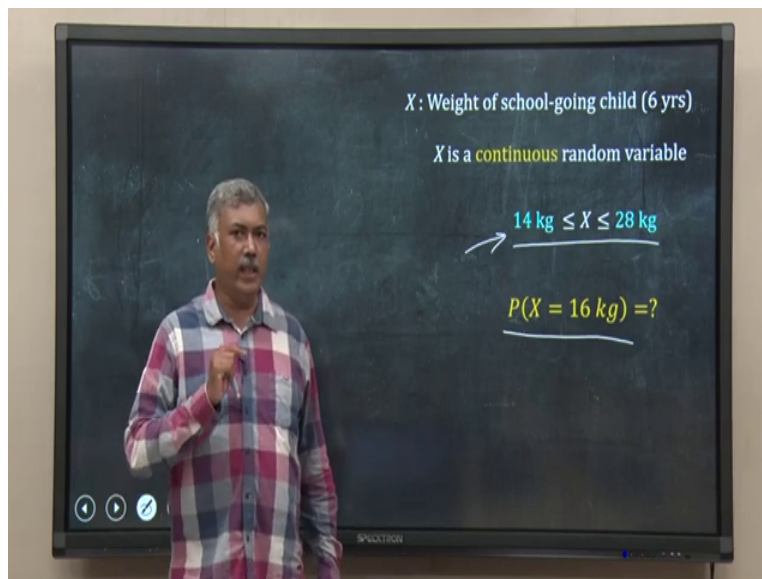


**Data Analysis for Biologists**  
**Professor Biplab Bose**  
**Department of Biosciences & Bioengineering**  
**Mehta Family School of Data Science & Artificial Intelligence**  
**Indian Institute of Technology, Guwahati**  
**Lecture: 3**  
**Continuous Probability Distribution**

Hello, welcome back. In the last lecture, we discussed about Discrete Probability Distribution, involving discrete random variables. In this lecture, we will discuss about continuous probability distribution. Nowadays it is commonly said that childhood obesity is increasing. So, suppose we are studying the weight of different age group of kids. So, we have some information on that.

(Refer Slide Time: 00:59)

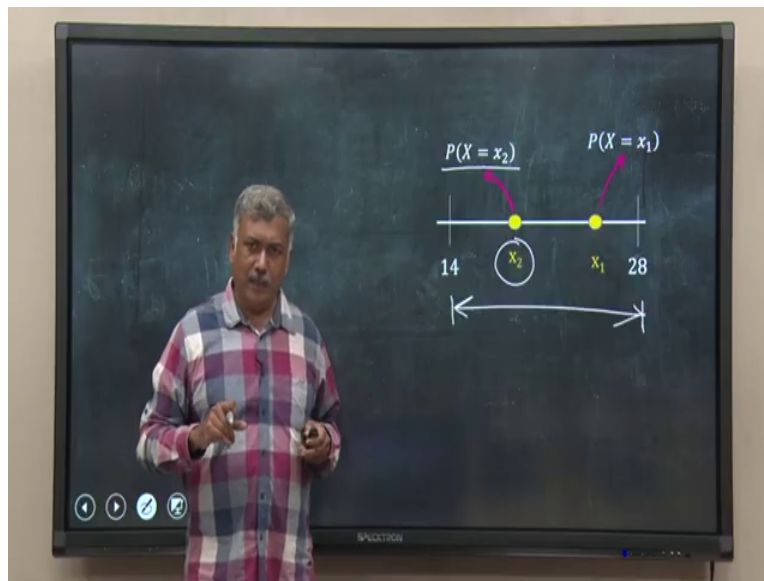


So, suppose  $X$  represent the weight of school going child children of age 6 years in the bracket of 6 to 7 something like that. So,  $X$  is a continuous random variable, because weight is not discrete, the weight can be 16 it can be 16.2 it can be 17, 17.53 something like that, it need not to be 16, 17, 18. So, weight of a child is a continuous random variable.

And suppose, we know that from our data that it varies from, 14 kg to 28 kg for this age group. Now, if I ask you a particular question that what would be the probability that  $X$  is equal to 16 kg? That means, you have to calculate  $P(X) = 16\text{kg}$ . Now, I have not given you any more data.

I have not told you the probability distribution or the raw data behind this question anything, only thing I have say told you is that what I have said you is that the weight varies from 14 kgs to 28 kgs and you have to calculate the probability that weight  $X$  is equal to 16 kg. Let me explain how we will answer this question and where it will lead us.

(Refer Slide Time: 02:13)

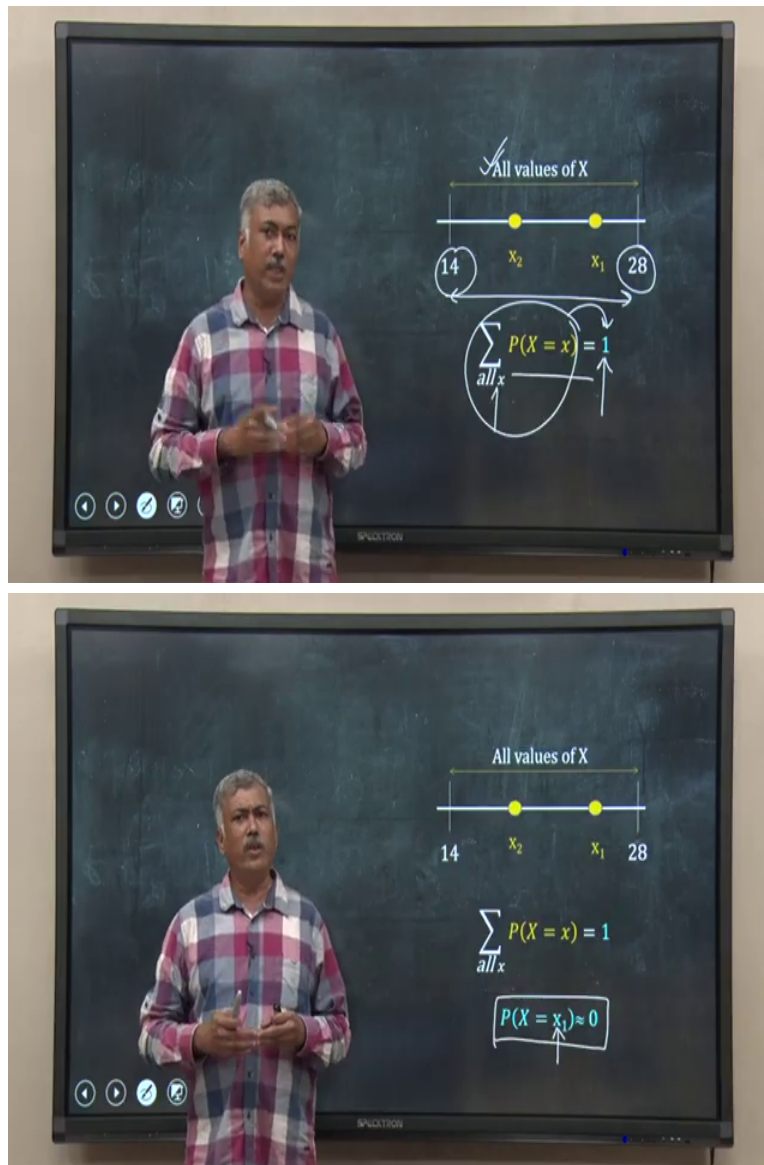


Take a number line. As the weight varies from 14 to 28, so, the number line starts here from 14 to 28, take a number between these two values 14 and 28. Suppose, that is  $X_1$  and suppose the probability of that  $X_1$  is  $P(X = X_1)$ . I do not know the numerical value, just assume that we have a probability value given to us.

Let us take another value between these 14 and 28. So, suppose that it is  $X_2$ . Now  $X_2$  also has associated probability, suppose that probability is  $P(X = X_2)$ . Again I do not know the numerical value and I am not too bother about that right now, just imagine that we know the probabilities. In this way take 100 and 1000s and lakhs of number between 14 and 28. All possible numbers not 1, 2, 1000, all possible number, take all values of  $x$  between 14 and 28.

Remember,  $X$  is a continuous random variable that means it is a continuous number, so, take all possible values of  $x$  between 14 and 28 and get their probabilities. Suppose somehow I know the probabilities.

(Refer Slide Time: 03:32)



Now, if I sum those probabilities, I take the value of X, all values of X from 14 to 28, and I sum the probability of all these values. What will be the result? Using the rule of probability I know that probability will be, the summation will be equal to 1.

$$\sum P(X = x) = 1$$

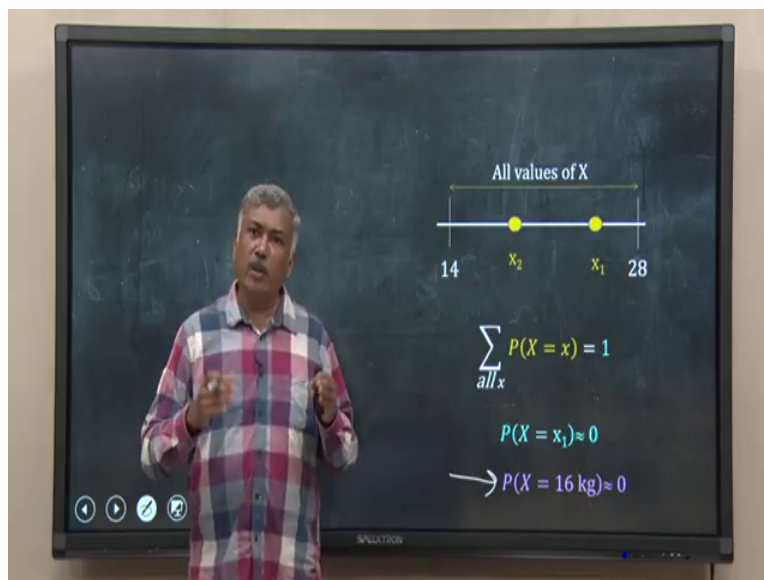
Now, how many numbers are there between 14 and 28? Infinite number, because I am not considering discrete numbers.

So, I can have infinite number of values of X from 14 to 28. And that means I have infinite number of probability values and I am summing all those infinite number here and that gives me equal to 1. So, summation of infinite terms giving me 1. So, what is the value of 1 of those term? For example, what will be the value of when P is taking X is taking a particular value of X, X1. So, if summation of infinite term is equal to 1 then a particular probability, X taking a particular value X1, must be infinitesimally small, very, very small and almost equivalent to 0.

$$P(X = X_i) \approx 0$$

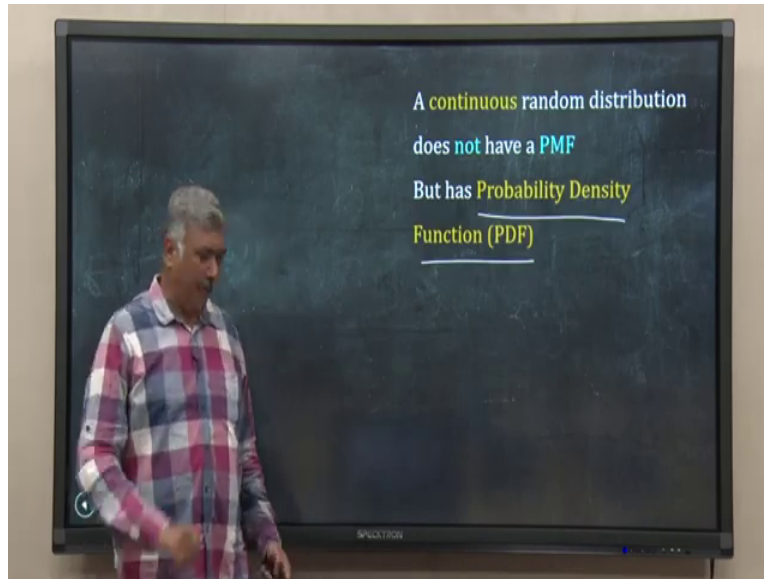
If I go back to my original question.

(Refer Slide Time: 04:51)



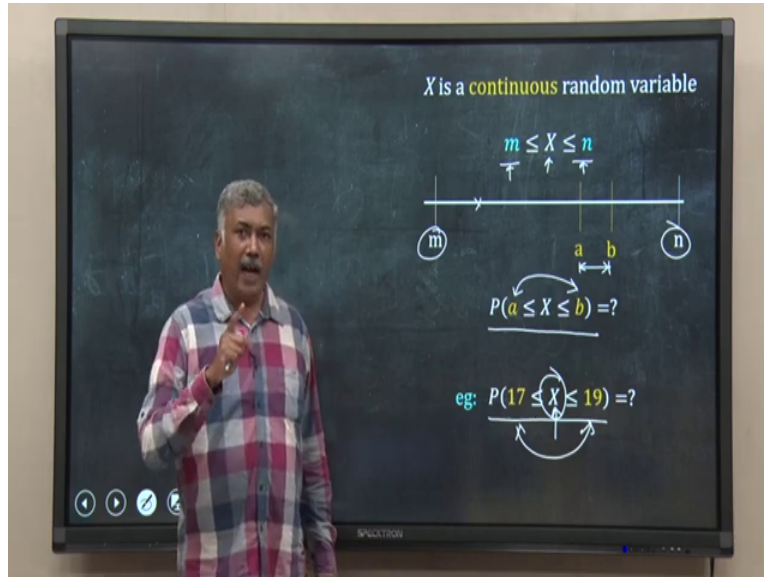
That means, the probability that X weight will be equal to 16 Kg will be infinitesimally small or equivalent to 0, for every practical purpose. So, in a way asking the question that calculate the probability that X will be 16 kg or in general term that probability of X taking a particular value when X is a continuous variable does not make sense, because that value will be 0.

(Refer Slide Time: 05:22)



So, that is why. If I am dealing with a continuous random variable just like weight, height, something like that, we do not consider probability mass function, because that does not make sense, because probability mass function, if you remember from the last lecture gives you the probability that a variable take a particular value, that does not make sense here. So, for a continuous random variable, we do not have probability mass function, we cannot have that. Rather we will have Probability Density Function PDF. And in this lecture, I will explain what is PDF. Let us start that.

(Refer Slide Time: 06:01)

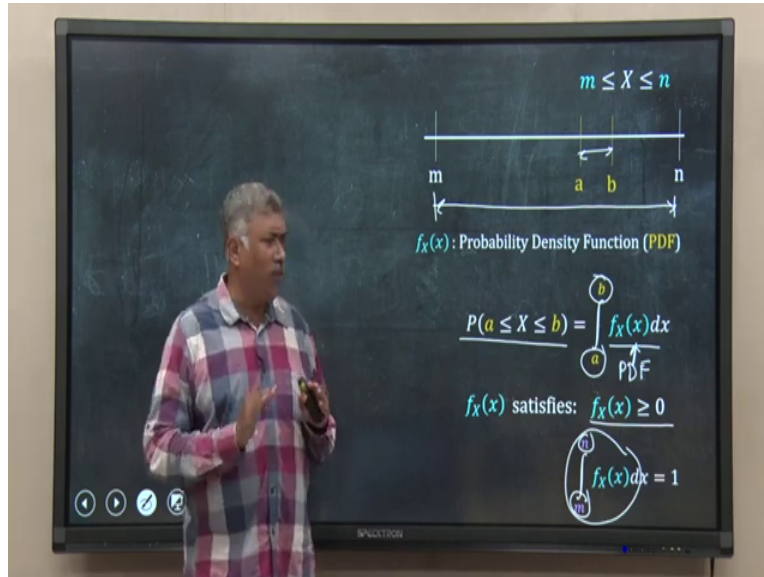


Again, take the number line, so, generalize it  $X$  is a continuous random variable that varies from  $m$  to  $n$ . So, on the number line I have marked  $m$  and also  $n$ . In this case, it does not make a sense to ask what is the probability that  $X$  will take this particular value somewhere there. I marked by the cross, it does not make sense.

But what if I asked you tell me the probability that  $X$  will lie between  $a$  and  $b$ ,  $a$  and  $b$  is a tangible length on this number line. So, that means the probability will be also something tangible bigger than 0, possibly, which can must not be infinitesimally small, it should be something a tangible value.

So, I want to calculate the probability that  $X$  lies between  $a$  and  $b$ . For example, if I go back to that childrens weight problem, I am asking you now to calculate the probability that the weight lies between 17 and 19. In the previous question, I asked you what is the probability that  $X$  equal to 16? That does not make sense. Now, I have changed my question I am asking give me the probability that  $X$  will lie between 17 and 19, in a interval from 17 to 19. That makes sense. And to calculate the answer for this question, I have to use Probability Density Function PDF.

(Refer Slide Time: 07:33)



So, imagine, I know the probability density function, it is a function, so it is  $f_x$  is a function, and this function will give me that probability that I am asking you to calculate. How it will give. If I integrate that function, this is the PDF. That is I am defining  $f_x$  is the PDF.

If I integrate this PDF,

$$\int f_x(x) \cdot dx = P(a \leq X \leq b)$$

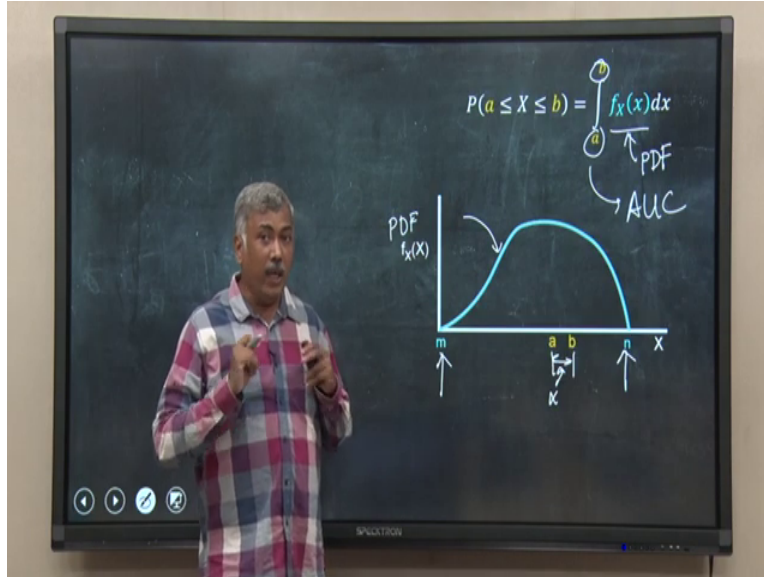
So, integration of this density function, which I have assumed I know somehow will give me the probability that  $x$  lies between this range, in this interval  $a$  to  $b$ .

Apart from this, this PDF,  $f_x$  that I have written should satisfy two other condition. One is, it should be a positive thing, the PDF should be a positive thing,  $f_x \geq 0$ . And if I integrate this function from  $m$  to  $n$ , that means for the whole range for  $x$ , that should give me 1.

That makes sense, because, if random variable varies from  $m$  to  $n$ , then the probability that it will lie between  $m$  to  $n$  must be equal to 1. So, the integration of my PDF from  $m$  to  $n$  must be giving 1, this condition must be satisfied. So, this is how we define a PDF. In this lecture, we will go into some very commonly used PDF. But before that, let me graphically explain what we are doing by this integration.

(Refer Slide Time: 09:23)



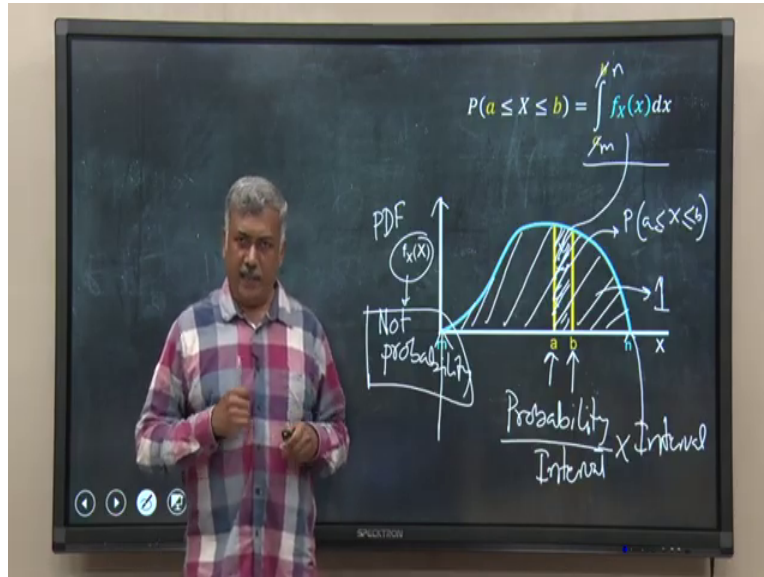


Let us take this plot. Suppose I know this  $f_x$ , this is my PDF, Probability Density Function, where  $x$  is a random variable, which is the continuous one, and it varies from  $m$  to  $n$ . So, I have  $x$  in the horizontal axis and the PDF on the vertical axis. And this blue line is my PDF. I have drawn it arbitrarily just to show you, explain to you, do not go by the shape.

And I have two numbers  $a$  and  $b$  here, and I want to calculate the  $P(a \leq x \leq b)$ . So, when I say I will get that answer by integration, by integrating from  $a$  to  $b$ , what I am doing, integration of a function is nothing but getting the area under the curve. So, integration is nothing but getting the area under the curve. So, I am integrating, in this case, from  $a$  to  $b$ . So, let us draw two line.

(Refer Slide Time: 10:30)





So, I have a line at a, and I have a line at b. So, the area under the curve in this region is the shaded region. So, when I integrate this function,  $f_x(x)$  from a to b, I get this area. So, that means this area is nothing but probability that x lies between b and a here in this interval. So, this area under the curve in that region is the probability.  $f_x$  is not probability, it is not probability.

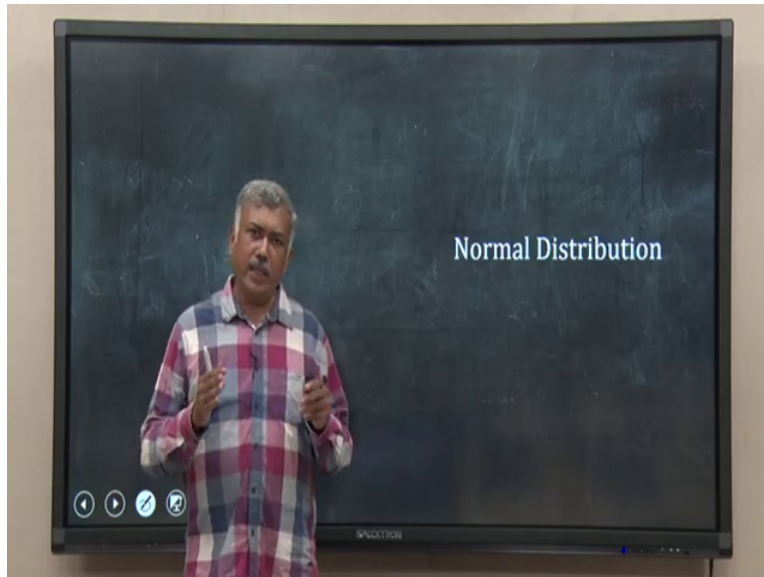
We have to understand the difference. In PMF plots, in the previous lecture in the vertical axis, suppose you take a binomial distribution, or poisson distribution, the vertical axis we have the probability, probability mass function, those are giving you the probability. Here in this plot, I have in the vertical axis, the PDF, and PDF does not give me the probability.

When I integrate and take the area under the curve by that integration, that gives me the probability. Now, you can easily imagine if I integrate from m to n, that means where I am integrating from this point to this point that means now I want the whole area under the curve. And by definition of PDF, this must be equal to 1.

Because the probability that the random variable which varies from m to n, that it will remain in that interval is obviously 1. So, this is what we call probability density function. Remember, it is not mass function, it is density function, that means it is probability divided by interval length. So, it is density. And when I have to calculate the probability, I have to multiply it with the interval. And that is what we are doing by integration to calculate the probability that the random variable stays in a particular interval. Now, I will move into some very commonly used

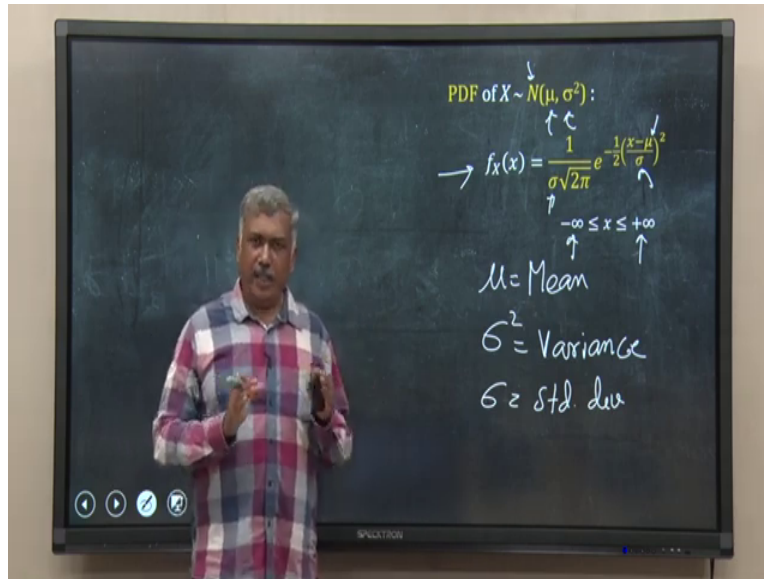
probability density function in data analysis and statistics. Obviously, we will start with normal distribution.

(Refer Slide Time: 13:05)



It is one of the commonest thing that I will consider where it is distribution will consider while discussing statistics and data analysis, so, let us start with that.

(Refer Slide Time: 13:17)



So, the PDF of a random variable that follows a normal distribution is defined in terms of two parameters, one is  $\mu$  and another one is  $\sigma^2$ . So,  $\mu$  is the mean value of that random variable, suppose we are talking about the weight of children. So, the mean weight of the children in that cohort, whereas  $\sigma^2$  is the variance. Obviously

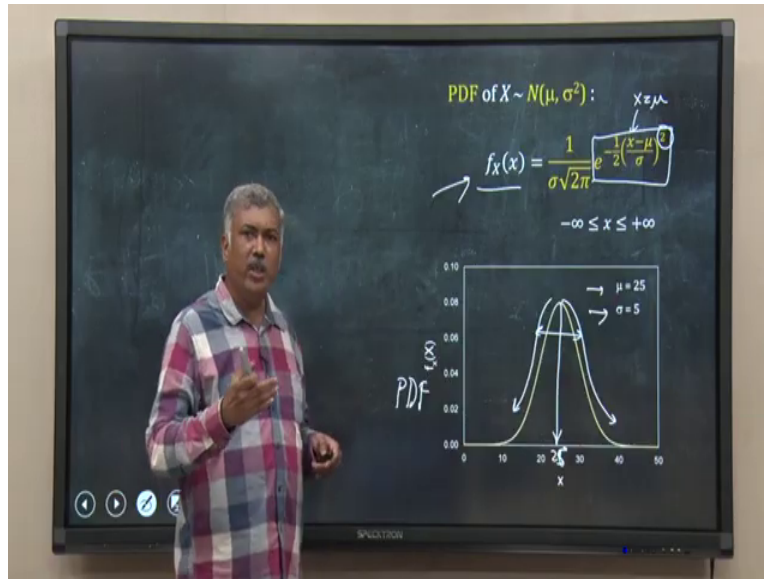
Then  $\sigma$  is my standard deviation. So,  $\mu$  and  $\sigma^2$  or in other term  $\sigma$  defines this PDF and the PDF is given here. So, you have  $\mu$ , you have  $\sigma$ , you have  $\sigma$  and, in this case remember we are talking about real continuous variable, so,  $x$  varies from minus infinity to plus infinity.

Usually, in notation term, if I am talking about a random variable or a measurement or variable in general, which I believe follows normal distribution, we write it like this

$$\text{PDF of } X \sim N(\mu, \sigma^2)$$

So, that means that this random variable or the measurement that I am making, that variable that I am measuring in experiment follow normal distribution. Now, normal distribution is very symmetric. Let us try to understand the behavior of that.

(Refer Slide Time: 14:45)

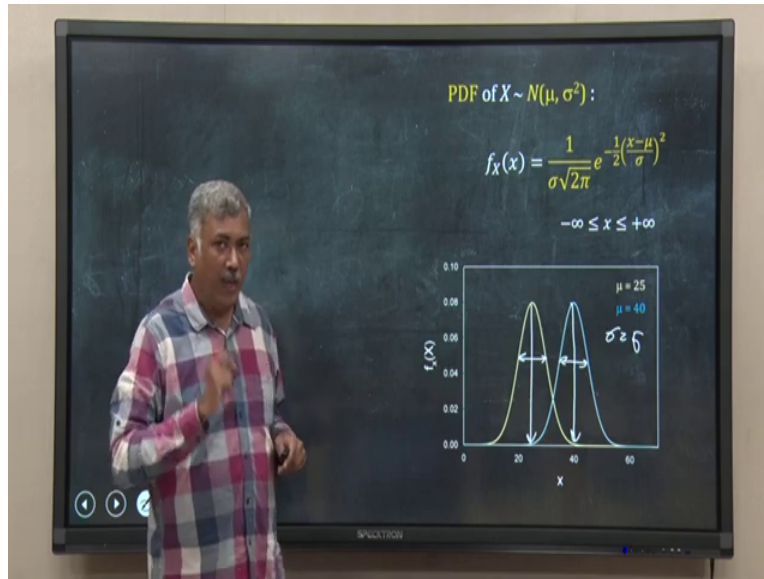


So, what I have done here, I have made a plot where I have decided that  $\mu$  is 25, mean is 25 and  $\sigma$ , that is a standard deviation is 5. So, now, if you look into the PDF so, when  $x$  is equal to  $\mu$ , when  $x$  is equal to the mean value, then it will  $x$  minus  $\mu$  is 0, that means  $e$  to the power something is 0,  $e$  to the power 0,  $e$  to the power zero must be 1.

So, actually, when the value of the random variable is equal to its mean, when  $x$  is equal to  $\mu$ , I should have the highest value. So, this functions will have the highest value when  $x$  is equal to  $\mu$ , and you can easily see it here, the highest value of this function. So, this is the PDF. So, the PDF has the highest value when I have 25,  $x$  equal 25. And now, think about the  $\sigma$ , what is happening besides these two sides, the on beside this mean value, the  $x$  should drop exponentially, because this is a exponential decay function,  $e$  to the power minus something.

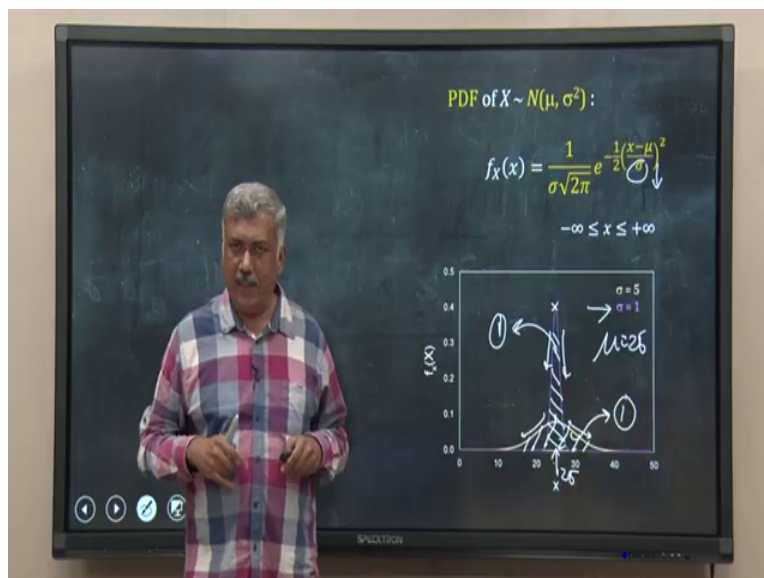
So, that means the value should decay exponentially very fast and not linearly exponentially very fast. And that is why around the mean, the values are falling exponentially. And as we have a squared term here, so the rate of fall on both sides of the mean will be identical. So, that is why my normal distribution is a symmetric bell shaped. Now, let us see what will be the effect of mean. We remember, we have said that the peak should be at the mean. So, now, let me change mean from 25 to something else.

(Refer Slide Time: 16:40)



A higher one,  $\mu$  equal to 40. So, obviously, now, my peak has shifted. So, my whole distribution has also shifted to the right hand side, but their variance is same. I have kept variance equal to 5. So, the rate of fall on both side are actually same, the thickness width of this bell shaped curve are same for both the curve because I have kept the sigma same. Now, if I change the sigma and keep the  $\mu$  same, what will happen, let me check that.

(Refer Slide Time: 17:13)



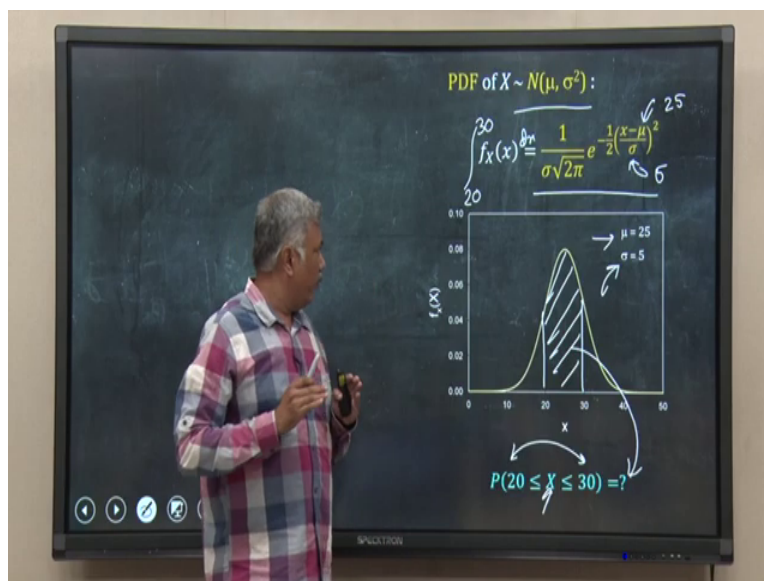
So, here what I have done, I have kept  $\mu$  equal to 25, constant. So, that is why all the peaks this one and this one are at 25,  $x$  equal to 25. But for the pink plot, if you can see clearly,  $\sigma$  is less, 1. So, if some, if I reduce the value of  $\sigma$ , the rate of fall will increase because  $\sigma$  is in the denominator and I have  $e$  to the power minus something.

So, the rate of fall will increase, that is why it is dropping faster. So sharp, whereas this one is dropping lazily, slowly, because  $\sigma$  is 5. You can understand it in another way, when  $\sigma$  is 1, that means most of the variance is very low, variance will be also  $\sigma^2$  will be 1, so, all the values of  $x$  are close to the mean. When the variance is 25.

Because  $\sigma$  is 5, the distribution will be fatter, so, the height of the peak should drop to accommodate this because the integration, that is the area under the curve, the whole thing this is also 1, and integration of this pink curve, below the area under the curve of the pink curve, this is also 1.

So, if one is fatter, its height has to be lower, if one is very thin its height has to be more, as simple as that. Now, we have understood the PDF, and how  $\mu$  and  $\sigma$  is affecting the shape of this symmetric bell-shaped normal distribution. Now, let us try to use this normal distribution to calculate probabilities because remember, we are using probability density function to calculate probabilities by integration or by area under the curve method. So, let us see few example.

(Refer Slide Time: 19:09)

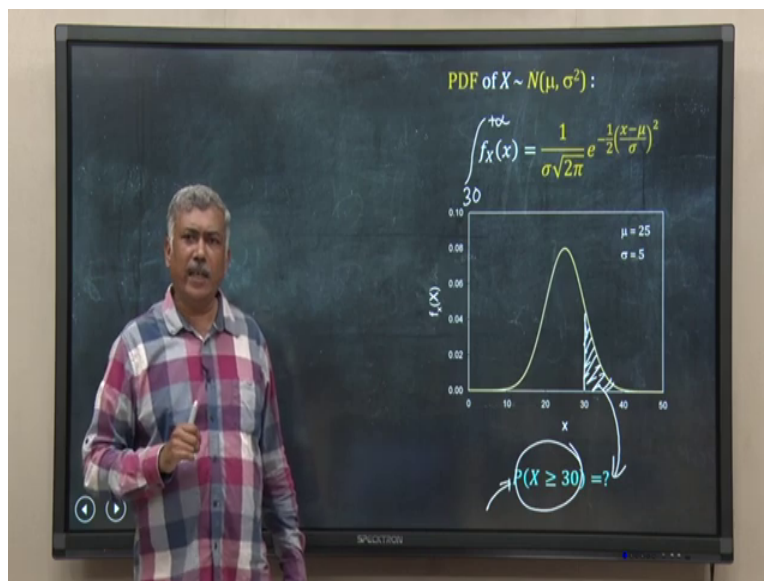


So, suppose this my plot, the plot I have shown here for  $\mu$ , the mean is 25 and  $\sigma$ , the standard deviation is 5. And this is given to you and we have told you that  $X$  is following a normal distribution, that means PDF is normal distribution. Now, I have to calculate the probability that the  $X$  lies between 20 and 30,  $P(20 \leq X \leq 30)$ .

How should I do that? From the concept of area under the curve, what do I want? To calculate is that from 20 draw a line, from 30 again, I draw a line up to this. And this area, this area under the curve, this is the probability that I am asking you to calculate. The probability that  $X$  lies between 20 and 30 is given by the area of the shaded region. And how do we do that? Obviously, we do not calculate this type of question graphically, what we do, we simply integrate this one.

So, I will integrate this one from 20 to 30, obviously, you have to put a  $dx$  there, I am not writing it explicitly. So, essentially, I will integrate the probability density function which has this form where  $\mu$  will be 25 and the  $\sigma$  will be 5. I will integrate that PDF from 20 to 30. If I now change the question and ask you to calculate a different probability.

(Refer Slide Time: 20:50)



Suppose I want you to calculate  $P(X \geq 30)$  means 30 to anything else, anything possible on the higher side. So, then what will be the graphical explanation. So, I have a line here. So, this is at 30. So, this area, this shaded area is this probability.

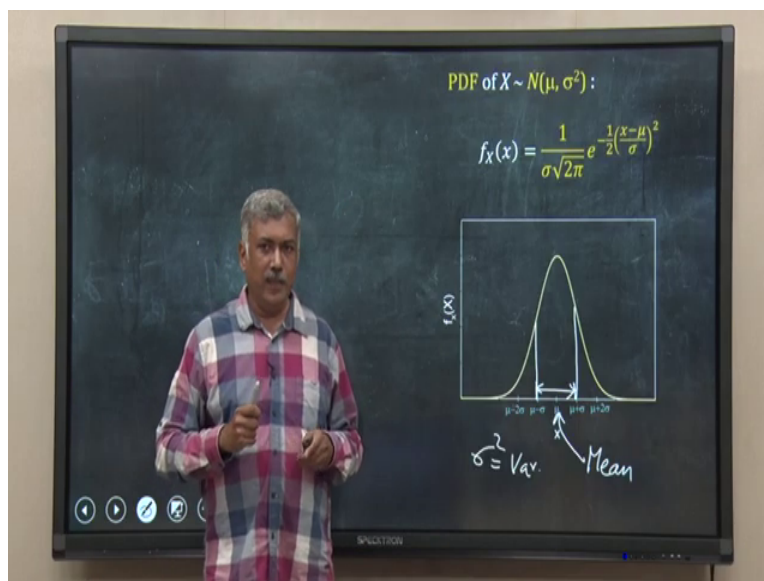


So, that shaded area starting from 30 towards the right hand side, towards positive infinity is the area which should give me the probability that  $X$  is greater equal to 30. So, usually, how you calculate, you calculate it by integrating it from 30 to positive infinity. The good thing is actually you do not need to do this integration yourself.

Even if you know how to do the integration, there are lots of calculators available online or some offline softwares that are available, where you can actually plug the value of  $\mu$  and  $\sigma$  and then you can actually ask this range, and they will integrate it for you and you will get the probability value, and most of the programming languages from Python to R you will have library function to do this one.

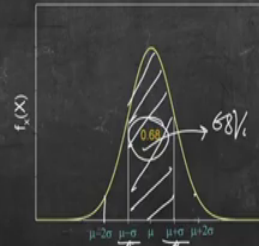
So, usually you have not to go into the integration, but you have to understand the basic concept that how we are actually calculating the probability from probability density function in terms of area under the curve or in terms of integration. Hope the idea is clear to you. Now, remember normal, we discuss the normal distribution is symmetric. And that is why it has some very interesting property, which is widely used in data analysis and statistics, let us discuss that.

(Refer Slide Time: 22:35)



PDF of  $X \sim N(\mu, \sigma^2)$  :

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$



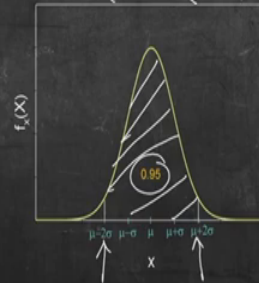
$$\rightarrow P(\mu - \sigma \leq x \leq \mu + \sigma) = 0.68$$



PDF of  $X \sim N(\mu, \sigma^2)$  :

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

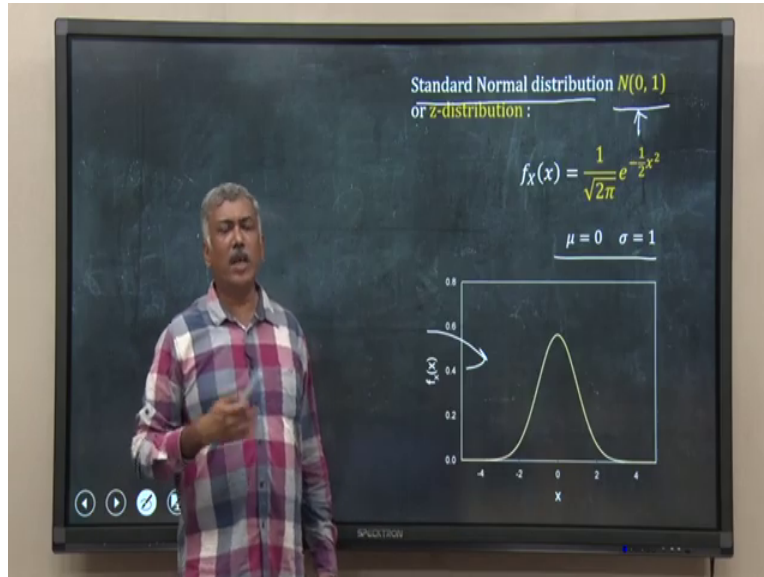
$$P(\mu - 3\sigma < x < \mu + 3\sigma)$$



$$\approx 0.997$$
$$= 0.997$$

$$\rightarrow P(\mu - 2\sigma \leq x \leq \mu + 2\sigma) = 0.95$$





See, for example, what I have done here, I have drawn a generic curve, generic normal distribution having mean at  $\mu$  and  $\sigma^2$  is the variance. Now, if I, this is a generic one, you can assume any value for  $\mu$  and  $\sigma$ . I am talking about a generic behavior of this normal distribution.

Now, if I ask you that what is the probability that your random variable which is following this distribution, a normal distribution, will lie between this region one  $\sigma$  on both side, in  $(\mu - \sigma)$  to  $(\mu + \sigma)$ . We say it one  $\sigma$  away from the mean in other words, either on the left hand side or on the right hand side.

So, I want to know what is the probability that this random variable will lie in the interval of one  $\sigma$  on the left hand side and right hand side from the mean. And you can do the calculation and we know roughly, this is equal to 68 percent or the probability will be 0.68. So, that means in this plot, this area which is lying between  $(\mu - \sigma)$  to  $(\mu + \sigma)$ , this is nothing but 68 percent of the whole area under the curve. So, in other words, the probability is 0.68.

So, the probability that the random variable  $X$  will lie between  $(\mu - \sigma)$  to  $(\mu + \sigma)$  that means, one  $\sigma$  interval, is 0.68, it is not exactly 0.68 slightly bigger than that, but for our all purposes thumb rule we can consider it 0.68 or 68 percent. And remember, this is true for whatever value of  $\mu$  and  $\sigma$  you assume, it does not depend upon the value of  $\mu$  and  $\sigma$ .

So, that is why it is so powerful. Now, if I ask something else, what will be the probability if I am  $2\sigma$  away from mean, that is also known and you can also calculate that yourself, and that will

be 95 percent, that is 0.95. So, this whole area where the number lies between  $(\mu - 2\sigma)$  to  $(\mu + 2\sigma)$ , that is  $2\sigma$  away from mean either on the left side on the right side.

So, this whole interval, the probability that a number will lie is 95 percent or 0.95. So, the probability that  $X$  lies between  $(\mu - 2\sigma)$  to  $(\mu + 2\sigma)$  is 0.95, again it is not exactly 0.95, you have few terms afterward, but for simplification, for ease of remembering we say it is 0.95. And in fact I have not shown here if I go further for example.

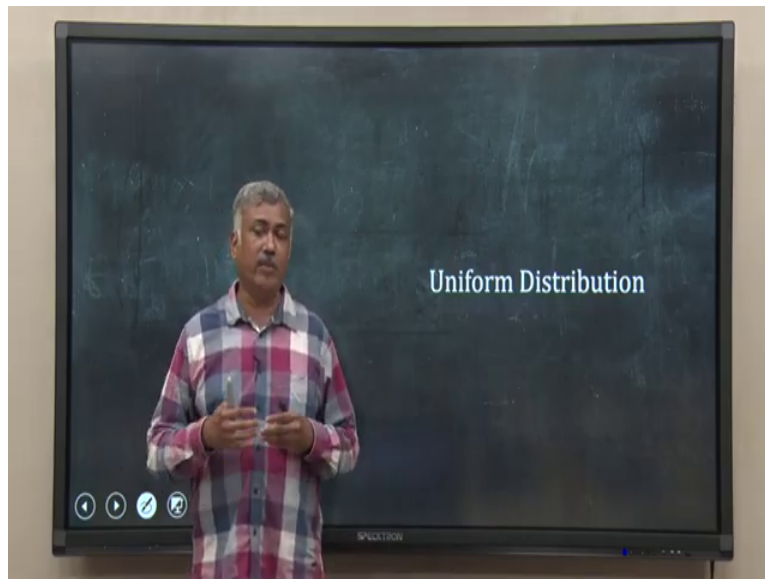
If I say probability that  $(\mu - 3\sigma)$  to  $(\mu + 3\sigma)$  that means, the probability that the number will, the value, variable will lie in within the interval of  $3\sigma$ , three standard deviation. This is roughly 99.7 percent so, this is 0.997. And again, I like to remind you, irrespective of the value of  $\mu$  and  $\sigma$  you have considered.

So, that is why this  $1\sigma$ ,  $2\sigma$ ,  $3\sigma$ , these interval based probability in normal distribution is called the rule of thumb of 68, 95, 99. If you remember this it helps you immensely in case of statistical analysis for lots of different types of data. There is another very useful form of normal distribution which is widely used in a data analysis and statistics, is called standard normal distribution. In a standardized normal distribution, what we are assuming here, we are assuming that the mean of the variable is equal to 0, and the standard deviation is 1.

So, you can actually bring mean to of any variable to 0, by subtracting the real mean. So, I can convert any variable in the form of 0 mean, 1 variance. So, that is what we do in most cases when we do statistical analysis. So, this  $N$  is the variable, the standard normal distribution  $N(0,1)$ . 0 is the mean,  $\sigma$  is equal to 1.

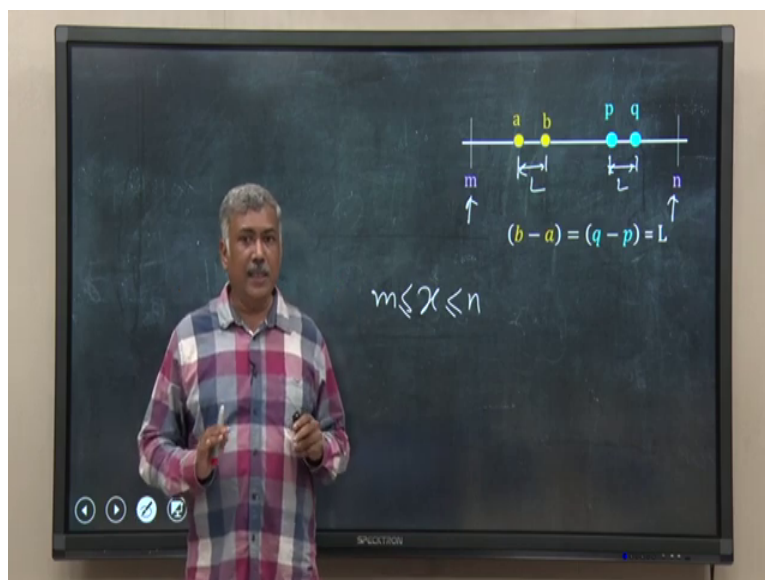
So,  $\sigma^2$  will be also 1. So, the shape of that will look like the one I have plotted in this diagram, this is called the standard normal distribution and some time it is also called  $z$ - distribution. Now, I am going to few other, particularly two other continuous probability distribution.

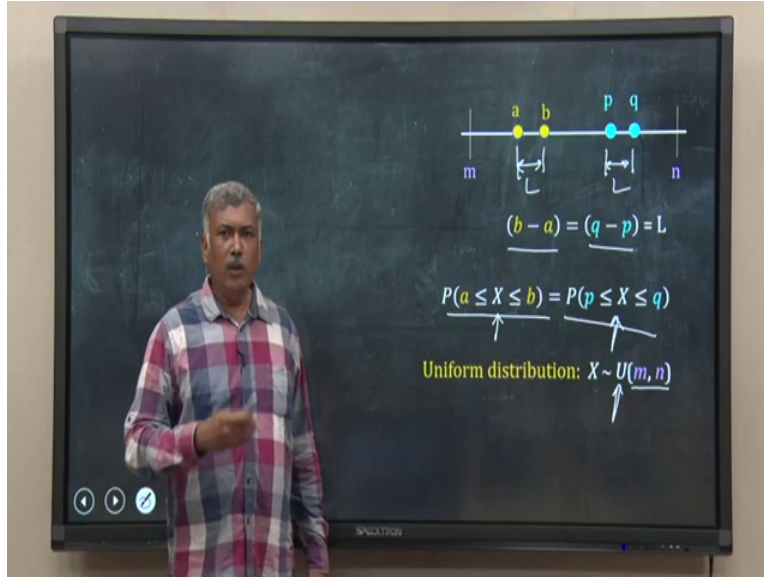
(Refer Slide Time: 27:38)



One is called Uniform Distribution and although it is not widely used in case of statistical analysis, but for many data analysis algorithms, where you have to sample, actually sample data from a large set of data something like that, we usually use uniform distribution, let me explain what is that.

(Refer Slide Time: 27:59)





Suppose, I have a number line and it start at  $m$  and it ends at  $n$ . So, that means  $X$ , a continuous random variable varies from  $m$  to  $n$ , and in this number line, I have two sets of data points, one is  $a$  and  $b$  and their gap between them is equal to  $L$ . And there is another set  $p$  and  $q$  and the gap between the interval between them is also  $L$ .

So, I have two short intervals between  $m$  and  $n$ ,  $a$  to  $b$ ,  $p$  to  $q$ , the length of them are same is equal to  $L$ . Now, if the random variable that we are dealing with,  $X$ , is such that,

$$P(a \leq X \leq b) = P(p \leq X \leq q)$$

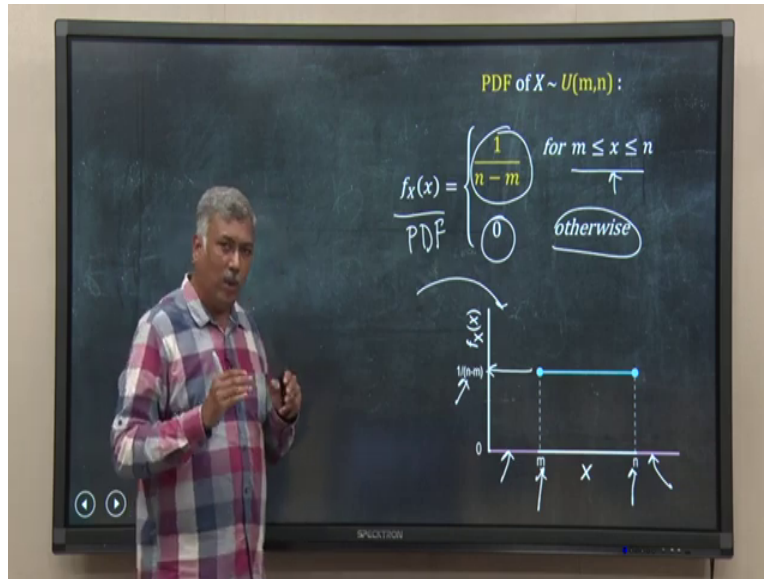
So, that means that the interval, its probability depends upon the interval  $L$ , it does not depend on where this interval is, whether it is  $a$  to  $b$  or  $p$  to  $q$  does not matter, it only matters on the length of the interval and that is  $L$ . So, if the probability that  $X$  lies between  $a$  to  $b$  is equal to the probability that  $X$  lies between  $p$  and  $q$ , where  $b$  minus  $a$  is equal to  $q$  minus  $p$ .

Then this random variable is said to be following uniform distribution and we usually write that in this way.

$$X \sim U(m, n)$$

Now, look at the PDF of this, we have to find out the PDF it satisfies these criteria.

(Refer Slide Time: 29:56)



And people have done that and the PDF for this type of distribution, a random variable which is varying from  $m$  to  $n$  and is uniformly distributed, that PDF, this is the PDF, that PDF will be equal to,

$$\frac{1}{n-m}, \text{ when } X \text{ is between this range.}$$

0, if it is outside this range

And that is what I have shown in this diagram, between  $m$  to  $n$  the  $f_x$  is  $\frac{1}{n-m}$ . Outside this range,

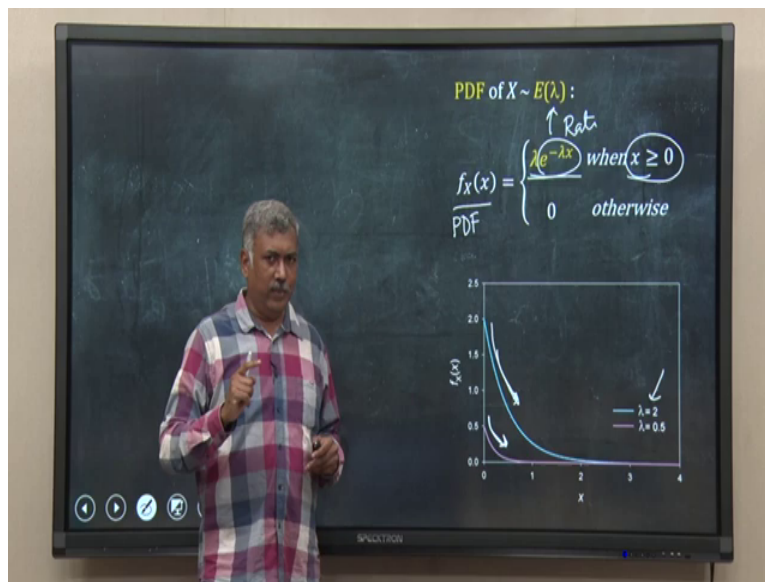
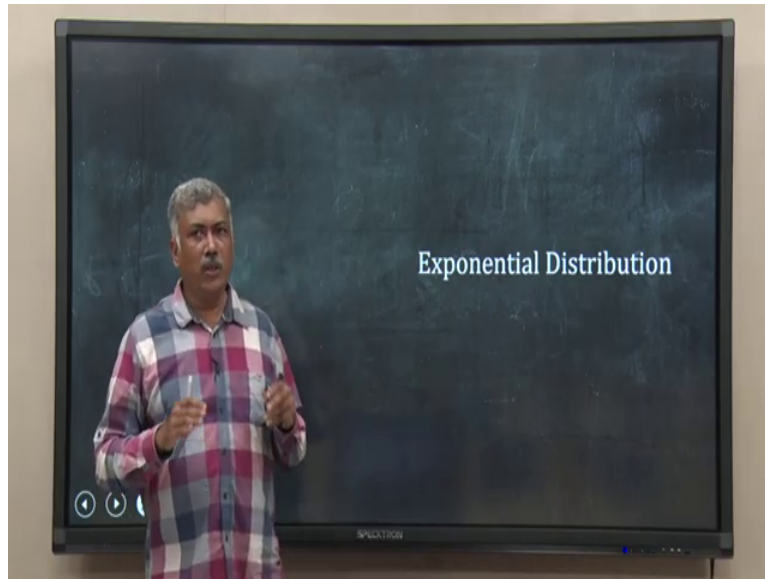
we have 0. Interesting part is that most of the computer programming languages, from the earliest one to the most advanced one, they will have library function to generate uniformly distributed random number, and that is immensely powerful tool, we can do lots of different types of simulation and data analysis considering random sampling, where it is do sampling randomly following uniform distribution.

And the other good thing is that if you can generate random numbers following uniform distribution, you can actually use those random number to generate a new set of random number following a particular distribution, which is non-uniform. For example, you want to generate exponentially distributed random number, normally distributed random number, you can generate from the data generated from a uniform distributed random number generator. So, that



is why understanding and learning uniform distributed random number is so important in data analysis and simulations.

(Refer Slide Time: 31:38)



Now, let me go to the last one for today's discussion, Exponential Distribution, this is also a continuous distribution and its PDF has the following form. It is a single parameter thing, we write it as  $\lambda$ ,  $\lambda$  is the call the rate parameter. So, the PDF, Probability Density Function of a exponential distribution is given like this,

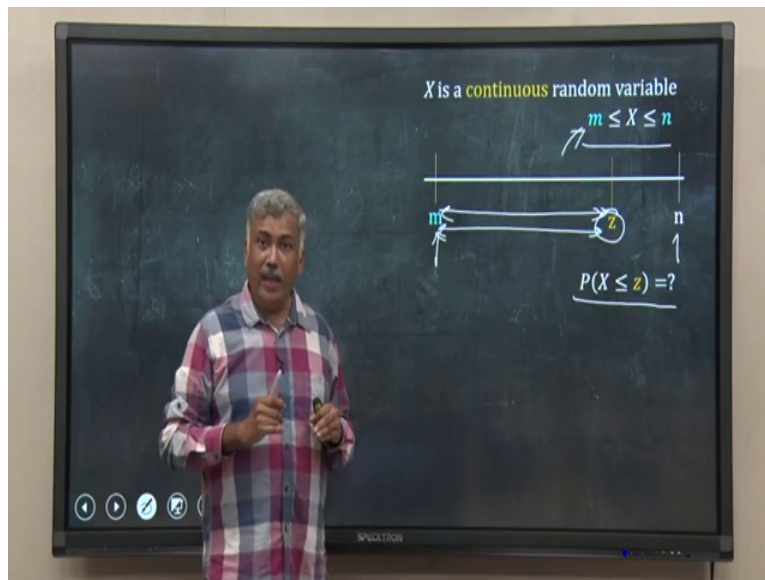
$$\lambda \cdot e^{-\lambda X}, \text{ when } X \text{ is positive}$$

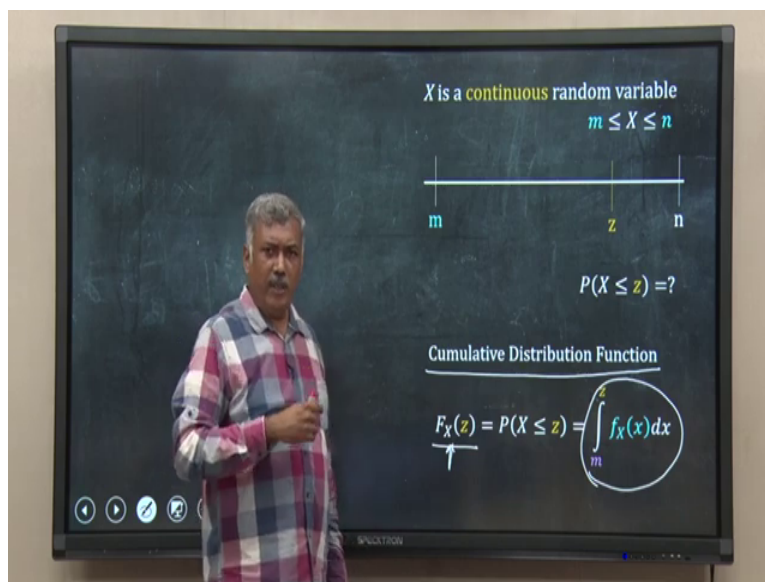
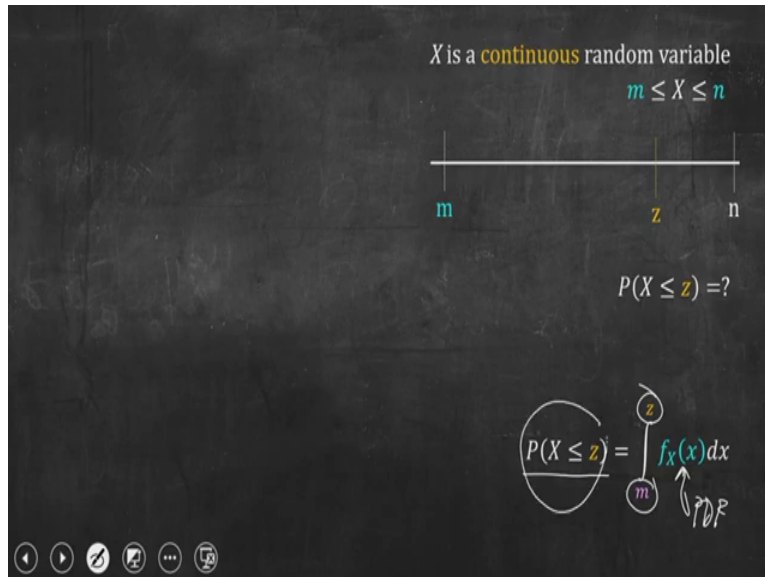
0, otherwise

So, if I plot it, what I have done here, I have plotted for two values of rate parameter,  $\lambda = 2$  and  $\lambda = 0.5$ . So, you can see we have exponential decay for both the cases because we have  $e^{-\lambda x}$ . So, it is the exponential decay and the rate of the decay is varying. So, obviously, the slope by which these curves are falling are also different.

Now, we have discussed till now, in this lecture about what is probability density function, the meaning of that, then we moved into three commonly used probability density function in data analysis and statistics, in particular in Biology, Normal Distribution, Uniform Distribution, and the last one Exponential Distribution.

(Refer Slide Time: 33:03)





Before I end this lecture, let me discuss another term called CDF. Cumulative Distribution Function, you may have seen or heard in the textbook. So, what is that? So, suppose I have a  $X$  is a continuous random variable, like the weight of children, or height, or age, something like that. And that varies between  $m$  and  $n$ . So, in the number line I have  $m$  here, I have  $n$  here, and I have another number  $z$  in this interval  $m$  to  $n$ .

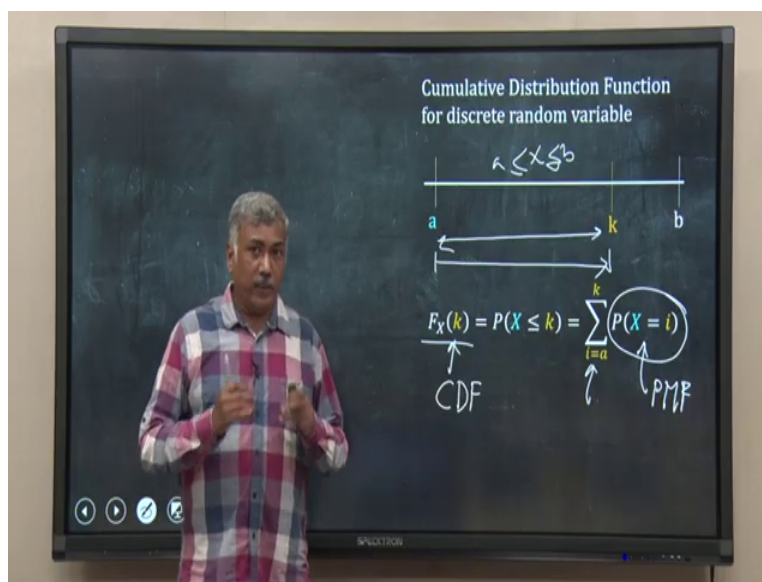
So, this is  $z$  somewhere. And I want you to calculate the probability that the random variable, my measurement will lie between  $m$  to  $z$ , in this interval  $m$  to  $z$ . So, I have to calculate,  $P(X \leq z)$ . I am not to say it is bigger than  $m$ , because we have already said that  $X$  cannot be lesser than  $m$

here. So, I am saying the probability that  $X$  will be less equal to  $z$ . To calculate this actually, what I need is called cumulative distribution function.

So, what I will do to calculate this probability, I will take the PDF and integrate it from  $m$  to  $z$ . So, that will give me probability that  $X \leq z$  and we will call this probability as cumulative distribution function. This capitals are  $F_x$ , this is cumulative distribution function. Explicitly,

This is nothing but the integration of PDF from the lowest possible value of the variable up to that value  $z$ . Now, you must be thinking this is for, what we are discussing is for continuous random variable, we have not discussed cumulative distribution function and for the discrete one in the last lecture, but actually, you can map it very easily. We will also have a cumulative distribution function for a random variable, which is discrete.

(Refer Slide Time: 35:00)



In that case, I will do a simple thing, I will not do integration, I will do simply summation. Because it is discrete, and I have not used PDF, I will use the PMF Probability Mass Function, so, I will take all values between  $a$  to  $k$ , because the variable  $X$  is varying from  $a$  to  $b$  in this example, so the  $X$  is less equal to  $b$ , greater or equal to  $a$ . So, I will take from  $a$  to  $k$ , I will take all discrete values possible for  $X$  and get their probability using this PMF probability mass function and then sum them all. That will be my CDF. That was the last thing for this lecture. So, let me now jot down what we have the key points of this class.

(Refer Slide Time: 35:58)

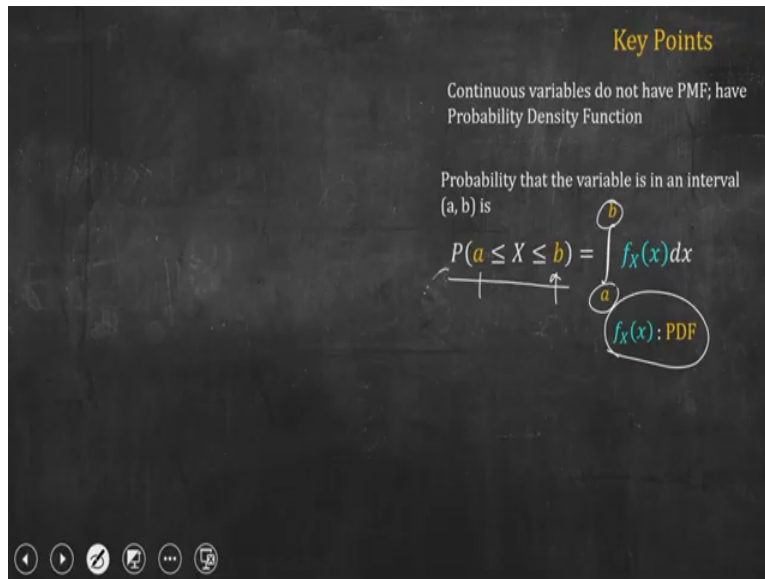
**Key Points**

Continuous variables do not have PMF; have Probability Density Function

Probability that the variable is in an interval (a, b) is

$$P(a \leq X \leq b) = \int_a^b f_X(x) dx$$

$f_X(x)$ : PDF



**Key Points**

Continuous variables do not have PMF; have Probability Density Function

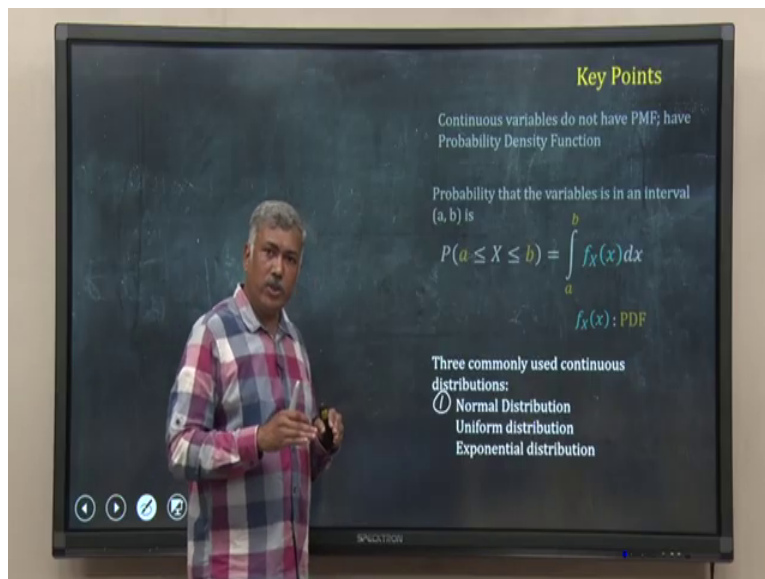
Probability that the variables is in an interval (a, b) is

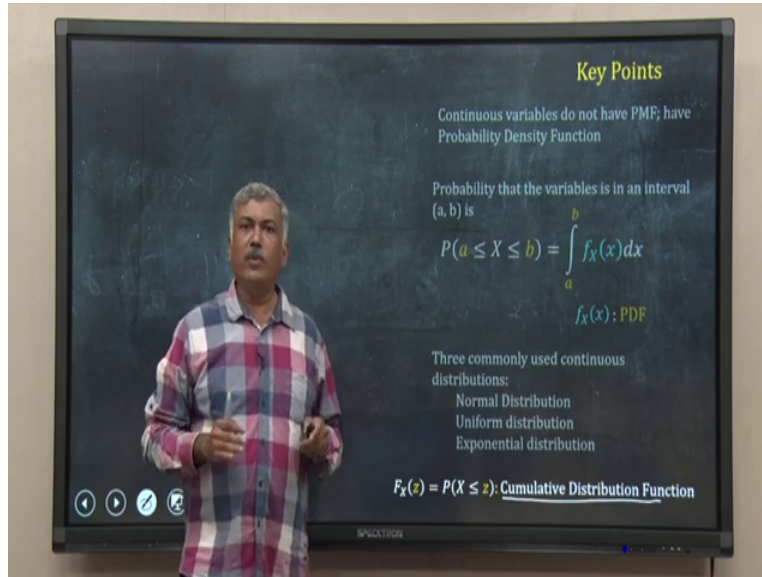
$$P(a \leq X \leq b) = \int_a^b f_X(x) dx$$

$f_X(x)$ : PDF

Three commonly used continuous distributions:

- ① Normal Distribution
- Uniform distribution
- Exponential distribution





So, what we have learned in this lecture, we have learned about that continuous variable, continuous random variables do not have PMF, it does not make sense to talk of probability mass function of a continuous random variable. Rather, if I have to calculate the probability, I have to ask what is the probability that, that random variable lies in a interval, for example, a to b?

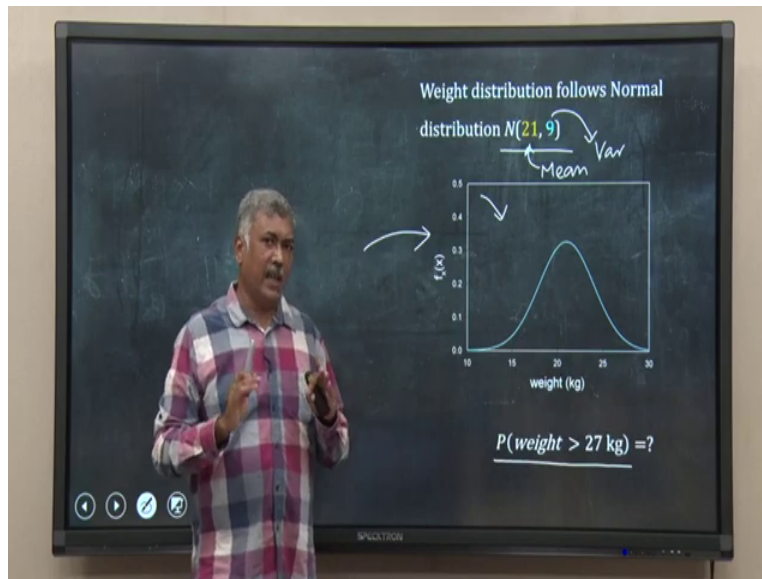
And to answer that question, I require a function called probability density function, which is not a probability but it is a density function. If I integrate that from a to b, I will get the probability, that we have learned. Then we moved into learning three commonly used continuous probability distribution one is normal, which we have discussed in length.

For example, we have discussed the 68, 95, 99 rule, we have discussed how to calculate probability using the normal distribution for a particular question. For example, probability that X will be bigger than that particular value something like that, we have also discussed about standard normal distribution or z-distribution, all these things are widely used in data analysis and statistics.

Apart from Normal Distribution, we have discussed about Uniform Distribution, and we have also discussed about Exponential Distribution. And lastly, in this lecture, we have introduced Cumulative Distribution Function, which is applicable for both continuous as well as discrete distributions. So, that is all for this lecture, but before we stop, let me leave you with a problem to solve.



(Refer Slide Time: 37:43)



So, this is again the problem related to the weight of children. So, suppose, we have collected a lot of data, we have done lots of lead research and we know that the weight of the children, the school going children in that particular cohort follow a normal distribution with  $\mu$  Mean, this is the Mean, Mean equal to 21 and the variance, variance equal to 9, and I have made a plot of that.

Because it is normal distribution, I know the PDF, the equation for PDF for normal distribution. So, I plugged the value of 21 and 9 and I asked the software to draw the plot. So, the plot is given to you. Now, the question that you have to deal is that, what will be the probability that weight is greater than 27 kg, the weight, is the  $P(\text{weight} > 27 \text{ kg})$ ?

Given the information that the weight follows a normal distribution with a mean 21, variance 9, try to solve this one. To solve this one by shortcut, what you have to use, you have to use the famous thumb rule of 68 95 99 percent. That is the clue, try to solve it. See you in the next lecture. Till then happy learning.