

Data Analysis and Decision Making - I
Prof. Raghu Nandan Sengupta
Department of Industrial & Management Engineering
Indian Institute of Technology, Kanpur

Lecture - 10
Distribution Function

Welcome back my dear friends; a very good morning, good afternoon, good evening to all of you. This is the DADM 1 course under the MOOC NPTEL series and we are covering the topics of Statistics in all its depth and this is the 10th lecture. So, as you know this is a course for 30 hours, 12 weeks spread over 12 weeks and each week we have 5 lectures, each lecture being for half an hour and 10th week means we are going to finish the 2nd week.

So, I am Raghunandan Sengupta from IME Department, IIT Kanpur. So, if you remember that in the 9th lecture just at the last few slides, we are discussing about continuous distribution. We did discuss about exponential and then, we went to normal distribution. Now, one thing if you remember when we are discussing the discrete cases, I did mention about the Poisson distribution and I did mention another important point is that then, it is basically the number of such readings you are going to take per unit time and that per unit time and mentioned time and again conceptually obviously, and mentioned the 2nd point also which I will come.

So, it was basically between any unit time that unit time can be 1 hour or can we say for example, 30 minutes, can be 2 seconds, 1 nano seconds. That is generally what I am trying to hint at what you mean by the concept of per unit time, but the actual distribution comes out to be true as that delta time tends to 0. So, hence the Poisson distribution comes out. Now, when I was discussing the exponential distribution, I did not mention that point, but I will try to definitely bring it into the discussion now before I again continue.

So, say for example, you are standing in front of a queue or you are basically trying to process materials or machine objects or some jobs which are coming at in for being processed in a machine, or you are standing in front of a teller machine or a teller window in the bank whatever it is or say for a counter, where food is being served whatever it is. Now, you have a stop clock in your hand and you measure two things

simultaneously. Consider that as possible for you number 1, you measure the number of people who are entering coming per unit time in that unit time you have to decide as I said and you are also measuring the time taken to process each and every unit which is entering.

So, say for example, it is a bank teller machine. A person is sitting and people are entering the queue per unit time. So, that would be discrete and as you decrease the unit time, that discrete distribution becomes PMF which is actually the Poisson distribution. Now, if you go to the case where you are trying to find out the amount of time taken to process each and every customer that may be somebody must have, may have brought his or her passbook to be updated, somebody wants to withdraw money, somebody wants to get a demand draft, somebody wants to basically deposit money. So, obviously everybody you would have different times to process the request. So, if I basically consider the time, then it basically becomes an exponential distribution.

So, there is a very underlying linkage between Poisson and exponential distribution. I will come to that later on also. Exponential distribution can also be utilized for the cases where you want to basically find out the working life of a machine, working life of say for example, bulb whatever it is. An exponential distribution is memory lost process in the sense that whatever the probability of failure, a probability of working was though instant basically, you want to again start counting; it would have no information set from the past. So, it is basically memory loss process.

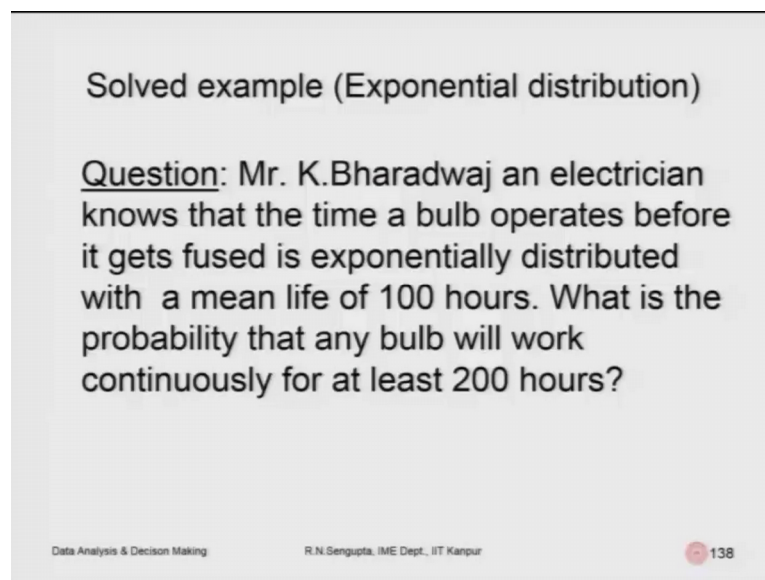
Another thing a very interesting fact is that I should have mentioned during the discussion, a Poisson distribution, but Poisson distribution basically came into the light a very interesting historical fact is that during the Persian war, so where cavalry was used as the main thrust for the worse. So many soldiers were killed by being trampled by the horses as they fell down. So, it was found out that the number of people who died in such accidents was basically the distribution was, distributed as Poisson distribution. So, there that is one of another historical reason. So, there are other historical backgrounds and also, but I thought I would mention to you that this.

Now so, these are some of the I would not say the missing link, some of the information set which I thought should be shared with you, so that it will basically be much more interesting for you to understand the concept. Now, if you remember I did discuss the

normal distribution. Normal distributions have the same mean. So, the same mode and the same median; mode if you remember is the value which is occurring the highest probability wise or frequency wise or relative frequency wise. Number 2 point is mean is basically the center of gravity or center of mass of that distribution.

So, mean mode are the same for the normal distribution and median is that value of x which divides the total distribution in two equal halves. When I mean these two equal halves, I basically implies that the probability that the cumulative probability or additional all the probability values till that value of mean median and after there, median are equivalent and they are equal to 0.5. So, let us consider first the case of one exponential distribution.

(Refer Slide Time: 06:14)



Solved example (Exponential distribution)

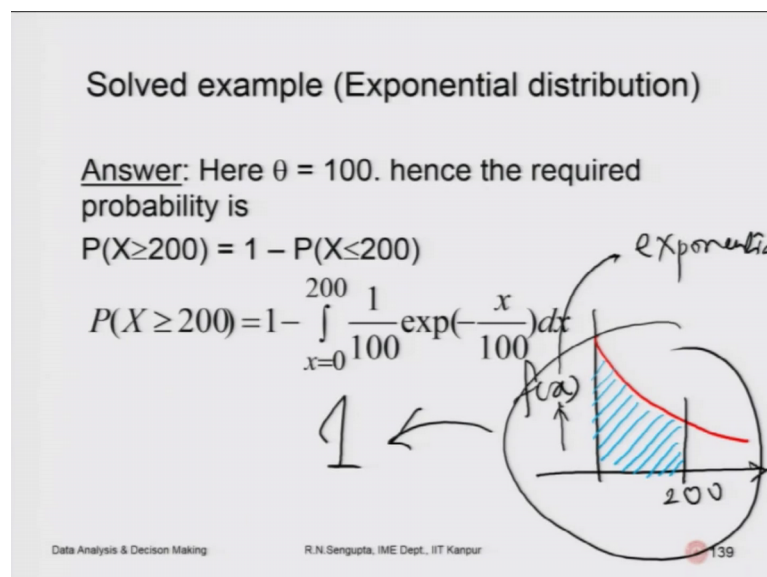
Question: Mr. K. Bharadwaj an electrician knows that the time a bulb operates before it gets fused is exponentially distributed with a mean life of 100 hours. What is the probability that any bulb will work continuously for at least 200 hours?

Data Analysis & Decision Making R.N. Sengupta, IIM Dept., IIT Kanpur 138

So, consider Mr. K. Bharadwaj an electrician knows that the time a bulb operates before it gets fused is exponentially distributed with the mean life of 100. What is the probability that any bulb will work continuously for at least 200 hours? Now, if you remember for the exponential distribution, I did discuss there are two parameters λ and μ . Now, a value is basically the minimum working life of that particular component for which you are trying to find out the distribution and the distribution exponential. So, if there is a bulb, consider that and the bulb may find the instantaneously the moment you put on the switch.

So, in that case A would be 0. So, in the example which you are going to consider the mean value which you know the formula is A plus lambda, here A would be 0. Hence, lambda is 100. So, based on fact we will basically proceed. It may be possible that for some machines where you want to find out the exponential distribution being used for the case of trying to find out the working life of that machines, but there is a warranty life. That means, the machine would never fail before say for example 6 months. So, in that case A would be 6 and you will basically calculate the corresponding probability and all the inset of some information which is required based on the fact that A is 6 months or half a year. So, in this case theta is 100.

(Refer Slide Time: 07:34)

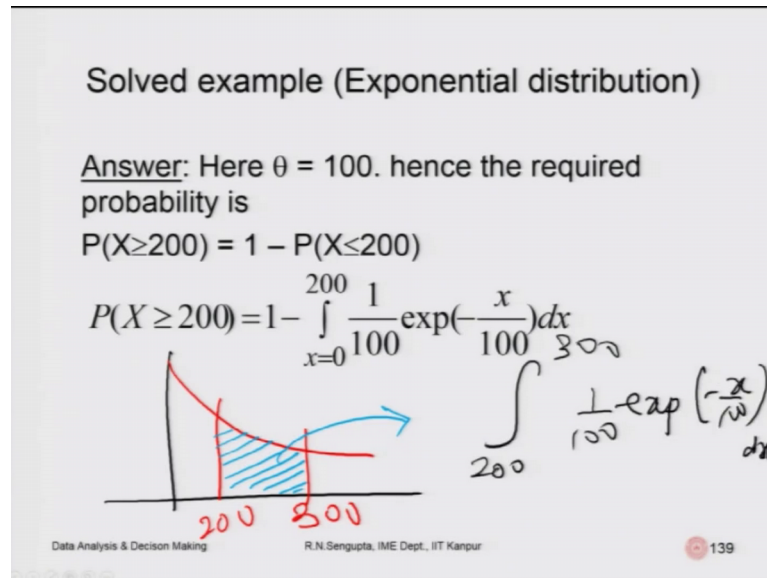


Obviously A is 0; hence the required probability would be X. Probability of X being less than is greater than 200 or it will be 1 minus probability of X being less than equal to 200. So, here less than and greater than would not make much of a difference. I will come to that later on. So, hence we want to find out the probability we integrate it. So, when you are it and just use the color so, this is the access the distribution is this. So, consider again you diffuse the black color. So, this is say for example, 200 and this is f of x; f of x is basically the exponential distribution.

So, you will integrate all the values. So, let me use again that another color blue. So, I will find out all the pdf summed up. So, the cumulative values and this whole area, that means starting from 0 to infinity, this whole area is basically 1. So, 1 minus that blue

area hashed area would give me the corresponding property. So, that what is given here say for example, if I want to find out, so obviously I will be coming to that. I want to find out the probability that the bulb will survive between your time period of 200 to 300 hours. So, obviously you need to do the calculation. Accordingly I will draw the picture that will make things much clearer to you.

(Refer Slide Time: 09:31)

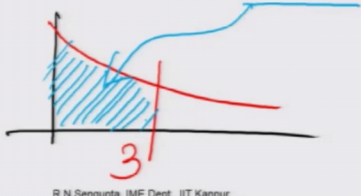


So, this is let me draw it in black, ok. Let me use the red one. So, this is 200, this is 300. So, the bulb is surviving between this time period. I need to find out this. Let me highlight it. So, this would be given by the corresponding area between 200 to 300 $\frac{1}{100}$. I am taking the same distribution exponential minus x by ω dx you find out and then, to your calculations accordingly on assignment again.

(Refer Slide Time: 10:21)

Assignment (Exponential distribution)

Question: Mr. Lyndoh the owner of an electronics shop knows that the average life of a tape recorder he sells is 1.5 years. What is the probability that any such tape recorder would function for at most 3 years?



Data Analysis & Decision Making R.N.Sengupta, IIM Dept., IIT Kanpur 140

An assignment I am telling it is based on the fact that they are to be solved by you, not to be created from our side. So, it is Mr. Lyndoh, the owner of an electronics shop knows that the average life of a tape recorder he sells is 1.5 years. What is the probability that any such tape recorder would function 4 at most 3 years? So, again we are considering that it may be possible the moment you put on the tape, it would have failed. So, obviously A would be 0. So, in that case you will consider λ or θ whatever it is equal to 1.5. The recorder would have function for at most 3 years and that means, I want to find out the total probability which is 3 years and beyond. So, it will be the distribution, this is the value of 3 years. See if I draw it, it becomes this. So, I want to find out till infinity.

So, we want to find out what is the probability that any such tape recorder would function for at most, ok. At most 3 years would be on the left hand side my apologies. So, I did not read it properly. So, at most 3 years would basically beyond on the left hand side, so maximum 3. So, in that case the erase it. So, it will be easier not to be confused yeah. So, in that case I use the pen. So, this is the probability.

(Refer Slide Time: 12:08)

Log-normal distribution

$[X \sim \text{LN}(\mu, \sigma^2)]$

$$f(x) = \frac{1}{\sqrt{2\pi\sigma_X^2}x} e^{-\frac{(\log_e x - \mu_X)^2}{2\sigma_X^2}} \quad 0 < x < \infty$$

- μ_X, σ_X^2 are the parameters where $\mu_X \in \mathbb{R}$ and $\sigma_X^2 > 0$
- $E[X] = \exp(\mu_X + \sigma_X^2/2)$
- $V[X] = \exp(2\mu_X + \sigma_X^2)\{\exp(\sigma_X^2) - 1\}$
- Example: Stock prices return distribution

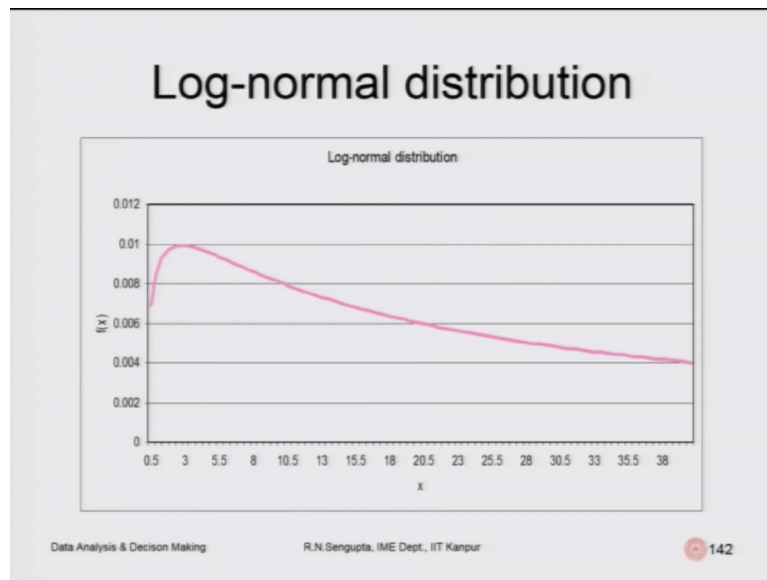
Data Analysis & Decision Making R.N.Sengupta, IIM Dept., IIT Kanpur

141

Now, we consider the log normal distribution. So, X where X is the random variable, LN is the log normal distribution and μ and σ^2 are the parameters. So, X value is for the log normal distribution. Remember log of values of negative are not possible, hence log of X value would be from 0 to infinity. μ_X and σ_X^2 , again the suffix basically denotes they are the parameters, where μ_X is a real on long the real line and σ_X^2 would obviously be greater than 0. The exponential, the expected value and the variance values are given as shown.

So, examples are stock prices returns distribution. They can be considered accordingly. So, in the log normal distribution look at the distribution pdf and it would be almost equal to the normal distribution. So, obviously the first part I will use the blue. Let me use the color black. So, this part is the same and obviously, you would have 1 by X and E to the power in case of the normal distribution, you had μ minus σ^2 . So, this is a log of X minus μ whole square divided by 2 σ^2 . So, these values are only different and based on that you can find out the log normal distribution.

(Refer Slide Time: 13:35)



This is the log normal distribution and again along X axis, your X values along the Y axis, the pdf values and you can calculate the corresponding probabilities whatever they are as required.

(Refer Slide Time: 13:49)

Relationship between Poisson and Exponential distribution

If a process has the intervals between successive events as independent and identical and it is exponentially distributed then the number of events in a specified time interval will be a Poisson distribution

Data Analysis & Decision Making R.N.Sengupta, IME Dept., IIT Kanpur 143

Now, if we remember I did discuss about just when I had just started the class and on the slide, the 10th lecture number was there, I did not mention the relationship between exponential and Poisson distribution. So, here it is. I will again repeat it. If a process has the intervals between successive events as independent and I did mention that, that idea

also. So, intervals and the arrivals time and the numbers coming or independent of each other between two intervals and intervals are of same equal length.

So, if a process is the intervals between successive events as independent and identical, so obviously these are ids and it is exponential distributed, then the number of events in a specified time interval will be a Poisson distribution. So, the relationship being numbers arriving is Poisson and the time taken to process them is exponential.

(Refer Slide Time: 14:52)

Cumulative distribution function (cdf) or the distribution function

We denote the distribution function by $F(x)$

$$F(x) = P(X \leq x) = \sum_{x_i \leq x} f(x_i)$$

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(x) dx = \int_{-\infty}^x dF(x)$$

$P\{a \leq X \leq b\}$ v/s $P\{a < X \leq b\}$

Data Analysis & Decision Making R.N. Sengupta, IME Dept., IIT Kanpur 144

The cumulative distribution function of the cdf, we denote the distribution function by F of x , capital F of x and F of x being the case where you have the probability of x less than equal to x , that is given by the summation of all the values for x till that value. So, if you basically you want to find out cdf values for the case when its the uniform distribution and it will between say for example, one to say for example, 10 and there only integers, I want to find out the cumulative distribution for the values less than equal to say for example, 3.

Then, I will add up the pmf values for value 1 plus 2 plus 3. If I want to find out the value of the cdf values for x values greater than 5 in the same case, so it will basically be the sum of the pmf values of 6 because 5 is not included and obviously, 5.5 5.7, all these things are not possible because of the discrete distribution. So, in that case I will add up the probabilities of value of 6, 7, 8, 9, 10 if I want to find out the values of the cumulative distribution for the uniform discrete case. For the case when it is inclusive of

3 and exclusive of 5 less than 5, but greater than equal to 3, then I add up the probabilities pmf values of 3 and 4 only. See in the case if it is a continuous distribution, so obviously in that case summation will be replaced by integration.

So, when I showed you when you have the exponential distribution log, normal distribution, in that case I will utilize the concept of integration, find out what is the cumulative distribution values for x between two values. So, if I write technically obviously I am not going to go in there details, but technically when I write for the continuously case and for the discrete case, the formula like probability of x between a and b and a and b both been inclusive. So, if I have say for example, this let me write it here. So, it is easy for so, if I write probability I am using Pr for probability.

Please here it is not written. It is only P . In the discrete case, it will mean that I include the probabilities of pmf of a and b both. That is one. So, in the discrete case, this versus this are different because in the later case, the value of let me highlight it. So, this value where e is not included would definitely change the overall cumulative values, but for the normal case it would not be case in the same because consider this I am trying to explain it in a very layman term for us to understand because in that case, there at between the value 1 and 10, there are infinite value.

So, whether you start from a or start from a plus delta, that would not matter and then, you basically some slight shift on to the right of a would not matter. So, the values which have written here considering x is between a and b both, the inclusive and x is between a and b , a not been inclusive for the case of the discrete case, it will be different for the case of the continuous case. It will be same.

(Refer Slide Time: 18:50)

Properties of distribution function

- 1) $F(x)$ is non-decreasing in x , i.e., if $x_1 < x_2$, then $F(x_1) \leq F(x_2)$
- 2) $\lim_{x \rightarrow -\infty} F(x) = 0$ as $x \rightarrow -\infty$ $F(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} f(x) dx$
- 3) $\lim_{x \rightarrow +\infty} F(x) = 1$ as $x \rightarrow +\infty$
- 4) $F(x)$ is right continuous $F\{x_1 \leq X \leq x_2\}$

$$\frac{dF}{dx} = f(x)$$

$$\frac{F(x_2) - F(x_1)}{x_2 - x_1} = \sum_{x \in [x_1, \dots, x_2]} f(x)$$

Data Analysis & Decision Making R.N. Sengupta, IIM Lucknow

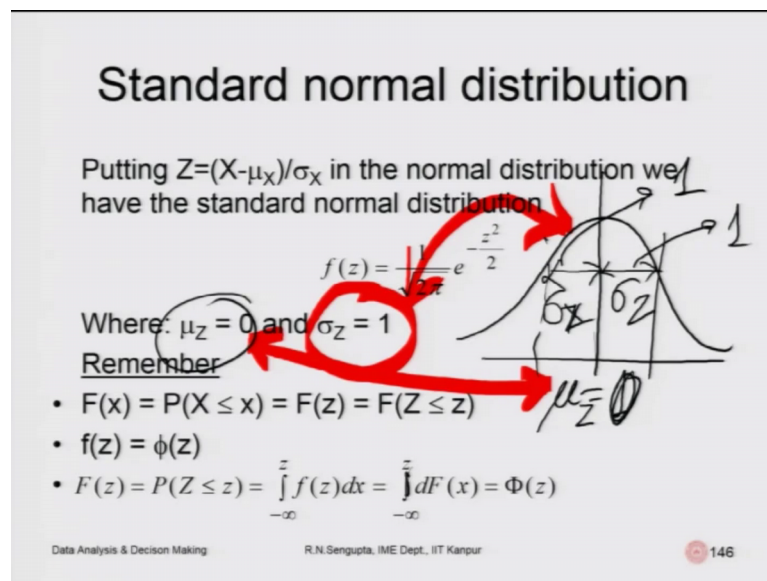
So, properties of the discrete of the distribution function, this distribution function basically means it is the cdf value I am talking about and I am using capital F of x. So, small f of x is basically the pmf for the pdf. So, F of x is a non decreasing in x that is if x 1 is less than x 2, then f of x 1, capital F of x 1 is less than equal to x 2 because for any pdf values of pmf values, the values cannot be 0. So, they are either increasing or at this fixed value.

So, if it is less than it means increasing less than equal to equality sign would be for true for the case, where it is true that there they are basically equal. Now, remember one thing if you consider the cdf values, what does cdf value means, I am basically adding up whether cumulative or the integration. Now, if you do the reverse mapping, that means you want to find out the value of small f given capital F. So, obviously it will be very simply we can understand that as F is equal to the integration between some values consider x 1 to x 2 f of x dx. So, in that case capital, let me expand capital X f 2 for your and for our better understanding.

So, it will be given as f of x 1 less than equal to capital X less than equal to small x 2. So, this would be true and obviously, when I want to find out the probability of x 1 less than equal to capital X less than equal to x 2, then in that case I sum them up. So, it will be all values of x, which is element starting from x 1 till x 2 both being included. So, if further less than equal to sign less than equal to sign, now when I go the reverse way, so

obviously it would mean df of dx for the case when it is continuous. So, let me use continuous case. So, obviously this would be true or else in the case many of the discrete case, this is the continuous case. The discrete case would be f of x 2 at x is equal to minus f of x 1 divided by x 2 minus x 1. So, this will give me the pmf value at that point. When it changes from x 1 to x 2, the standard normal distribution would be utilized in all the problems.

(Refer Slide Time: 22:07)



So, putting Z is this standard normal deviate is equal to X minus mu by sigma and using very simple one dimensional Jacobian transformation. So, it can be done for multivariate case. You will use the multivariate Jacobian transformation putting Z is equal to X minus mu divided by sigma which is basically the normal distribution. We will have the standard normal, but in this case the corresponding standard normal distribution would have not mu as the mean, but it will have 1 as 0 as the mean value and it will not have sigma square as the variance. It will have 1 as the variance of that Z for which obviously, the standard deviation value would also be 1. So, if you look at Z distribution, it will look at the diagram is going to come within few minutes, but I will just draw it.

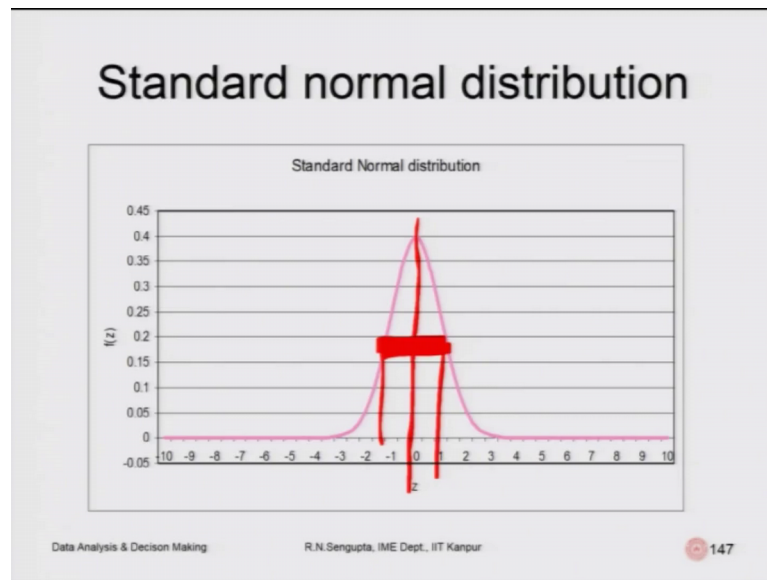
So, this would be the standard normal, this would be the mean value which is 1 as I mentioned that. So, this would be the mean value corresponding to the standard normal deviate and the variance which you will have, oh sorry. This will be 0 my apologies. So, it will be 0 and the variance values for the standard normal deviate. So, this is sigma Z

this is sigma Z, this value would be 1, this value would also be 1. So, plus minus 1, you will have. So, here new Z is the color change would help. So, this is what I am talking about and when you have this, this is the value which I am talking about. So, remember F of x which is the cdf value is for x, probability x being less than equal to x, you will also have the same values as Z less than equal to Z because you are doing one to one transformation.

So, F of z which is the pdf value for the x as the normal distribution, the corresponding pdf would basically be given by small phi of z and when we have the cumulative distribution values. So, F of x, that means you add up all the values for x starting from minus infinity to 0. So, obviously the corresponding values in the case when you have the standard normal deviate would also be from minus infinity to the small z value, where x and z have one to one correspondence that would be given by capital Z. So, remember one thing if you look at Z distribution, the mean value being 0 which means the values equidistance on to the left and the values on equidistant onto the right are equal. That means, it is a symmetrical values are there on the left hand side and the right hand side.

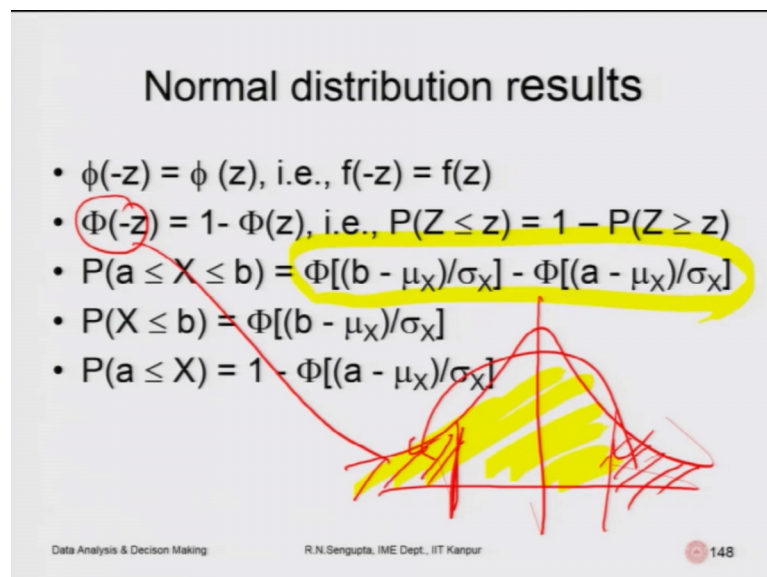
So, if you go plus 1 sigma on to the left and plus 1 sigma on to the right, the corresponding Z values would be equally magnitude, but they would be of opposite sign because if you are considering from I am my side I am considering. So, 0 is here if I go left, you will have basically minus Z values. If we till in minus infinity if you go to the right, it will go to plus infinity with all positive values.

(Refer Slide Time: 26:08)



So, this is the Standard Normal Distribution, this is the 0 value, this is basically I would have variance here. So, that is basically plus minus 1 and again along the x axis, e of the x values along the y axis, e of the corresponding pdf whereas, being drawn in a way where you can find out the correspond and n equal plot the corresponding pdf for the standard normal.

(Refer Slide Time: 26:38)



So, normal results in distribution rules, the value of small phi minus z corresponding the fact that it is equally on to the left equally on to the right. So, small phi minus z, it will be

equal to small phi plus z because let me draw it. So, this is here. See if I go minus 2 or minus 3 on to the left and if I go plus 2 plus 3 to the right, the values of the height because small phi means basically corresponding to the probabilities. So, they would be same. So, this value is equal to this value, this value is equal to this value and so on and so forth. The corresponding capital phi minus z, so that means I am taking let me delete it. Let me delete it. I am taking the pen.

So, if this is z minus z is here. So, if I find out this, so this would be all the values till minus infinity to plus infinity. If I get positive 1, so obviously these values are already equal so, I have 1 minus this would basically give me the corresponding probabilities which I am trying to find out for the case of this. So, these values would basically give me the cdf values starting from minus infinity to plus z. So, 1 minus the corresponding value on to the right would give me the cdf values from minus infinity to minus z.

Probability of a being less than x being less than minus beta, so obviously when you do the transformation in the standard normal case, it will be x becomes z. So, the corresponding values of a would become a minus mu x divided by standard deviation of x and b would become b minus mu x standard deviation of x and if a tends to minus infinity, then the actual corresponding for values of capital phi would basically be 0. So, in that case probability of x less than beta would basically be all the values added up from minus infinity to beta value and if beta becomes positive infinity, so obviously you would add up all the values from minus infinity to a and 1 minus that would give you this value. So, I will again come back to using this problem.

You will understand it in much details as you do the assignments and understand and read the book. So, with this I will close the 10th lecture which is the end of the 2nd week and wish you all the best of luck.

Thank you very much.