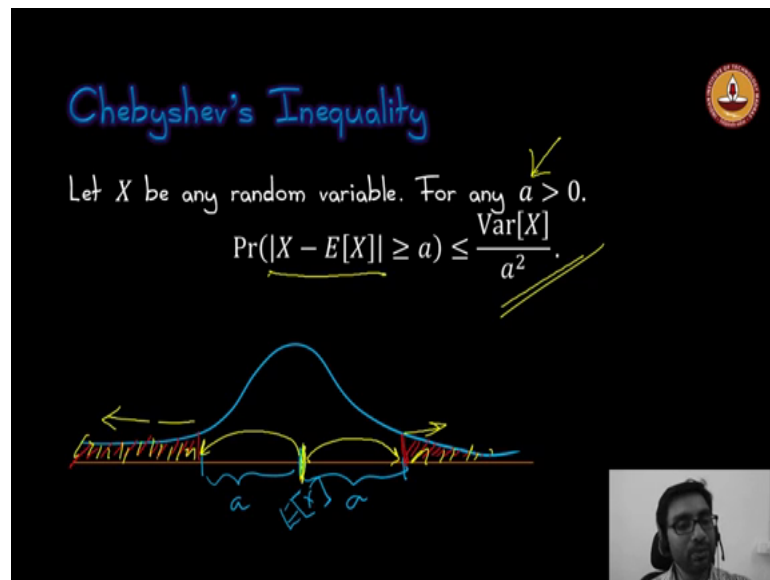
flip is heads then  $X$  is going to be 1 because you have seen the heads and therefore, you are not going to toss anymore. So, then  $X$  squared is also going to be 1; so, this whole term becomes a 1 ok. And we know the probability of flipping a heads is  $P$  because it's a biased coin ok. So, that leads us to just the term  $P$  because this is just going to be a 1; so, the second term here is just going to be  $p$ .

Let us now consider this first expression; well the first the probability that the first flip is a tails is  $1 - P$  and that we have and let us consider what happens when the first flip is tails. After the first flip you really have to this is the geometric random variable; so, there is this memoryless property. So, you have to set aside the first coin toss and then now again all start all over and wait for the first heads to occur. So, in essence after the first flip has been set aside you really are in a situation where you have to repeat the geometric random variable from the start.

So, that is why we have  $X'$  which is another geometric random variable with the same parameter  $p$  plus 1 this is this 1 is where you set aside the first coin toss. And then you start over with a geometric random variable  $X'$  that is and this whole thing is within a square. So, it is going to be  $X' + 1$  the whole square and we want the expectation of that that is that is going to be this part ok.

And now we can since it is a square, you can apply the formula and then apply linearity of expectation over and we will get be getting  $E$  of  $X$  squared plus  $2 E$  of  $X'$   $E$  of  $X'$  square rather plus 1. And so, this when we expand it out we are going to get this expression for the expectation of  $X$  square ok, but that is just the expectation of  $X$  square; the variance is given by the expectation of  $X$  squared minus the expectation of  $X$  the whole square that is the formula for variance if you recall. So, if you apply that you are going to end up with the variance of a geometric random variable with parameter  $p$  being  $1 - p$  over  $p$  square ok. So, with that we have seen a couple of examples of the of computing the variance for random variables.

(Refer Slide Time: 17:52)



We are now ready to talk about this state inequality called the Chebyshev's inequality. And as you may expect the Chebyshev's inequality depends on the variance of a random variable. So, you in order to be able to apply this inequality you need to know the variance of that other random variable ok. So, let us take  $X$  to be any random variable now notice that  $X$  need not be non negative  $X$  can be anything ok. And now  $a$  is some parameter in this context we are not just going to bound the upper tail like we did in Markov's inequality, we are going to bound both the upper and the lower tails ok.

So, in precisely speaking we are interested in  $X$  minus  $E$  of  $X$ ; the absolute value being larger than  $a$ . And if you think about that it corresponds to these regions within the distribution. So, this is your expectation and you want to know you want to talk about  $X$  minus the expectation of  $X$  ok. So, that can fall anywhere on this line and you are in particularly interested in the event that this  $X$  minus  $E$  of  $X$  is greater than  $a$ .

So, its it has to be greater than  $a$  or because it is we consider the absolute value  $X$  minus  $E$  of  $X$  has to be less than  $a$  these are the two things that we care about and that is why these areas under these shaded portions is what we care about. And that tail is given to be the variance of the random variable  $X$  divided by  $a$  square; so, this is Chebyshev's inequality.

(Refer Slide Time: 19:48)

**Chebyshev's Inequality**

Let  $X$  be any random variable. For any  $a > 0$ .

$$\Pr(|X - E[X]| \geq a) \leq \frac{\text{Var}[X]}{a^2}.$$

Proof.

$$\Pr(|X - E[X]| \geq a) = \Pr((X - E[X])^2 \geq a^2)$$
$$\leq \frac{E[(X - E[X])^2]}{a^2}$$
$$= \frac{\text{Var}[X]}{a^2}. \quad \blacksquare$$

Markov's Inequality

Positive

And it is not hard to prove this. So, let us see how this can be proved; so, this is what we want it is just straight from here. And inside this probability we have this  $X$  this  $E[X]$  minus  $E$  of  $X$  absolute value is greater than or equal to  $a$ . Now just square the terms on both sides of this inequality of this event ok; it is not going to change the event at all. In fact, it has a nice property that it gets rid of the absolute and this square on the left side is going to be non negative ok.

And so, this  $X$  my now what we have is the probability this term becomes the probability that  $X$  minus  $E$  of  $X$  the whole square is greater than or equal to a square. And because this is a positive term the non negative term, we can simply apply Markov's inequality and you can now see that all I mean this state inequality; this more fancy looking tail inequality essentially depends on Markov's inequality.

So, this is a positive term and there I an a squared I mean a parameter here a square in this context. So, then we can apply Markov's inequality and we get the expectation of this is by the way random variable right. So, the expectation of that random variable divided by a square that is this Markov's inequality and what is the expectation of  $X$  minus  $E$  of  $X$  the whole square? That is nothing, but the variance of  $X$  that is just another formula for the variance of  $X$  and so, it is going to be variance of  $X$  divided by a square which is exactly Chebyshev's inequality ok

(Refer Slide Time: 21:50)

Useful Alternative Forms

For any  $t > 1$ ,

$$\Pr(|X - E[X]| \geq t \cdot E[X]) \leq \frac{\text{Var}[X]}{t^2 E[X]^2}$$

std. deviation

Define:  $\sigma[X] = \sqrt{\text{Var}[X]}$ . Advantage: same unit as  $X$ .

$$\Pr(|X - E[X]| \geq t \cdot \sigma[X]) \leq \frac{1}{t^2}$$

And the same inequality can be slightly rewritten. So, you can rewrite it in a couple of forms; so, here is one form. So, if you write this tail as  $t$  times the expectation of  $X$  you simply get the in over here it is going to be the square of whatever you put in over here.

So, it is going to be variance of  $X$  divided by  $t$  squared times  $E$  of  $X$  the whole square. And now another form let us define a term called the standard deviation you know often denoted sigma of  $X$  this is nothing, but the standard deviation ok. And that is nothing, but the square root of the variance of the random variable. The nice thing about the standard deviation is it is in the same units as that of the random variable  $X$  suppose  $X$  measures say distance in meters variance is  $E$  of  $X$  squared minus  $E$  of  $X$  the whole square.

So, when you think about it the unit is going to be beta squared whereas, the standard deviation is going to take a square root of the variance and therefore, it is going to bring it back to meters ok. So, that is one of the nice features about the standard deviation and so, it is quite commonly used in practice and the common question people ask for random variables is you know well whether they relate to how far I mean how large standard deviation is because that tells you how much the random variable is likely to deviate from its expectation.

So, what is the probability that a random variable will deviate from its expectation by more than  $t$  times the standard deviation of  $X$ ; that is going to be at most  $1$  over  $t$  squared again simply by applying the Chebyshev's inequality. Now here it will be what will we

have over here this will become variance of X divided by t squared times sigma of X the whole squared, but that is essentially just variance is nothing but sigma of X the whole squared. And so, those two cancel out you will be left with 1 over t squared that is how you get this formula this inequality.

(Refer Slide Time: 24:44)

The slide features the following content:

- Title:** Back to the Binomial Distribution
- Equation 1:**  $X \sim B(10000, 0.5)$
- Equation 2:**  $E[X] = 5000$
- Equation 3:**  $Var[X] = 2500, \sigma[X] = 50$
- Probability Statement:**  $Pr(|X - E[X]| \geq 3 \times 50) \leq 1/9$
- Annotation:** "Much better than Markov's" with an arrow pointing to the  $1/9$  result.
- Visuals:** A small circular logo in the top right and a video inset of a man speaking in the bottom right.

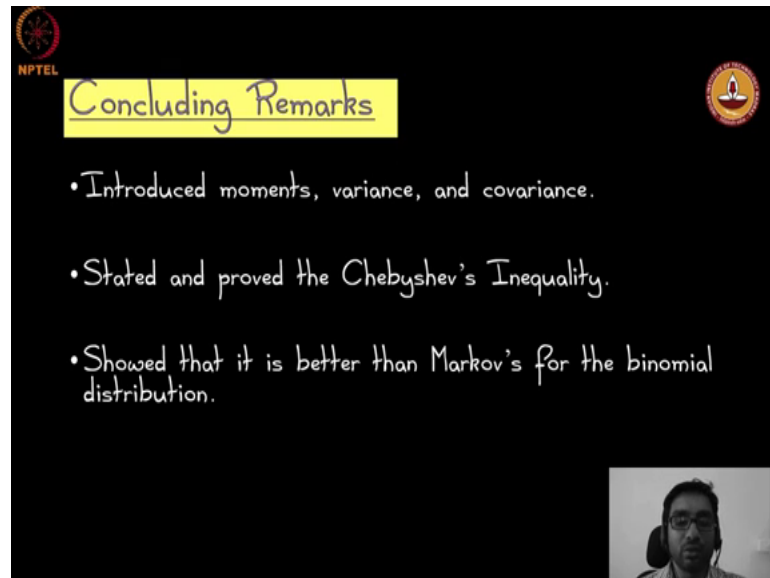
So, let us go back to the binomial distribution and test out this new tail bound that we have this Chebyshev's inequality. Remember that in our experiment in the last segment we drew several samples from the binomial distribution with parameters 10000 and bias 0.5 in particular I mean. So, this means that the expectation is going to be 5000 and you can work out that the variance is going to be 2500 which means that the standard deviation is going to be 50 just 2500 square root of that ok.

So, now you can ask what is the probability that X is going to deviate from the expectation by more than a 150 ok. If you recall there was almost never I mean the number that that we drew from the binomial distribution almost never went beyond 5000 650 I mean.

So, sorry 5150 or it almost never went below 4850. So, it was within the plus or minus 150 range and that is what we are asking and if you work it out it this is going to turn out to be 1 over t squared. So, t here is 3 and that is going to be 1 over 9 and this is a much better more accurate probability and Chebyshev's is clearly more powerful than mark

curves in this context, but in a in a strange way it is essentially markers applied in a more appropriate way that is exactly what Chebyshev's says ok.

(Refer Slide Time: 26:40)

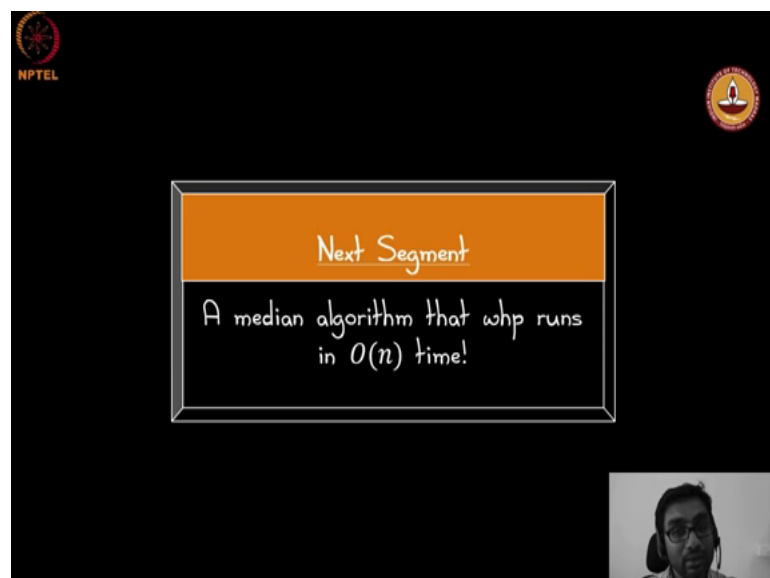


The slide features a black background with a yellow highlight box for the title 'Concluding Remarks'. The text is written in a white, handwritten-style font. In the top left corner, there is a small red and white NPTEL logo. In the top right corner, there is a circular logo with a lamp. A small video inset in the bottom right corner shows a man with glasses and a headset.

- Introduced moments, variance, and covariance.
- Stated and proved the Chebyshev's Inequality.
- Showed that it is better than Markov's for the binomial distribution.

So, with that we come to the end of our segment on Chebyshev's inequality ah. What we did was we introduced some definitions and some terms that were crucial to understanding Chebyshev's inequality. And we stated and prove it and we showed that at least in the context of binomial random variables it is it is a better way to bound the random variable than Markov's inequality ok.

(Refer Slide Time: 27:14)



The slide features a black background with a white-bordered box containing an orange header and white text. In the top left corner, there is a small red and white NPTEL logo. In the top right corner, there is a circular logo with a lamp. A small video inset in the bottom right corner shows a man with glasses and a headset.

Next Segment

A median algorithm that whp runs in  $O(n)$  time!

And in the next segment, we are going to go back to the problem of finding I mean of selecting the  $k$  element. In fact, it stays a little bit simple well just be interested in finding the median of a given array of numbers, but this time the good thing is we are going to ensure that the running time is  $O(n)$  with high probability not just on expectation, but with high probability. So, that is that is going to be the focus of next segment.

Thank you.