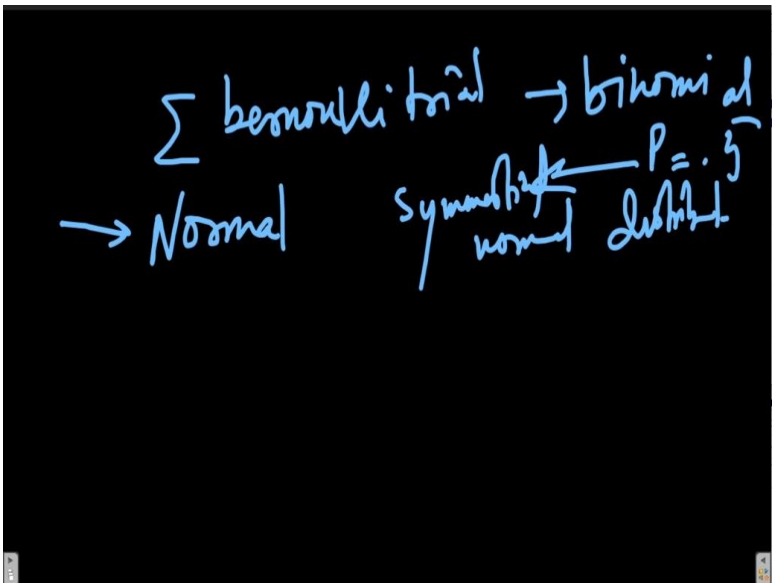
Applied Econometrics
Prof. Tutan Ahmed
Vinod Gupta School of Management
Indian Institute of Technology - Kharagpur

Module - 3
Lecture - 31
Normal Distribution (Contd.)

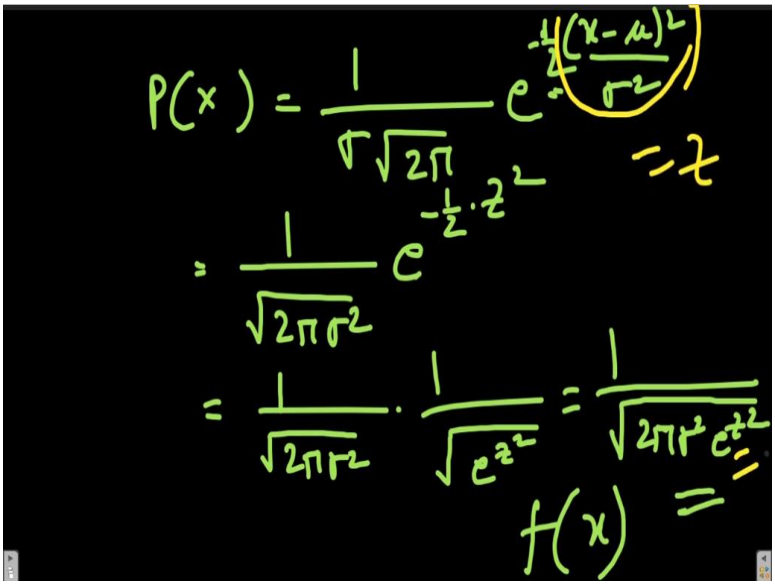
Hello and welcome back to the lecture on Applied Econometrics. We are into the third week of our lecture.

(Refer Slide Time: 00:32)



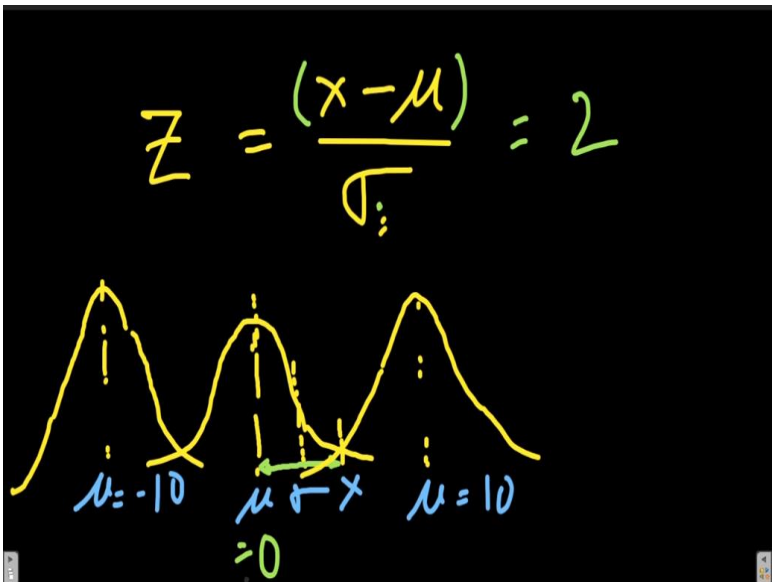
So, we have been talking about normal distribution. And normal distribution is one of the most important distribution that we see in nature, and that is basically widely used in statistics and econometrics. Now, in this lecture, we are going to see the empirical side of it. We are going to use numbers; we are trying to make sense of the normal distribution with numbers. Now, to do that, we remember that we actually used something called Z, we used something called, this part, we call as Z.

(Refer Slide Time: 00:59)



And we substituted that Z and we got this formula. Now, the point here is, what is this Z and why it is important?

(Refer Slide Time: 01:14)



So, we define Z as, Z is equal to X minus mu by sigma. Now, if I try to understand the Z using the normal distribution curve, so, it would mean; let me actually have; let us say, this is the mean, the mu; a different colour; this is mu. And let us say I have my X which is here; X is, let us say, here. And I have a sigma S, let us say S is here; sigma is here. Now, when I subtract mu from X, I get this difference.

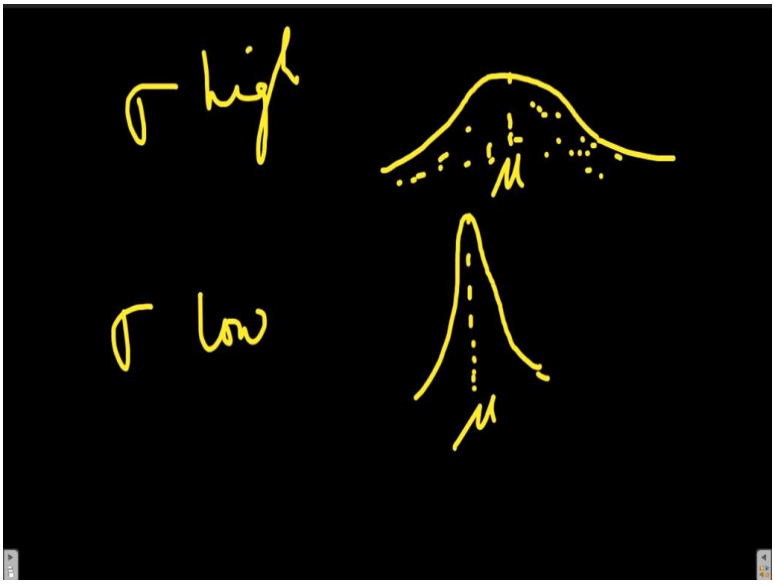
Now, when I divide this difference with sigma, so, basically, I see how many standard deviation away, X is from mu. That is basically the definition. When I do X minus mu by sigma, it shows how many standard deviation away, X is from mu. So, if, let us say, this

distance is 2 sigma; so, then, Z is basically 2, which means that Z is 2 standard deviation away from the mean, mu. So, that is basically the idea.

Now, if I have my mu shifted; let us say I have, mu is 0 here, and I want to substitute mu is equal to, let us say, 10. And then, if I substitute mu as 10, so, in the number line, the normal distribution will shift, let us say here; of course, they are not; I could not draw them equally, but let us say they are equal. And this is my new mu here, and mu is equal to 10 here. And I can also shift it towards the left, and I can, perhaps, let us say mu is equal to -10.

And it will shift to the left; let us say mu is equal to -10 here; and the normal distribution would be symmetrical around mu. So, this is how we actually can construe the shift of normal distribution if we change the value of mu. Let us try to understand what happens if we change the value of sigma. So, if sigma is high, so, which means the dispersion of the standard deviation is high; what happens is that, the normal distribution gets wider. So, that is not really a very highly expected scenario where the sigma is high.

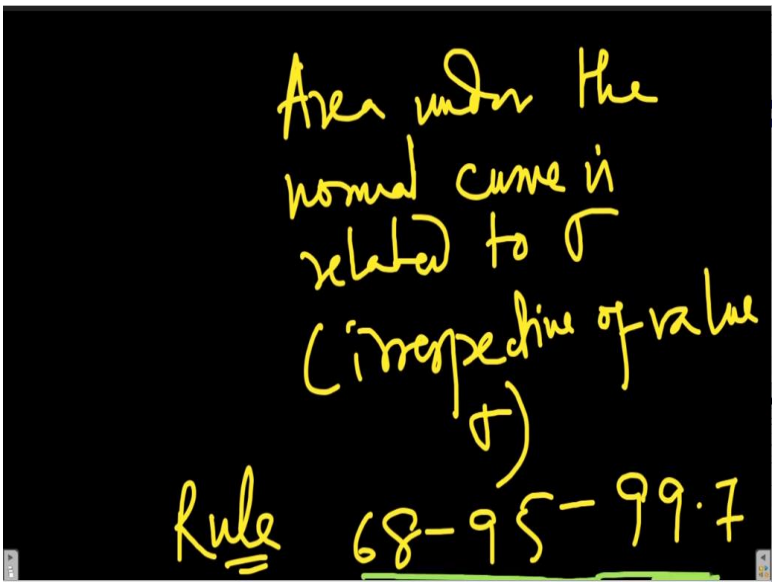
(Refer Slide Time: 04:15)



For a sigma high, we will have the normal distribution something like very flat, flat sort of normal distribution. We do not want that kind of distribution. If sigma is low, we have a narrow sort of normal distribution. So, all the observations are close to mu. And this is kind of a desired distribution, because we want our X's to be close to mu; not to be very far away from the mu.

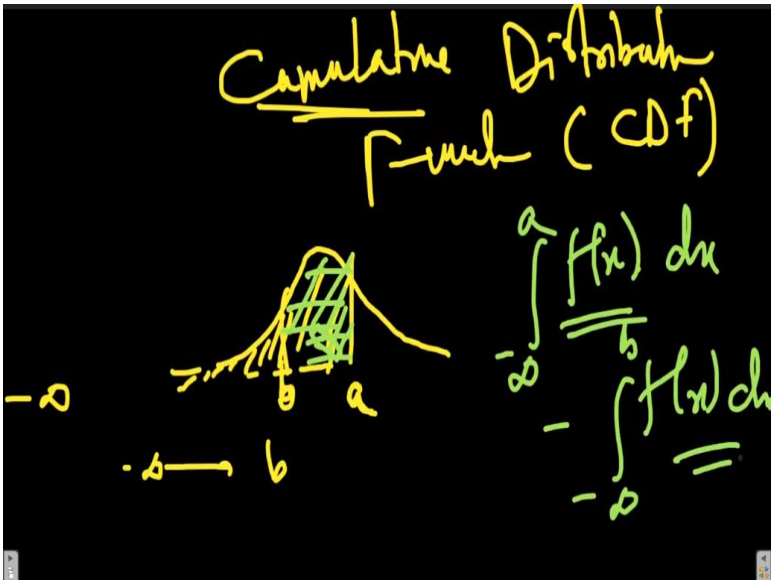
In this case, all the X's are quite far away from the mu, and possibility of; you will actually end up on; it is difficult to actually detect the mu from all these different dispersed observations; but whereas, it is easier here. So, what happens is that, irrespective of whatever the value of sigma is, it is big or small, what happens is that, we will see that the normal distribution is actually the area under the normal curve has some sort of relationship.

(Refer Slide Time: 05:24)



Area under the normal curve is related to sigma; so, irrespective of the value of sigma. Whatever be the value of sigma, it really does not matter; the area of the normal curve will have some relationship with sigma. And we will come to that, and we call that as 68, 95 and 99.7 rule. So, this is the rule. We will just talk about this rule, but before that, let us talk about something a little more fundamental, and that is basically how we understand the area under the normal curve, how we actually obtain the area under the normal curve.

(Refer Slide Time: 06:23)

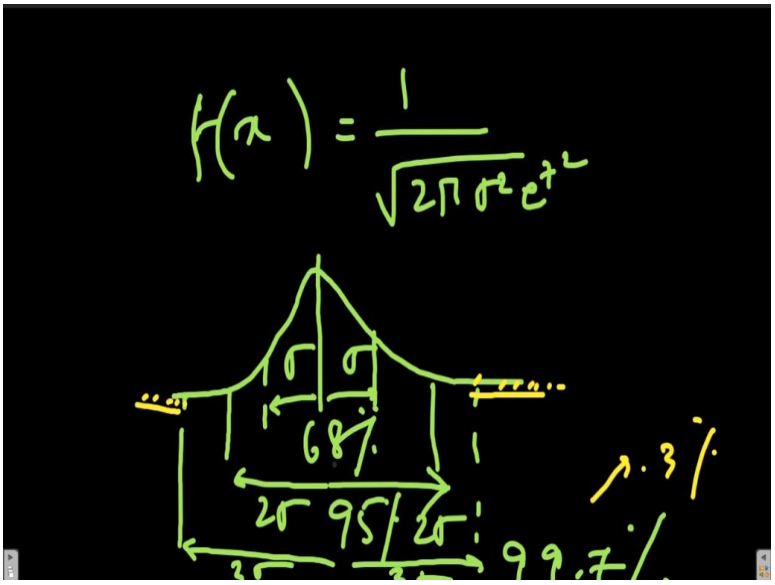


And to do that, we have to again go back to our previous lectures, where we actually talked about cumulative distribution function, CDF. So, when we talked about CDF, it is basically, when we talked about cumulative, so, we basically add up all the probability densities that we have, and you get the cumulative distribution function. And let me actually explain how we do that. So, let us say this is a normal distribution curve.

Let us say, this is my a, and this is my b. And if I want to get the CDF up to the point a; so, essentially, from the left side, I basically add up this area under this curve from minus infinity up to a. This is how I get the area under the curve. If I say the area under the curve for X is equal to a, so, anything less than a would be considered under, when I consider the cumulative distribution function.

But whereas, if I say cumulative distribution function, you want to estimate the area under the curve up to the point b, so, you take minus infinity to b. Now, what do you do when you have to get the area between a and b? So, it is simple. Basically, what you do is, you do from minus infinity to a f x dx, and you basically subtract minus infinity to b, b f x dx. So, basically, you take 2 cumulative distribution functions and you subtract these two; basically, the same cumulative distribution function across two different ranges, and you subtract these two to get the area under this curve. So, depending on what range you want, so, you basically define that way.

(Refer Slide Time: 08:39)



Now, actually, if you remember the formula we derived, so, f x in our case is basically 1 by root over 2 pi sigma square e Z square. Now, if I put this, substitute this f x here, and if I try to integrate this f x over this range; so, it is a little difficult, because the function is not really very easy form of function. So, that is something like difficult of course, but there is some easy way out.

And the easy way out is; one is that, basically that the Z value that we have seen here. And the Z value, where we use this Z; there are certain standard formula or standard rules that we already have. And that is, the area under the curve will actually depend on the standard deviation, irrespective of their value. So, let us say, if I take a standard deviation is equal to 1 standard deviation away; so, if my normal distribution is like this, and if I take 1 standard deviation away in both left and right, so, this basically covers 68% of the area.

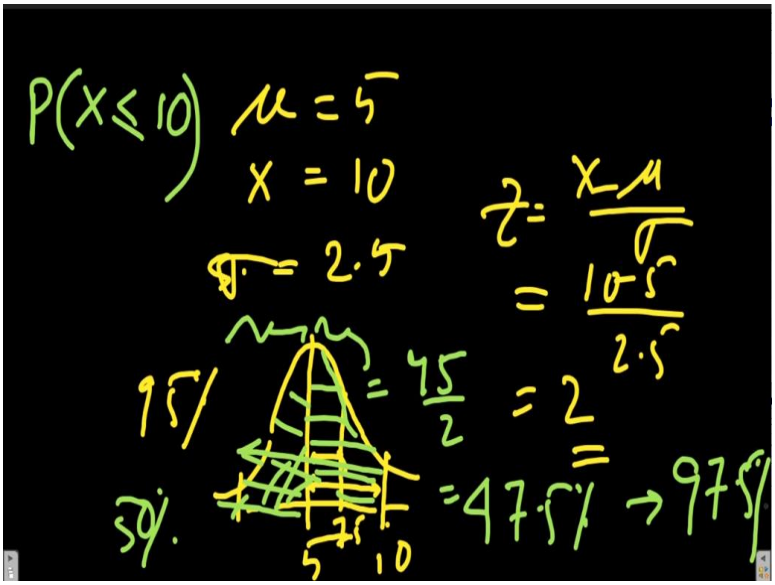
So, sigma, sigma. And if I take 2 standard deviation, so, it is going to be 95%. So, that is basically 2 standard deviations on the left, 2 standard deviation on the right. And if I take 3 standard deviation; so, 3 standard deviation left, 3 standard deviation right; it is going to be 3 sigma, 3 sigma; it is going to be 99.7%. So, basically, if you take 3 standard deviation on both the sides of mu, then the area under the curve is actually 99.7% basically of the entire area.

So, which would mean that; use a different colour; only this little part we are leaving out here; only this little part goes up to infinity; and here, only up to this little part goes up to infinity. So, which would mean, that consists like the rest of the area, which is 0.3%. So,

now, using this, you can actually solve many problems; of course, you can also have; not necessarily it has to be a 1 sigma, 2 sigma, 3 sigma; and in those cases, we can use this standard normal table; this is the z-table.

But let us see how things are easier if I have this formula 1 sigma, 2 sigma, 3 sigma, and if we have this standard area known to us. So, we will try to solve some problems. So, let us do one problem. So, let us say you have your mu is equal to 5.

(Refer Slide Time: 11:41)



And you have your X, for a given value of X, let us say 10. And you have your SD of sigma is equal to 2.5. So, I have chosen this number conveniently, so that I do not have to use a calculator or any z-table. So, now, if I get a Z, so, X minus mu by sigma would give me; so, 10 - 5 by 2.5, which is equal to 2; my Z is equal to 2. So, I get the sigma. And then, of course, I know; if my Sigma is 2; so, if my Z is 2, so, how much area this X would be covering?

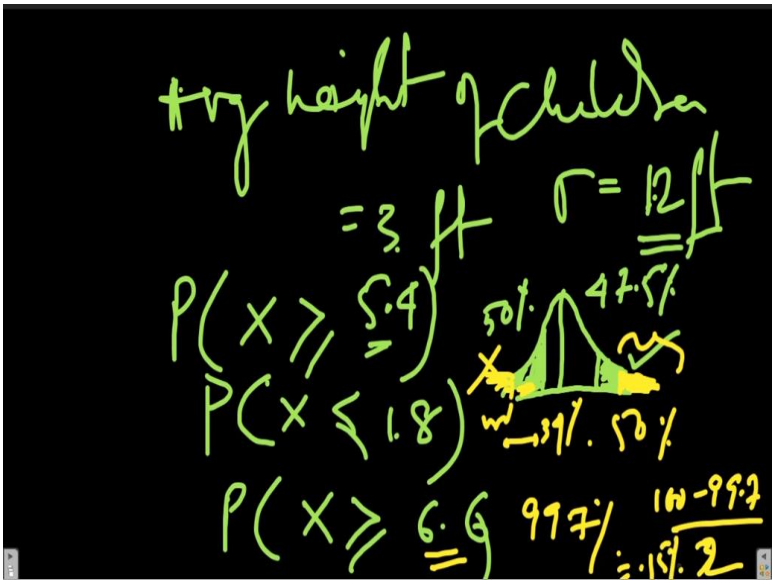
So, I know that my sigma is; so, let us say this is 5 mu and this is 10 and sigma is 2.5. So, basically, if I take 1 sigma away, so, this is going to be 7.5, and this is 10. So, essentially, Z 2 means, it is 2 standard deviation I am talking about. So, this area is actually 2 standard deviation area. So, if this is 2 standard deviation area, and I know from my rule that this whole, if I have 2 standard deviation on both the sides, so, that would have been 95%.

Now, here what I have? I have 2 standard deviation on one side, and then on the entire this side; 50% of the curve is on the other side. So, if I divide this; so, for this part is going to be 95 by 2, which is 47.5%. And this part is going to be 50%, other half of the normal

distribution. So, if I say any number; let us say, if I have to get a probability of X less than equal to 10, for any value of X which is less than 10; so, that means, any value of X which is this side; so, that would mean the entire area, entire 50% and this part; so, which is 50 + 47.5, which is, total is going to be 97.5%.

So, this is how we basically, our life becomes easy when we know the area under the curve. So, let us do another problem. And the problem is, let us say, we are getting, we are actually measuring the height of children.

(Refer Slide Time: 14:36)



And let us say the average height of children is 3 feet. And we have a SD of sigma is equal to, let us say, 1.2 feet. Now, if I want to get; let us say I want to know probability of X greater than is equal to 5.4, probability of X less than 1.8 and probability X greater than 6.6. So, I will give this task for you to solve it. So, I will just give you the hint. So, here, what is happening is that, you have; if I am asking about X greater than 5.4, so, that means, if I add 3, so, this is a mean value; and if I add 2 standard deviation, 2 sigma, so, that will be 5.4.

So, this is my mu, this is my 2 sigma; and this is basically going to be the area here. I need to know the area here. So, here it is 50%. And here we have already seen in the previous example, this is 47.5%. So, what is the rest of the area? So, that is basically the answer. So, that is quite obvious. Here, if X is less than 1.8, again, like 3; so, you have the 3 is mean. And then, if you subtract 1 sigma, so, this area is basically your answer.

Now, we know that 2; so, basically, if we have 2 standard deviation, 1 standard deviation on both sides, so, that is 68%. So, this part is going to be for the second problem, this part is going to be 34%; and here, this one is going to be 50%. So, if I add these two and subtract, so, whatever is left is the answer. So, that is basically 16%. What about the third one? So, third one is, I actually have 3 sigma, so, X has to be more than 3 sigma; so, it has to be here.

So, I know, up to 3 sigma, this area is 99.7%. And I asked for only this area. So, rest 0.3% is basically covering both these areas. So, I do not need this; so, I only need this, because I am only concerned about the heights which are above 6 and 6.6 feet. And that is going to be, so, basically 100% - 99.7 by 2, which is basically 0.15%. So, we are, I think, pretty much now clear about this. I am going to show you some Excel sheet here.

(Refer Slide Time: 17:53)

		PDF	CDF
	-10	0.0000881489205918614	0.0000884172852008147
5	-9	0.0002181706737614 4	0.0002326290790355 4
4	-8	0.0005072620432494 2	0.0005770250423907 66
	-7	0.0011079621029845	0.0013498980316301
	-6	0.0022733906253977 6	0.0029797632350545 6
	-5	0.0043820751233921 4	0.0062096653257761 6
	-4	0.0079349129589168 5	0.0122244726550447
	-3	0.013497741628297	0.0227501319481792
	-2	0.0215693297066279	0.0400591568638171
	-1	0.0323793989164729	0.0668072012688581
	0	0.0456622713472555	0.105649773666855
	1	0.0604926811297858	0.158655253931457
	2	0.0752843580387011	0.226627352376868
	3	0.0880163316910749	0.308537538725987
	4	0.0966670292007123	0.401293674317076
	5	0.0997355701003582	0.5
	6	0.0966670292007123	0.598706325682924
	7	0.0880163316910749	0.691462461274013
	8	0.0752843580387011	0.773372647623132

And this is basically, I will just show what is the probability distribution function, probability density function, and then, cumulative distribution function, and how we can derive the normal distribution curve from here. So, let me actually just write down some numbers. So, let us say I start from -10, -9, -8 and so forth. And let me actually get all the numbers in this way. So, let me get at least 20 numbers.

So, if I just continue with this, so, let us say I get up to 10. So, I get from -10 to +10. And let us say I have a mean is equal to 5; we can keep mean as 5; and sigma is 4, let us say. So, now, when I have to get a PDF; so, basically, we use a function called NORMDIST. So, for NORMDIST, if for a probability density function, I have to give the value of this. The

average is 5; so, since I want it to be constant across all the values, so, it is going to be dollar signs.

And then here, the NORMDIST, the number is going to be this; the average is going to be this; and we put 2 dollar signs. And then we have the standard deviation which is this; again put 2 dollar signs, because we want it to be constant. And the form type is important because form type is talking about whether I want a probability density function or cumulative form. So, let us say I want probability density function, so, which basically means I will get the height; close it I guess; what happened?

So, here is a value. And then, I use the formula for all the cells and get the values. So, I got the values here. So, these are basically the PDF. This NORMDIST is actually giving me a PDF. So, this particular form NORMDIST is giving me a PDF. Now, I get a CDF, cumulative density function. And the same thing, let us say, this is equal to this. So, I will just use the same NORMDIST and then I use the number; then I have the mean.

NORMDIST; so, now I use a number, number is this; and the average, I will use this; the mean; now, I have to put a dollar sign; sticks in the cell; the standard deviation is going to be this; and I have to put a dollar sign; 4 and 8. And to the form type, I need to give the cumulative form, and then I press. So, basically, I am getting a cumulative; all the values from minus infinity to -10 have been added up, when I am using a cumulative form.

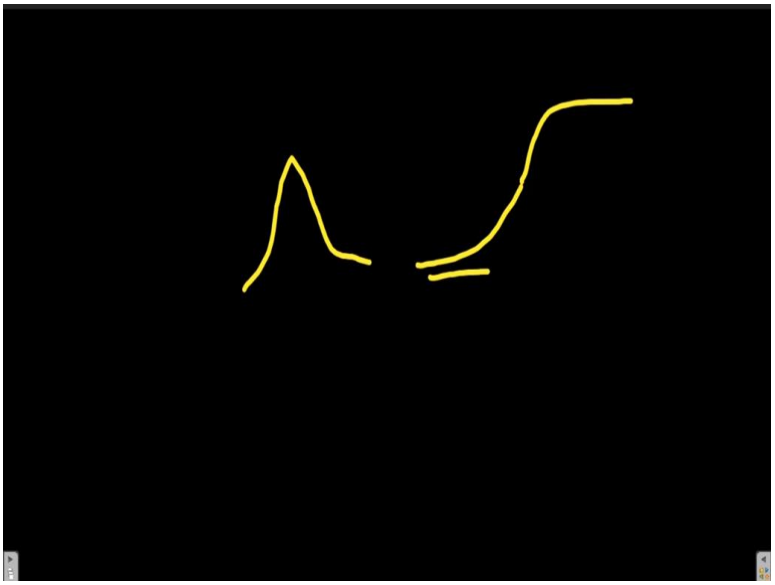
So, let us see. So, it should ideally give me a 0.5 at some point; at 5, I guess, I will get 0.5, because it has to be symmetrical around 5. So, here, I get 5. And then, if I keep, I get other values. So, now, let us say, if I actually plot this, we can have an interesting thing.

(Refer Slide Time: 22:41)

		PDF	CDF	
	-10	0.0043820751233921 4	0.0062096653257761 6	
0	-9	0.0079349129589168 5	0.0122244726550447	
4	-8	0.013497741628297	0.0227501319481792	
	-7	0.0215693297066279	0.0400591568638171	
	-6	0.0323793989164729	0.0668072012688581	
	-5	0.0456622713472555	0.105649773666855	
	-4	0.0604926811297858	0.158655253931457	
	-3	0.0752843580387011	0.226627352376868	
	-2	0.0880163316910749	0.308537538725987	
	-1	0.0966670292007123	0.401293674317076	
	0	0.0997355701003582	0.5	
	1	0.0966670292007123	0.598706325682924	
	2	0.0880163316910749	0.691462461274013	
	3	0.0752843580387011	0.773372647623132	
	4	0.0604926811297858	0.841344746068543	
	5	0.0456622713472555	0.894350226333145	
	6	0.0323793989164729	0.933192798731142	
	7	0.0215693297066279	0.959940843136183	
	8	0.013497741628297	0.977249868051821	

If I make it 0, so, you see, the mean has come to 0; the mean value is, the CDF is 0.5 at 0. Basically, that makes sense, because the normal distribution is symmetric around 0. Now, if I plot a normal distribution curve, I can actually take the different values, the PDF values, and at the same time, we can take different values of the CDF. So, for PDF, I will see the normal distribution curve; whereas for CDF, I will see something like; let me actually draw it.

(Refer Slide Time: 23:31)



For PDF, if I plot the PDF, I will get something like the normal distribution; but if I plot the CDF, I am going to see something like this, the sum total of all the probabilities will be represented by the cumulative distribution function. So, I will not do it here; and I would ask you to do it on your own; get some imaginary numbers, plot some numbers and play with the value of SD, the value of mean and of course use the NORMDIST function to see how the normal distribution is changing its shape and basically with the value of mean and standard

deviation. So, with this, we end this lecture. And a couple of other distributions, we will cover in the next couple of lectures. With this, we end the lecture here. Thank you.