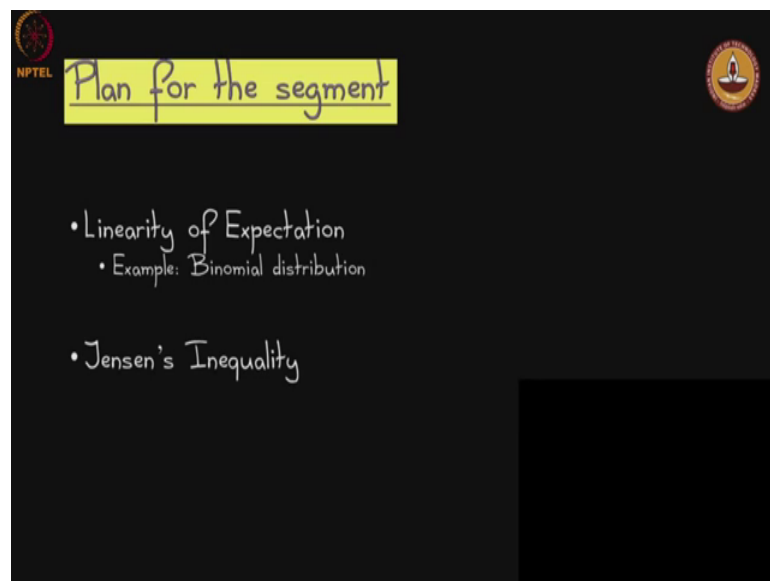


Probability & Computing
Prof. John Augustine
Department of Computer Science and Engineering
Indian Institute of Technology, Madras

Module – 02
Discrete Random Variables
Lecture - 10
Segment 2: Linearity of Expectation & Jensens Inequality

Ok. So, now, we can we start the second segment in module 2, where we will be continuing our discussion on random variables, and we will be looking at two important properties of expectation, one is the Linearity of expectation and the other is called the Jensens inequality ok.

(Refer Slide Time: 00:37)



Linearity of expectation we will study it with the binary will. So, in studying that we will introduce this distribution called binomial distribution, it can a very fundamental distribution shows up heavily in computing. And we will introduce Jensens inequality by trying to understand what the area of a square random square is.

So, let us start with the linearity of expectation, it is a very useful theorem let me state it.

(Refer Slide Time: 01:07)

Theorem: Linearity of Expectation

Let X and Y be random variables with finite expectations and let a and b be arbitrary constants. Then,

$$E[aX + bY] = aE[X] + bE[Y].$$

Proof.

$$E[aX + bY] = \sum_x \sum_y (ax + by)(\Pr(X = x) \cap \Pr(Y = y))$$

The slide also features the NPTEL logo in the top left and a circular logo in the top right. A small video inset in the bottom right shows a lecturer speaking.

So, let us consider two random variables X and Y and they have finite expectations, they do not need to be finite in the sample space, but their expectations must be finite and a and b are some arbitrary constants. So, we are interested in what is the expectation.

So, now when you think about it, this itself this aX plus bY itself is a random variable. You can take two random variables and you can add them you can apply some functions on them and outcome will again be a random variable. Why because it is just a composition of functions, remember X itself is a function, Y itself is a function you can you compose them and you get this new function on the sample space, it is a function on the sample space and therefore, it is a it is a random variable ok.

So, this expectation E of aX plus bY is a random variable. So, you can ask what is the what is the expectation of this random variable aX plus bY . And as it turns out you have you it equals a times E of X the expectation of X , plus b times expectation of Y . So, this is a most natural thing that you can you would suspect it is value to be and that is exactly what we get ok. What is surprising is that, later on we will see other quantities other notion and other ways to understand the random variable for which this would not hold unless you have other conditions like independence of random variables and things like that. So, here in this theorem the important thing is there is no restriction, it is any two random variables X and Y you take expectation of aX plus bY , you get a times E of X plus b times E of Y ok.

So, this is the linearity of expectation and the proof is also quite simple what is the expectation of a X plus $b Y$? Now what you have to do is consider all possible values of X and Y . So, possible values of X is lowercase x , possible values of Y is lowercase y and it is just the sum the weighted sum of the those values right. So, $a x$ plus $b y$ weighted by the probabilities remember a and b are constants.

So, they do not have probabilities associated with them, but we can take what is the probability that random variable X takes this little x and the random variable Y takes this little y and you just sum up over all possible values that x and y can take and we will play with the summation.

(Refer Slide Time: 03:41)

Theorem: Linearity of Expectation

$$= a \sum_x x \sum_y (\Pr(X = x) \cap \Pr(Y = y)) +$$

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

Now, what we do? We arrange it so that, what happens over here is we put the summation x first in the summation y first and ah. So, basically what we do here is we when we take this probabilistic term and multiply it by $a x$ then probabilistic term multiply it by $b y$ that is the two terms that we get, but then when we write it down, we put the in the first term we put the x coming first. So, that what the advantage this has is that this x should be inside over here.





But it can come out of the y we summing over y the x is going to stay common. So, you can bring it out, and a is going to be common throughout it it is not going to be affected by the values that x takes all the values at y take. So, a comes out all the way outside the summation x comes out, but cannot come out of a summation of x , it can come out of the

summation of y and then you have the probability inside because this probability is dependent on y as well. So, you cannot bring it out of this summation.

So, that is the first term that you have over here, and then the second term similarly you teach the constant b comes out all the way, in the summation you put the y first and then the x comes. So, the y is able to come out of the inner summation, but it cannot come out of the outer summation. So, this is a very straightforward thing to do anything about it.

Now, what about this quantity? A and you think about it, the x equal to x is one even, and you are summing that intersected with y equal to y for all possible values of y what does that remind you of? it is the law of total probability, you the y equal to y is going to arrange the entire sample space.

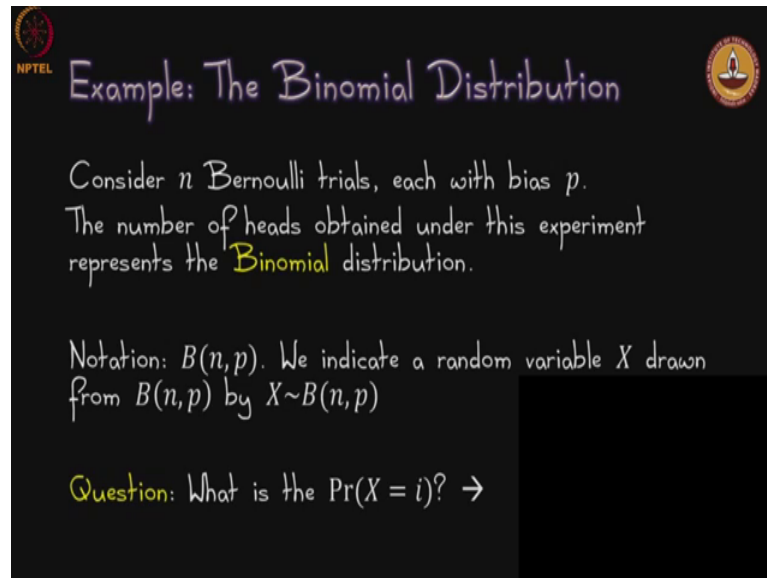
(Refer Slide Time: 05:40)


$$= a \sum_x x \Pr(X = x) + b \sum_y y \Pr(Y = y)$$
$$= aE[X] + bE[Y]$$


So, what do you get out of this? This is nothing, but the probability that X equal to x and similarly you get probability that Y equal to y over here ok, but of course, now what is this quantity? This is nothing, but the definition of the expectation of x ok. So, that is what you get over here, similarly this is the definition of the expectation of y you get a times expectation of X plus b times expectation of Y and that ends the proof of this theorem.

So, we will try to now apply this theorem um. So, very simple application, but nevertheless you know please bear with me, because simplicity means it is important and this is generally going to be true, when things have a very simple clear explanation; that means, they will show up again and again and again. So, there; that means, is important ok.

(Refer Slide Time: 06:36)



The slide features a dark background with white and yellow handwritten text. In the top left corner, there is a small red circular logo with the text 'NPTEL'. In the top right corner, there is a circular logo with a lamp. The main text reads: 'Example: The Binomial Distribution', 'Consider n Bernoulli trials, each with bias p . The number of heads obtained under this experiment represents the Binomial distribution.', 'Notation: $B(n, p)$. We indicate a random variable X drawn from $B(n, p)$ by $X \sim B(n, p)$ ', and 'Question: What is the $\Pr(X = i)$? \rightarrow '.

So, the binomial distribution simply this. So, you consider n Bernoulli trials each with bias p we say bias p is the success happens with probability p or you can think of it as heads occurring with probability p , the number of heads obtained under this experiment represents the binomial distribution. So, you toss a coin n times biased coin with probability p of heads count the number of times the heads appears that is the binomial distribution. And you should denoted the $B(n, p)$ the two parameters that define this distribution are the number of times it tosses the coin n , and the probability p ok. And when we think of a random variable x that has this binomial distribution we denoted as X drawn from $B(n, p)$ that is the notation we use $X \sim B(n, p)$ ok.

Let us gets a (Refer Time: 07:27) what is the probability that X equal to i ? It's very easy to see how x what are we asking. So, X can take the values either zero when no heads appears or n when all the tosses outcome are heads, we pick a particular value i and ask what is the probability that X can take that value i .

(Refer Slide Time: 07:49)

Example: The Binomial Distribution

$\Pr(X = i \mid \text{a particular set of locations for heads})$

T	T	H	T	T	H	T	H	H	T
$1-p$	$1-p$	p	$1-p$	$1-p$	p	$1-p$	p	p	$1-p$

$= p^i(1-p)^{n-i}$

How many ways?

$\binom{n}{i}$ mutually disjoint events

$\Pr(X = i) = \binom{n}{i} p^i (1-p)^{n-i}$

So, let us focus on a particular situation, where there is a set of locations in particular I locations where heads we want heads to occur. So, the rest of them tails ok.

So, if you think about it, each location where we want tails to occur what we are expecting what we want is an event with probability $1 - p$ to occur because we want tails to occur over there, and wherever we want each heads to occur we were that will that event will occur with probability p ok.

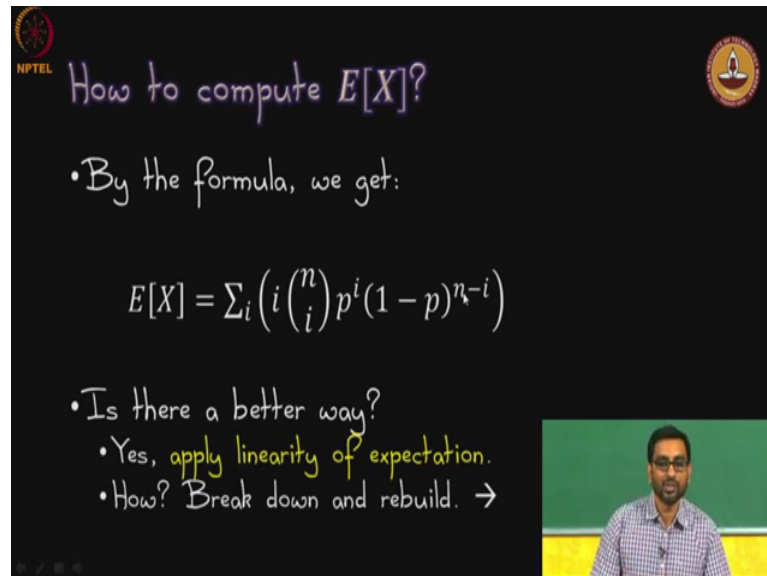
So, we just pick some I locations for H, and you put probabilities p probability p for all of them the remaining ones would be $1 - p$. So, if you if you specify the locations for the heads the outcome has the probability p raised to the I and these are all independent tosses. So, you can multiply the probabilities p raised to the I times $1 - p$ raised to the $n - I$ for all the remaining locations ok. This probability works when you have specified the locations for the heads ok.

But that; obviously, you cannot do you do not know where the heads will occur. So, you have to choose you have to find out all the we were to consider all the n choose I possible ways in which the I heads can occur right. As n choose I mutually disjoint events if it is mutually disjoint you do not multiply you add them up. So, and there are n choose I of them. So, you basically n choose I times p raised to the I times $1 - p$ raised to the $n - i$. Basically you taken p raised to the I or $1 - p$ raise to $n - i$

If you have added it and choose I times. So, that is where the multiplication comes from. So, this is the property that X equal to i .

Now, let us look at computing the expectation of this random variable.

(Refer Slide Time: 09:47)



NPTEL

How to compute $E[X]$?

- By the formula, we get:

$$E[X] = \sum_i i \binom{n}{i} p^i (1-p)^{n-i}$$

- Is there a better way?
- Yes, apply linearity of expectation.
- How? Break down and rebuild. →

NPTEL

. So, by the formula we get this. So, you basically sum over all possible values of I that is going to be I is the values that X can take. So, it can be 0, when no heads occurs or n . So, when all of them are heads and you sum over all i 's, but then weighted by their individual probabilities, and I do not know how to deal with this summation. I am sure there is some way you can get this to work, but it is a little bit messy ok. So, what saves the day for us is that we can apply linearity of expectation.

(Refer Slide Time: 10:42)

The slide features a black background with white and yellow text. At the top left is the NPTEL logo, and at the top right is the Indian Institute of Technology Bombay logo. The title 'How to compute $E[X]$?' is written in a purple, handwritten-style font. Below the title, there are two bullet points. The first bullet point states: 'Let $X_i = 1$ if i th toss is 1. Otherwise, $X_i = 0$.' followed by the equation $E[X_i] = p$ in yellow. The second bullet point states: 'Notice that $X = X_1 + X_2 + \dots + X_n$. Thus,' followed by the equation $E[X] = E[X_1 + X_2 + \dots + X_n]$. Below this, it says 'From linearity of expectation,' followed by the equation $E[X] = E[X_1] + E[X_2] + \dots + E[X_n]$ and $= np$. In the bottom right corner, there is a small video inset showing a man in a checkered shirt speaking in front of a green chalkboard.

So, there is a very simple way to compute the expectation of the binomial distribution, and what we do is we break down the binomial distribution into individual Bernoulli trials, and then we somehow rebuild using linearity of expectation. So, what we do let us break it down first. There are n coin tosses we use X_i equal to 1 it is a new random variable that we are defining x . In fact, we will define n such random variables X_i will be equal to 1, if the i th toss is a 1 otherwise X_i is 0 ok. And we know E of X_i is p . So, now, we know that the number the total number of heads that we going to get is simply the summation of X_1 plus X_2 and so on up to X_n basically take each one of them if you get a heads, it accounts towards this capital X otherwise not.

So, it is just a summation. So, clearly you can take expectation on both sides there. So, E of X is equal to E of X_1 plus X_2 plus and so on up to X_n and so, now, well at this point we do not know what to do with the right hand side, but if you just apply linearity of expectation immediately you know what to do with it you each one of them E of X_1 is a p , E of X_2 is a p and E of X_n is a p there are n such terms. So, the total expectation is n times p let us a linearity of expectation for you is very powerful very simple and very useful for us in the same context of random variables just look at Jensen's inequality.

So, let us motivate that by a simple example suppose you consider a random square and what do I mean by that ah.

(Refer Slide Time: 12:16)

Example: Area of a random square

- Random square with edge X UAR from $[1,99]$.
- What is the expected value of its area?
 $E[X^2]$

$E[X^2] \stackrel{?}{\leftrightarrow} E[X]^2$

The slide features the NPTEL logo in the top left and a circular logo in the top right. A video inset in the bottom right shows a lecturer with glasses and a blue shirt speaking in front of a green chalkboard.

So, the edge length X is chosen uniformly at random from the range 1 to 99. So, let us say it is an integer value um. So, questions that we can ask, what is the value expected value of it is area? So, you take X squared and you know what is you ask what is the expectation of X squared? And one question that should come to mind is E of X squared the same as E of X the whole square and in this context, we are going to just see how these two relate to each other.

(Refer Slide Time: 12:56)

Convex Function

Another perhaps easier way to think about convex functions is that the line joining any two points on the curve will always be fully above the curve.

Consider $f: \mathbb{R} \rightarrow \mathbb{R}$.
The following statements are equivalent.

- f is convex.
- For any $x_1, x_2 \in \mathbb{R}$ and $\lambda \in [0,1]$,
 $f(\lambda x_1 + (1-\lambda)x_2)$
 $\leq \lambda f(x_1) + (1-\lambda)f(x_2)$.
- $f''(x) \geq 0$.

The slide features the NPTEL logo in the top left and a circular logo in the top right. A graph on the right shows a convex curve f with points x_1 and x_2 on the x-axis, and $f(x_1)$ and $f(x_2)$ on the y-axis. A line segment connects the points on the curve, and a point on this line is labeled $\lambda x_1 + (1-\lambda)x_2$. A video inset in the bottom right shows a lecturer with glasses and a blue shirt speaking in front of a green chalkboard.

So, let us for that let us think noticed this X squared is a convex function. So, let us just be precise about what we mean by a convex function um. So, here is the definition. So, this way of defining it as three equivalent statements ah. So, f is convex, and then three other two statements equivalent statements are defined the notion of convexity is the following.

Suppose you have. So, we have shown this function f over here right. So, it is just this curve of shown over here it is convex, if for any two points say x_1 and x_2 that you choose you look at some intermediate point, and that is defined by this parameter λ . So, we take λ times x_1 plus $1 - \lambda$ times x_2 you get a point somewhere in the middle.

So, that is shown in this blue is that that is the blue expression here $\lambda x_1 + (1 - \lambda)x_2$. And now if you take you apply f for that function that is the left hand side and the claim is in the for a convex function what happens when you take the f you get to this point on the function, that is this purple point over here you compare that with the point ah.

So, now if you think of f of x_1 that is f of x_1 over here, f of x_2 is over here and then you take the corresponding point parameterize by λ in the seg between the segment in the segment connecting f of x_1 to f of x_2 . Now the comparison of these two points gives you a definition of whether the where the curve is convex or not. For convexity what you need is that this point on the curve should be less than or equal to this intermediate point by just joining there then those two points f of x_1 and f of x_2 by a line segment.

So, and that should be true for any choice of x_1 and any choice of λ ok. And many of these functions that they all of these x squared x to the 4 and all are going to be convex functions. Another way to think about it is if you take the second derivative and if the second derivative is non-negative then again you can say that the function is convex ok.

(Refer Slide Time: 15:22)

Jensen's Inequality

If f is a convex function, then,

$$E[f(X)] \geq f(E[X])$$

Proof (assuming f has a Taylor expansion).
Let $\mu = E[X]$. (Standard notation.)

$$f(x) = f(\mu) + f'(\mu)(x - \mu) + \frac{f''(\mu)(x - \mu)^2}{2}$$
$$\geq f(\mu) + f'(\mu)(x - \mu) \quad (\text{since } f''(\mu) \geq 0)$$

So, what is Jensen's inequality saying? If f is a convex function then the expectation of f of X is at least f of the expectation of X . So, this is a quite fundamental inequality that you see every once in a while and so, let us quickly look at the proof and the proof we are going to do the simple version where there is a Taylor expansion for this function the proof holds more generally just to be clear. And you can actually try it out try proving it more generally there is an exercise in the textbook which gives you a hint to prove it more generally as well, but for our purposes we will just quickly prove it under the assumption that there is a Taylor's expansion.

So, another standard notation expectation of X is often denoted as just μ when the context is clear. So, I am just going to use μ to represent expectation of X and remember μ is just essentially a scalar value. So, f of x can be written as f of μ plus f' of μ times x minus μ plus f'' of μ times x minus μ squared over two this is based on standard Taylor.

Note you take the Taylor series and just capture the first two terms as is, but then the rest of the terms are captured by this third term over here. The nice thing you have is recall that the convexity gives you that f'' the second derivative is always going to be greater than or equal to 0. So, this third term over here is going to be a positive term or at least non-negative term.

So, what do we do? We simply get rid of the third term and replace the equality by a inequality.

(Refer Slide Time: 17:25)

Jensen's Inequality

Thus,

$$\begin{aligned} E[f(X)] &\geq E[f(\mu) + f'(\mu)(X - \mu)] \\ &= E[f(\mu)] + E[f'(\mu)(X - \mu)] \\ &= f(\mu) + f'(\mu)(E[X] - E[\mu]) \\ &= f(E[X]) + f'(\mu)(\mu - \mu) \\ &= f(E[X]). \quad \blacksquare \end{aligned}$$

So, now, we apply expectation on both sides. So, basically we have f of x is greater than or equal to f of μ plus and so on we just apply expectation on both sides, on the right hand side you have expectation over a larger term I mean at two summation of two terms, you apply linearity of expectation. So, far we are doing something things that are quite straightforward.

Now, here what are we doing this f' of μ is just a scalar value ok. Remember E times a of x is a times E of X we have already seen that. So, we simply get the scalar value out over here. So, it is just f' of μ become a times E of X and E of X minus E of μ . Expectation of a constant is just a constant itself E of X is anyway just μ and then there is this f' of μ and here what are we doing we are taking expectation of f of μ is again just f of μ because f of μ is just a constant or scalar value ah, but μ is a is essentially E of X , here you of course, have this μ minus μ . So, this term cancels out you are left with f of E of X .

(Refer Slide Time: 18:51)

NPTEL

Example: Area of a random square

- Random square with edge X UFR from $[1,99]$.
- What is the expected value of its area?

$E[X^2]$

Variance $E[X^2] - E[X]^2$ measures X straying from $E[X]$.

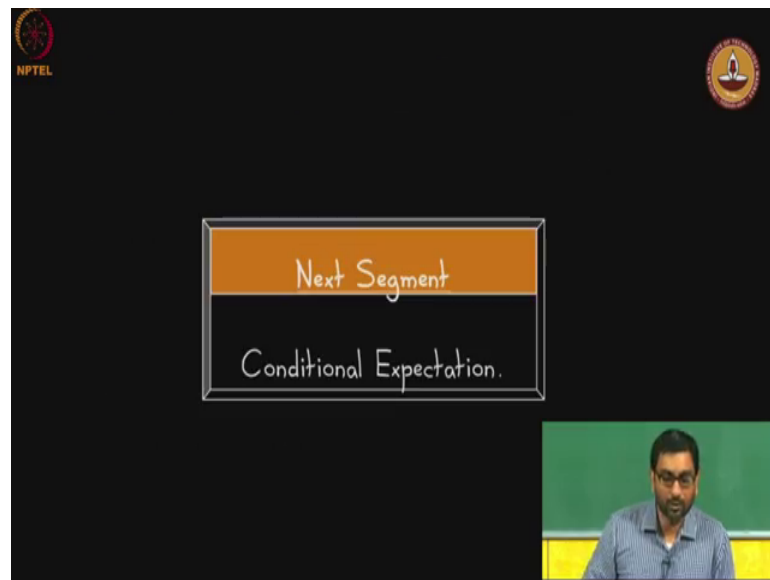
$E[X^2] \geq E[X]^2$

$\frac{9950}{3} > 50^2 = 2500$

So, this establishes the Jensen's inequality ok. Just going back to the example if you apply this what you will get is well what is E of X ? E of X is going to be 50 X ranges from 1 to 99 right. So, like E of X is going to be 50, 50 squared is something some 2500 ah, but E of X squared if you work it out is going to be a larger quantity. So, that is just an example where it works out that way.

The important thing to keep in mind is that we are actually what we are doing is, we are working our way towards understanding of some other measures. So, for example, this if you notice that E of X squared is larger than E of X the whole squared. The difference as it turns out is actually an important measure it tells you how much a random variable tends to deviate from its mean value from its expected value ok. So, that itself is an important measure.

(Refer Slide Time: 19:51)



So, with that we end this segment, where we studied the expected value of a random variable and just proved something called it and proved the linearity of expectation and now the Jensen's inequality. And so, with that we will we will have to get ready for conditional expectation, which is some more understanding of how expectation works.