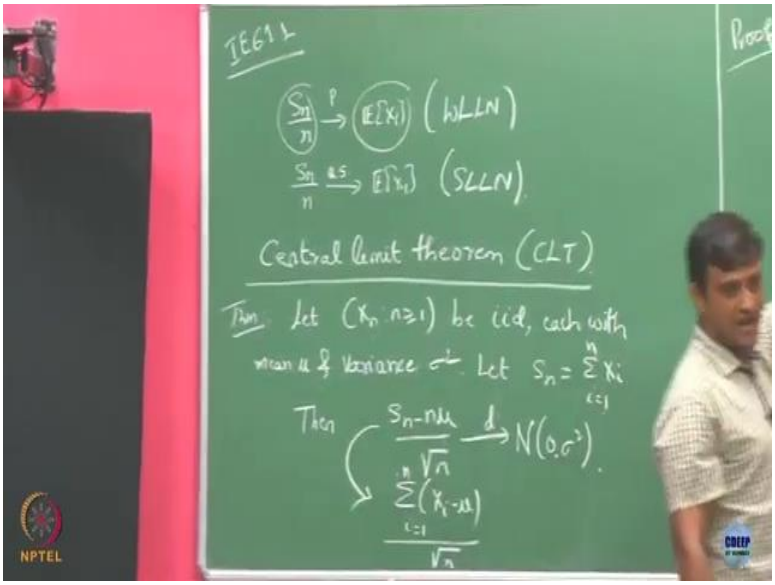
Introduction to Stochastic Processes
Professor Manjesh Hanawal
Department of mathematics
Indian Institute of Technology Bombay
Lecture 27
Central Limit Theorem

So let us start our discussions. So in the last class, we started discussing about converges of some of random variables. So we said Sn is some of n random variables and then we were interested in the case where Sn by n converges in to what value.

(Refer Slide Time: 0:52)



So we noticed that, so if a sequence of random variables which are IID, then this value of Sn by n converges to expected value of the random variables, which is expectation of x1.

So notice that here what we are saying is the limiting random variable is a constant and what is that constant? That constant is the mean value of the random variables. So this we call it as the weak law of large numbers. And we could also show that this is indeed converges also almost shows sense, but we put it only under the special case that the fourth moment is finite and this function we call as strong law of large numbers.

So this limiting distributions are very important in applications. So can somebody think of why it is so? So why these results are, I am saying they are very useful. Can you think of an example why this should be useful?

Student: To predict the future.

Professor: To predict the future, what future?

Student: (())(2:23).

Professor: But what distribution you are talking about. I am saying it is an IID sequence. Every time it will have the same distribution or to predict there. What you are saying is what is Sn by n, this Sn is the sum of the random variables and n. So like he give an example of fertility rate, let us say number of children per woman. We have it for many, many women and we want to calculate what is the average number of children per woman. What you can do?

You can just count take many, many samples. Talk to many, many women and take a survey and just and then average then, if you have taken sufficiently many samples, maybe this is going to converge to that value, an another simple case could be, Let us say the voting pattern. So everybody has a preference for one party or the other, let us call party A and party B.

And now you want to see that, what is the, with which probability each guy is going to prefer? Let us say all of them has the same kind of preference for like if I have preference for party A with probability P and preference for party B, probability minus one minus P, let us say that if it is same across all the people, how I am going to calculate that value P?

I will just ask everybody, tell me what is your preference? You prefer party A or party B and then just I will take large enough samples and then just do the average, then I am going to get this value. So you might say seen like all this after elections many people will just say, this guy is going to get these many seats and this many this and they are going to say this is going to happen in certain probability.

How they will do, maybe they will just mostly some variant of this results they are going to apply and usually when they say this, they are going to say what is the sample size they use to make this list? They say, we talked 100 people. And based on this our survey says that this party is going to win this with this probability. So and most of the times these surveys are believable if their sample size is large, that they have got this input from more people?

If their surveys like if they have just got and talked to hundred people and from that they are concluding with it, we may not take it seriously. But if they have talked to maybe like ten

thousand one lakh people, maybe that is more reliable because our research says that, this average converges to this and this convergence is more accurate if my n is large.

And similarly like as simple as if you want to decide what is the success rate of a coin, what is the probability that when I drop a coin, it comes with head. How you are going to do, how you are going to find it. So let us say I will give you a coin and it has some bias for coming up head, and you want to compute this head. What is this probability? How you are going to do that?

You just keep throwing it and see that how many times head comes map that head to let us say one and map tails to zero. Just count how many times you get one and then divided by N then this is telling that that should go close to as you take more and more samples that should be go to that value P. So whenever we have such things to be done whenever we have a kind of parameter estimations, you are going to do this.

So suppose let us say all my random variables are exponentially distributed random variables x1, x2 the sequence is IID, but each random variable is exponentially distributed with parameter lambda. If I take many, many samples of them, let us say I do experiments, let us say ten thousand times, get the samples of them and then do the average. What will this value be? What is that average will be?

Let us say I have an experiment which has a distribution, which is exponential. I am going to repeat it many times. Each time I am going to repeat it independent of the previous one. Whatever the sample I got first time. Let me call that as x1. Whatever the sample I got it second time let us call it x2 like that. Like this I have ten thousand samples.

If I just all add them 10000 samples and divide it by 10000 approximately what is the value that you expect, one by lambda. So in this case suppose if apriority know that my distribution are exponential, but you do not know what is the rate. Then what you can just do is do this experiment many, many times and then take average, you already got close value of lambda.

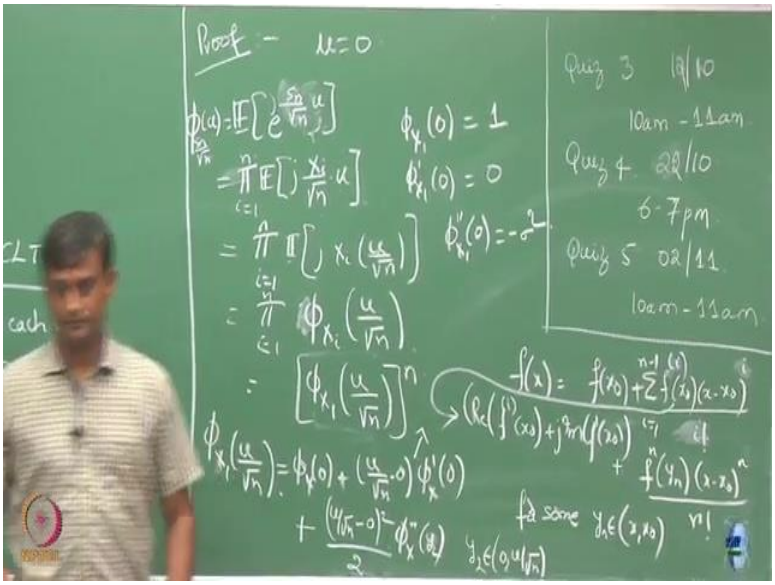
So for example, all this you might have studied in class 12, like life of atoms. There will be like kind of exponentially distributed. But you do not know the rate at which they are going to decay. Then maybe you collect many, many samples and doing this, you can estimate their rate, the lambda rate. Today, we are going to see one more important result of this converges of some of random variables called Central Limit Theorem.

First, let me write this result and then we will discuss. So what it says let us say the sequence of random variable that are IID, and they have this mean and variance Sigma Square. If I look at there, sum like this and then look at that normalized sum. So I am going to call this normalized sum because what is the mean of Sn, Nu.

So I am basically subtyping mean already from them in a way it is centered because each of them are now it is a zero mean random variable Sn minus nu and then I am normalizing it by square root n. So this is Centered and normalized random variable here and we say that this goes to whatever sequences is, whatever distribution it has if it is IID then it goes converges and distribution to normal distribution with variances sigma square.

So what does this result say? Yes. Now say earlier when I did it in Sn by n just took the average. It always converges to a fixed constant. But now here the limit is not a fixed constant. It is a distribution, which is caution, the parameter zero and Sigma Square. So let us try to understand why this result is true. So to this result, the characteristic function comes to her help and we just use it here.

(Refer Slide Time: 12:04)



Let us say, so again, I am going to approve this result, assuming u to zero. If u is not equals to zero, you can straightforwardly still conclude this by saying that, so if mu's are zero. This is Sn by n. So if mu's are not zero, I could as well write the numerator as Xi minus mu i at one to n. So I have just replaced Sn minus nu by Xi minus Mu. So this is by definition.

Now each of this random variable Xi minus mu is I can treat this as a new random variable Xi minus Mu. What is the mean of this random variable, zero. So then what I am basically

adding is n random variable where each one of them has zero mean and still their distributions are same. So because of that I can as well assume that the mu's are all zero and try proceed through this theorem.

Now So because of that, I just need to worry about Sn by Mu n and let us try to find out what is the characteristic function of Sn by n through the point u. So this is a characteristic function at u, so this is Sn by square root n. And this is a characteristic function of Sn by square root n. Now, if you just expand this guy, what you are going to get.

So Sn is sum of n random variable, where each of them is independent. So I am going to get it as product of J Xi square root n of u and this i from one to n. And then what I am basically doing now next is, so now I am going to treat it as characteristic function of Xi computed value u by square root n. That is my argument here so I can write it as i one to n, expectation of j Xi and u by square root n so which is nothing but.

And now because these are all identically distributed. I could write this as well as phi of x1, u by square root of n to the power n. So notice that if they have identical distribution there characteristic functions should be necessarily the same. So that is why we have used this property. All of them should have the same character distribution of that of x1.

Now let us try to understand how this guy behaves. So now I am going to use this characterization. So how many of you know Taylor expansion of a function? So let us say I have a function f of x. And I want to expand to a Taylor expansion of this function around a point x0. So then I can write it as, so in general I can write it as and what is this, this to the power?

And what is the n indexing happens here?

Student: 1 to infinity.

Professor: 1 to infinity but there is one alternate characterization of this stellar function, which says that I could do this as instead of taking the infinite sequence? I will look at the truncated version of this. What does they say, then the nth one is going to be…

So what should be here, it should be i here. What is this Yn?

Student: (())(17:58).

Professor: For some Yn belongs to x and x0. So this is the truncated version of my Taylor series. But to see this, so you understand what I mean by this superscript i here?

Student: ith variable.

Professor: This is ith derivative of this function F? Now this stellar function involves the first derivate, second derivative and higher order derivatives. But we will be interested only first and second derivative because I want to I have only first order and second order statistics here and when I first order and second order statistics, my moment generating function, I need to only care about first derivative and the second derivative.

So to write this let us say what is the value of this function at zero, phi of X1 at zero when the argument is zero. Whatever the argument is for any random variable x1 this guy is going to be what one and what will be first derivative of this at zero? So what is I am looking at? The first derivative of the characteristic function x1 at the point zero. What is this?

So should you differentiate the characteristic function and put that argument 0. What you are going to get?

Student: (())(19:50).

Professor: You get 0. So how do you are going to get the compute the mean of a random variable from the characteristic function?

Student: (())(20:03).

Professor: I cannot hear you.

Student: (())(20:07).

Professor: Expectation of…

Student: (())(20:10).

Yeah. I mean, I am asking how you are going to derive the mean of a random variable from if you know its characteristic function.

Student: (())(20:18).

Professor: You take first derivative

Student: Zero.

Professor: Zero. So what exactly I am doing, I took the first derivative and put zero.

Student: Expectation of (())(20:28).

Professor: And that I have assumed to be zero. That should be mu. And that I have assumed to be zero here. And what is if I take the second derivative and put zero? What I am going to get, expectation of x1 square. But in this case, our mean is zero. So it is variance, but with what sign?

Student: It is negative.

Professor: It is going to be negative with a negative sign. Just check it when second derivate and because of this complex term involved there you are going to as sig minus Sigma Square.

So now for this characteristic function, I know what is the mean value. I mean, I have been able to express this the mean value in terms of the first derivative and the second derivative of the characteristic function and I know these things. So let us try to do that air expansion of this only for n equals to 2.

Now and I am going to do the Taylor expansion of this function around origin around 0. So then what will be the Taylor expansion of this? So it is going to be of x1 of 0 plus…

Student: (())(22:00).

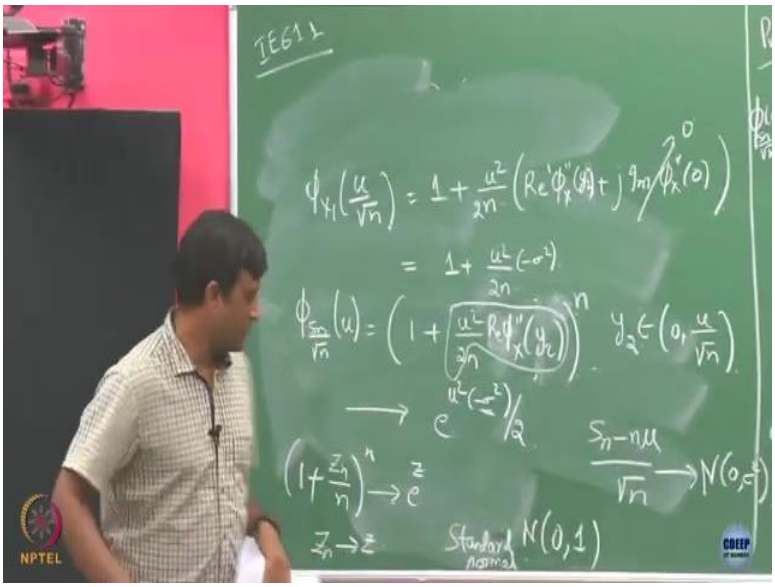
Professor: u of minus 1 times I am just writing, expanding for you. But I could have as such skipped this. And here it is going to be not zero. But a value y2 here, which is going to be dependent on where that value of y2 will lie? 0 and u by n.

It this correct? So when I wrote this Taylor series expansion, I kind of expanded it, assuming that my function F is real here. But here my phi function could be all this real or it could be complex as well. It could be complex as well. So when I have a complex function. All I need to do is this guy is here is wherever I have this phi i x naught.

I have to make it as a real part of f of i x naught plus j times imaginary part of x0. Then my function F is a complex function. So this guy here. I have to replace it like this. because my function phi is complex. Now let us look at this. This phi x bar so for the phi of X1 its derivative at zero is always it is going to be zero either it is real or both it is real and

imaginary parts are going to be zero. So because of that, this guy is going to get zero. Now, what remains is, this part.

(Refer Slide Time: 24.38)



So what we have is phi of X1 of u of root n is going to be, again phi of x0 is one plus this is going to be u square by n and 2 there in the whether 2 there and this one I have to write it as real part of phi of x double prime 0 plus j of imaginary part of phi of X double prime zero.

So you see this when I write this phi of x double prime, its value is minus sigma square. So it has only real part. There is no imaginary part in this. So because of that, I will ignore this part. What is this going to be, one plus u square two n and what is this part? This is simply going to be minus sigma square.

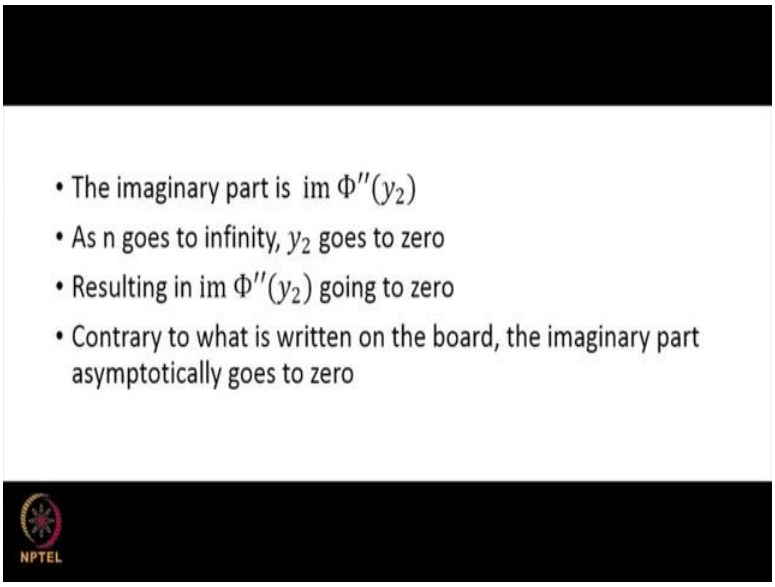
I just substituted the value for this. So what I could get is only this part, but what I am interested in is this quantity raised to the power n. So that is what my phi of Sn by n. So now let us go back to that phi of Sn by n at u is this whole quantity, this side? The real part of this is what? It should be for y2, so just let me correct this.

What is this guy is going to be? This is going to be yn here, which is going to be zero. Yeah, so I want this to be exactly this. This is fine, so where is this y2 is going to be u upon divided by n because that is what our that is what we said here, y2 belonging to zero to u by n and this is going to be still the real part of phi of X double prime y2.

fine. Now and there is a power n here. So now can you look into the sequence and try to see, you know. Are you familiar with this sequence as n goes to infinity what this sequence

converges to? E raise to this guy. But this itself is changing. This is as n if you let n go to infinity. This itself is changing and what this y2 goes to zero. And what is the value of this phi of double prime at zero? Is minus Sigma Square.

(Refer Slide Time: 29:00)



And then what is the can you then what as this n goes to infinity where this limit will go to? E raise to, it is going to u square. So you people understand why I got minus sigma square here. Now so yesterday we had discussed that if I have a sequence like this, it goes to e to the power z where Zn converges to Z.

If I had a sequence like that Zn converges to Z, then this sequence converges to e to the powers z. Now, look at this. Look at this part, think it has Zn. u square by 2 remains like that, as n goes to infinity. This guy goes to phi x Prime double prime at zero. And that value what is its value? That value is exactly minus sigma square.

And that is why we get u square by 2, this minus sigma square. Now this characteristic function has converge to this characteristic function, this function and what is this function?

Student: (())(30:58).

Professor: What is this? This is a characteristic function of a standard normal distribution. That means this has zero mean and variance sigma square.

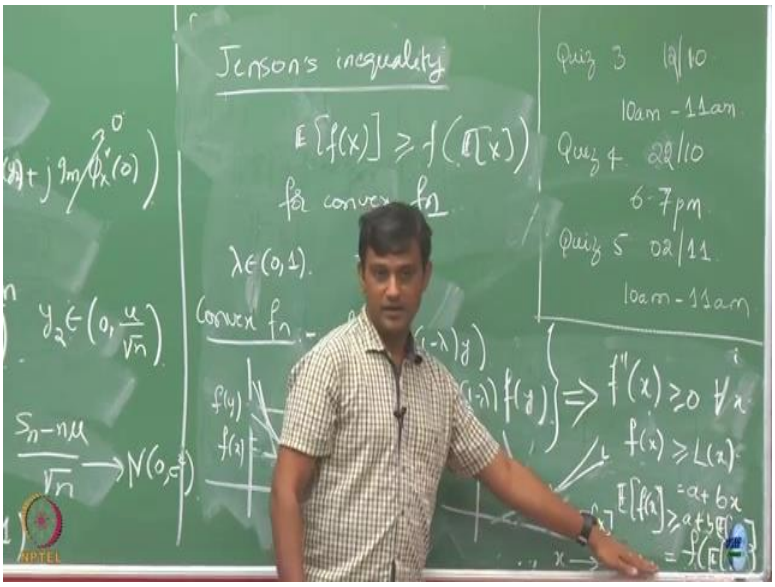
So, this is a this is converging to a distribution, which is which has mean zero and variance Sigma Square. So that is why we are saying that my Sn, so if that is the case, we know that. So one more terminology.

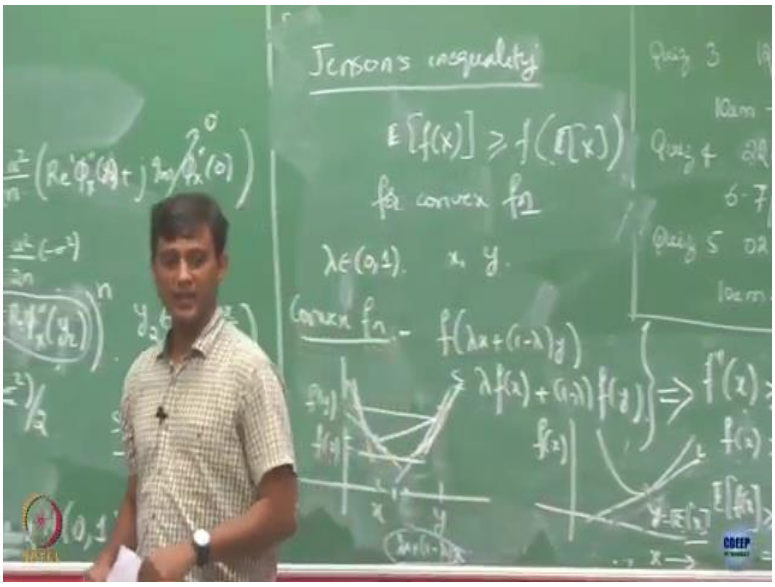
So in general, we are going to call if it is variances with zero and mean is one. Then we call it as standard normal. So in this case it is just going to mean 0 and variance sigma square. So is it clear? Now, so we will come to see you like fine. What is this? Why this center limited theorem is important. So in a way it says that you have aggregated large samples.

And if you center them and normalize by square root n that aggregation, you can treat it to be like Gaussian distributed if your n is sufficiently large. So whatever be the Xi distributions I do not care Xi all that could be exponentially distributed or all of them could be uniform distributed, whatever they are.

Then if I am going to look at their aggregation, normalized aggregation, then for, if I take sufficiently many samples to aggregate, they will behave as if they are normal distributed. So in a way I kind of already know how this with sufficiently many samples, how that distribution going to behave like, fine.

(Refer Slide Time: 33:41)





One last thing I want to cover on this topic is Jensen's inequality and the (())(33:47), let us quickly discuss the Jensen's inequality. So how many of you know Jensen's inequality? What is Jensen's inequality? Expectation of fx. What is this F? convex function, what is a convex function then? You know what is a convex function. Yes defined it.

So let us take any lambda between 0 1 and let us take 2 values. So if I take so I am defining a convex function now. So if I am going to take lambda X plus one minus lambda Y. What is this going to be? So if a functions satisfies this for all lambda and X, Y where X, Y are in the range of the function sorry domain of the function.

Then we are going to call this as a convex function. So how does the convex function look like? Let us take two points X here and let us say Y here. So what is this guy here? And what is this guy here? So where is lambda X and one minus lambda Y is going to lie? And where is Lambda f of x plus one minus lambda f of y is going to lie?

Student: (())(36:21)

Professor: So this is a linear combination of the point f of x and f of y. So f of x is here, f of y is here this is going to be the linear combination of them and what is the function value at this guy? The curve falling, yes, so let us say may be just to be more careful. Let us say this is my lambda X, one minus lambda Y. So what is my function value at this point?

This value and what is my lambda Fx in to one minus lambda Fy? It is going to be on that line, and what is we are saying? Whatever the value of this, this is a limited this value I am going to get that is going to be smaller than the linear combination of these two points. So if I

have a convex function, it is always the case that my function value, like the linear combination of those two points is going to be smaller than the linear combination of the function itself has those two points.

Now the question is if my function F is like this, why it is a case that this is true? So now we understand how the convex function looks like. So that is a consequence of this definition. So I know that. So there are properties one can think of for any convex function for all X. So why this is true, actually?

So I have a connect function like this. So if you start looking at the slope of this function, so the slope is somewhere here at this point. And if I look at the slope it will be like this. And at this point, slope is going to be zero. At this point, the slope is going to be like this. At this point, the slope is going to be like this.

So what is this slope here? It is negative. And here this is again, negative. But this is having a larger negative slope at this point. What is the slope of this?

Student: Zero.

Professor: And what is the slope of this guy?

Student: (())(39:14).

Professor: And what is the slope of this?

Student: More positive.

Professor: More positive so what is happening? The slope itself is increasing. So if the slope is increasing means what there is double derivative is increasing. So that is one property. Another property of this convex function is, at every point I can draw a tangent that touches this function at one point. so I can draw a linear line at any give me any point on this function.

I can draw a linear line that acts as a tangent to this function f. So from this property, actually we can write away like this. So let us see why that is true. So take a point x, some point let me take this X. Let me take this Y. So this is X and this is my function f of x at given point Y I can draw a tangent here. That means this line here. Here is a lower bound on my function. So what does this mean? I think maybe this much space is enough.

So f of x is going to be greater than or equals to L of x and this is true for all x. Because this line is take any point this line is going to be lie below this function F? Now let us take expectation of this functions and in particular I would be interested in taking this Y to be expectation of X. So expectation of X is going to be somewhere in the middle of my domain.

So let us take that point to be f y here. And on this point, we know that this function and this function both at the same value. So what is this L of x it will have some its own slope and it will be like this. Now, let us take an expectation of this. And this is going to be A plus B, an expectation of X. Is this true? I just upload expectations to both sides.

And what is this? At this point this is the line A plus Bx? This is A plus Bx. Now if I complete this function at the point Y, which is expectation of X at that point this value is going to be same as my function f of x. So this is my f.

So on this this is simply the expectation of X. So I am up on this property because this line at the point, So what does this mean? A plus B into expectation of X means this is nothing but this linear function computed at point X which is nothing but the expectation of X.

So this is the linear line which is computed at different point A plus Bx. But now if I am saying this is A plus the expectation of X that means this function being computed at expectation of X that expectation at this point is A plus B is nothing but f of expectation of X. So that is why expectation of x of f is going to be greater than or equal to be f of expectation of X.

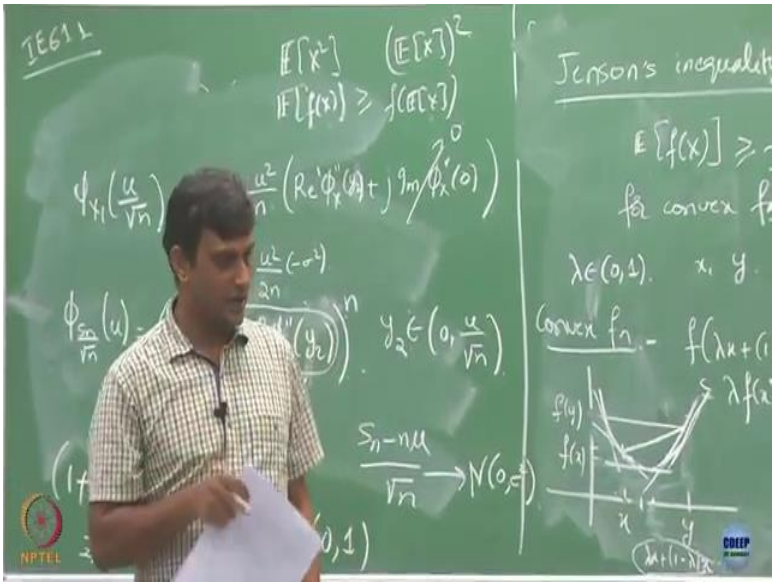
So this is exactly is what we said as Jensen's inequality.

Student: L is the tangent.

Professor: Yes L is the tangent. At this point, I have chosen this tangent to be at a particular point, which is the expectation of X here. I know. Give me any point on this curve that convexity properties tells that at this point, I should be able to draw a tangent.

What I did is I chose a point Y which is the expectation of X and on that point, I have a linear function like this. And I know that this linear function, always lower bounds my function f of x. So that is the first property I used here and then just Make sure that this F and this L function, they have the same value at expectation of X, so that is why Jensen's inequality holds.

(Refer Slide Time: 44:59)



So because of this, you can some derive some of the properties very quickly. For example I want to find a relation between expectation of X square and expectation of X whole square, which is bigger. Expectation of X squarer is greater than this, why?

Student: (())(45:14).

Professor: Can you use Jenson's inequality?

Student: (())(45:20).

Professor: So what function you are going to use here? X squared. So I can think of, this as f here, where f is X square. So if I had to do this, then I know that this is nothing but expectation of X. And F is nothing but square function. So that is there. And as you said that we already know that is true because variance is a non-negative quantity and variance is the difference of expectation of X square minus expectation of X whole square.

So like this, this comes handy to prove this Jensen's inequality comes handy to prove many inequalities of this part. Some of them are already there in our exercise and your assignment. Try to solve them.