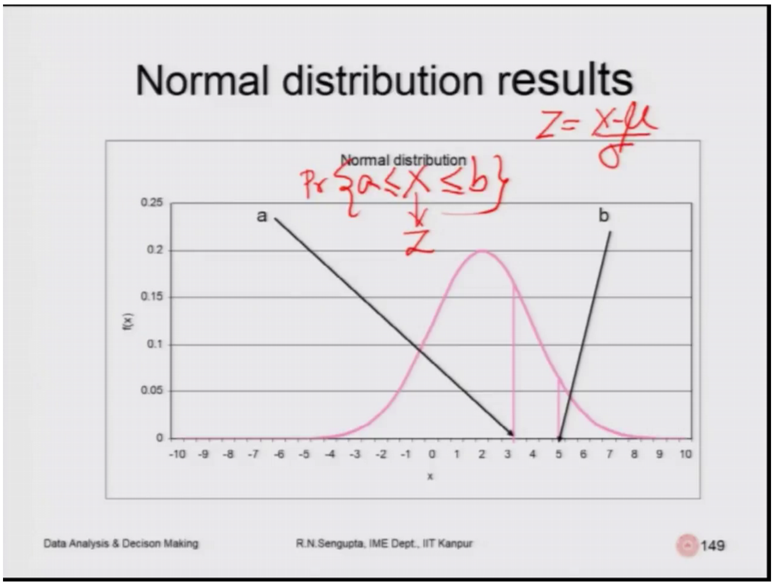
Data Analysis and Decision Making – I
Prof. Raghu Nandan Sengupta
Department of Industrial & Management Engineering
Indian Institute of Technology, Kanpur

Lecture – 11

Very warm welcome to all my dear friends and students, a very good morning, good afternoon, good evening to all of you. I am Raghu Nandan Sengupta from IME department, IIT, Kanpur. And this is the DADM-1, which is the Data Analysis and Decision Making-1 course under the NPTEL, MOOC series. This is a course for 30 hours, which is 12 weeks. And each week as you know and there are 5 classes, each being for half an hour.

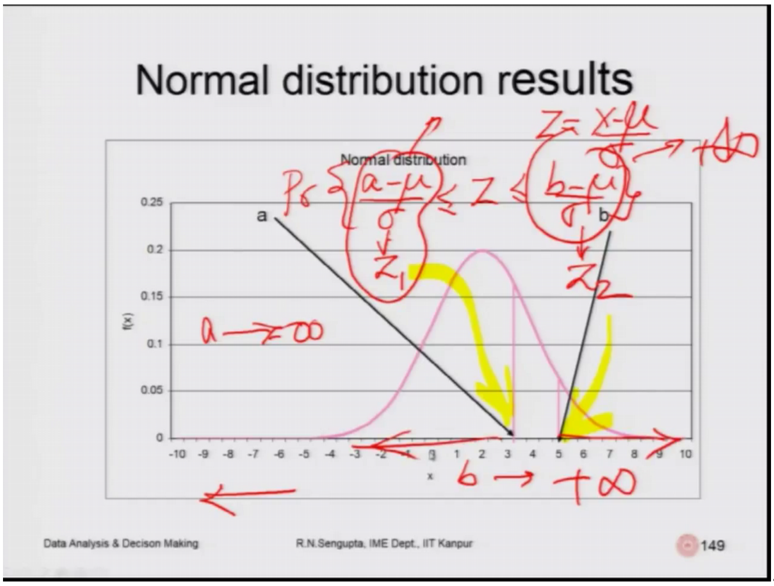
So, we will start the 11th class, which is basically we are going to start the 3rd week, 1st class, because 5, 5 - 10 are already over. So, if you remember that, we were discussing about the normal distributions. And I did say that what does the small pi, capital Pi mean the if X is normally distributed, how it can be converted into a standard normal deviate using Jacob simple Jacobian transformation. And for the case, where X has a mean value of mu X, and variance as sigma square, which two are the parameters for the normal distribution. It can be converted it into a standard normal with Z, where the mean values Z is 0, and the variance is 1.

(Refer Slide Time: 01:51)



Now, consider the normal distribution. And if you remember also I have the in the last slides of the 10th class, I had written this, which I will write again for let me use the red color, so it will be clearly visible in this slide. So, what we analyze was the probability of a being less than equal to X being less than equal to b we have to find this. So, what technically we do is that we convert X into Z. So, when you convert that you use the formula Z is equal to X minus mu. So, this mu is basically the mean value of X by sigma. Correspondingly, the a value also gets converted, b also value also gets converted. So, I am going to write that, erase that considering there is no space. I will erase it, and write it accordingly.

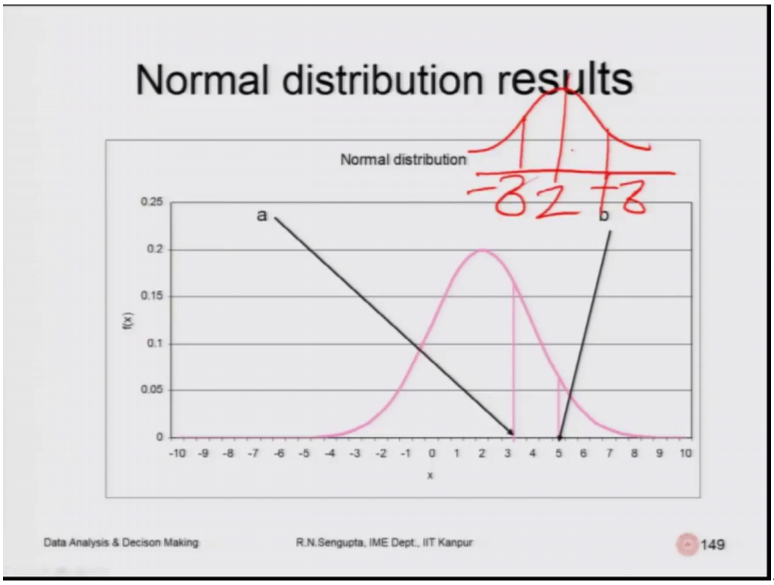
(Refer Slide Time: 02:28)



So, probability a now becomes a minus mu by sigma less than equal to Z. So, Z as already been converted less than equal to b minus mu by sigma. So, this would be Z 1 a realize value. This would be Z 2 a realize value. So, technically this Z 1 is the transformed value for a. And this Z 2 is the transformed value of b.

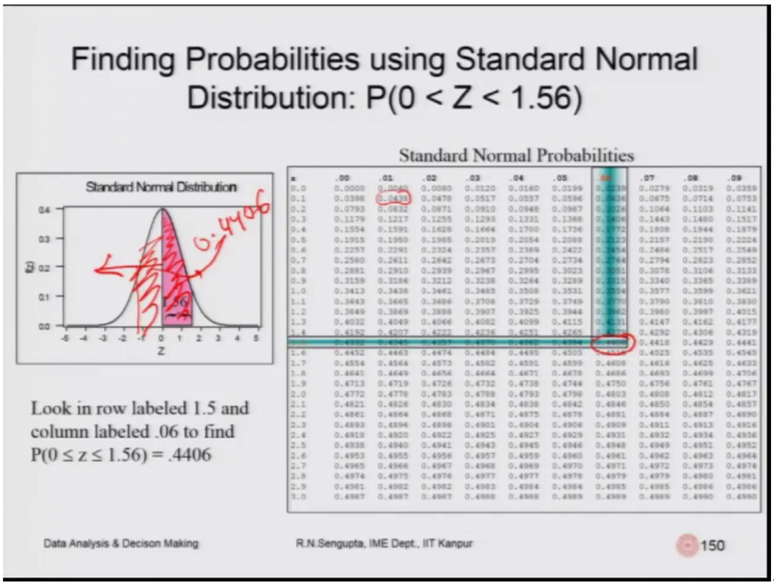
Now, as you know we did mention that in the limiting cases a tends to 0. The probability or the value, actually this value basically becomes it sorry, if a tends to minus infinity my apologies is minus infinity that means, it is tending towards the extreme left hand side in this case. so, obviously this will vanish in the case, when b is tending to plus infinity, this will actually become plus infinity, so that the corresponding probability will that means, this is shifting to the right. And this is shifting to the left. So, we can find out the values.

(Refer Slide Time: 03:48)



So, this being a standard normal deviate being Z. The value onto the right hand side, if it is plus Z, the value on the left hand side would be minus Z, because they are equally dispersed symmetric.

(Refer Slide Time: 04:01)



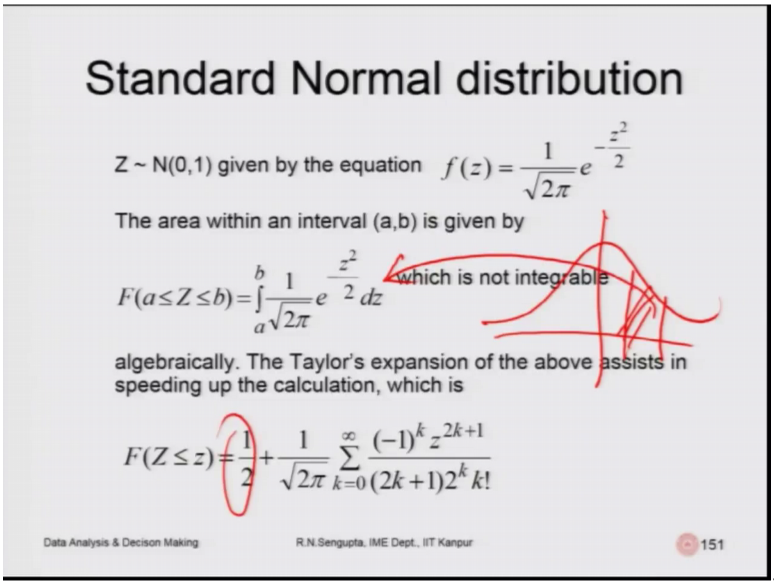
So, now we want to find out the standard normal deviate and the corresponding CDF values. CDF is the cumulate distribution values based on that we can find on the probability. So, concentrate on the on the diagram on the normal distribution on the left hand side; and corresponding standard normal probability, which I given on the right hand table. So, now what you do, this pink area will actually means this is the overall CDF value for finding out the summation of all the probabilities starting from 0 to a value of 1.56. So, if that is the value, how do we find out?

Now, come and obviously it would also mean, that if we move on to the left that means this side, then the corresponding CDF value or the sum of the probabilities from moving from 0 to minus 1.56 or in another way of moving from minus 1.56 to 0. The overall area would be exactly equal to the pink area, which is shown. .

Now, let us refer to the tables. So, the standard normal table has Z values along, so I am just hovering my pen over that Z value starting from 0.1, 0.2 till 3.0. And the decimal places, which means that it is 0.000 that is second decimal 0.0102 till the last value is 0.09. So, if I want to have a value of 1.1s 0.11, then the corresponding Z value would be the not the Z value. So, I am saying the corresponding cumulative values would be this, this, which technically means that if I find out the whole area starting from 0 to 0.11 the overall cumulative value is 0.0438.

Now, let us come back again to this diagram, which is which shown in pink the shaded region. And try to find out that what is the corresponding probability distribution sum that means the CDF value. So, the value is 1.56. So, let us go to 1.5 first. So, this is 1.5. This is 0.1.56. So, the corresponding value, which you have here is 0.4406 that means, total area of this pink shaded region is given as 0.4406. So, the corresponding value if I go on to the left hand side, it is from minus 1.56 to 0 again the overall area here also would be 0.4406.

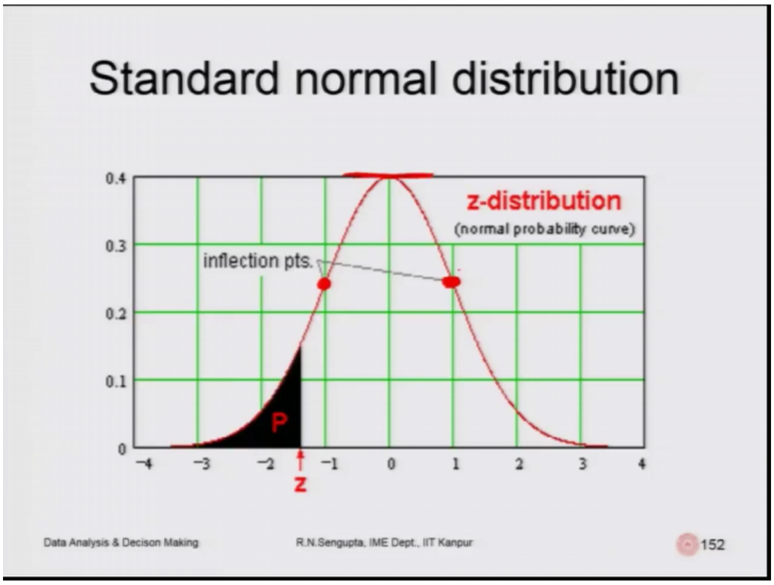
(Refer Slide Time: 06:45)



So, standard normal distribution has a mean value of 0, variance of 1 is given by the equation f of z, 1 by square root of 2 pi, because sigma is 1, so obviously it would not be there. And in the numerator what you had in the in the actual normal distribution is e to the power minus in the bracket X minus mu whole square by 2 sigma square. So, mu is 1, so obviously in that case, mu is 0, so obviously in that case it will be z square.

And in the denominator, you had basically 2 sigma square. So, sigma is 1, so it will be 2. The area within an interval a, b is given by this equation. So, you want to basically find out the overall area a, b. This area is given by the integral from a to b 1 by 2 pi e to the power minus z square by 2 d z, which is not in integrable algebraically. So, hence many of the cases, we use Taylor series expansion of the above in order to basically do the speed calculation. So, this is the Taylor series expansion based on which we will do. So, this half basically is coming, because it is on the left hand side we have already covered, from minus infinity to 0, which is 0.5.

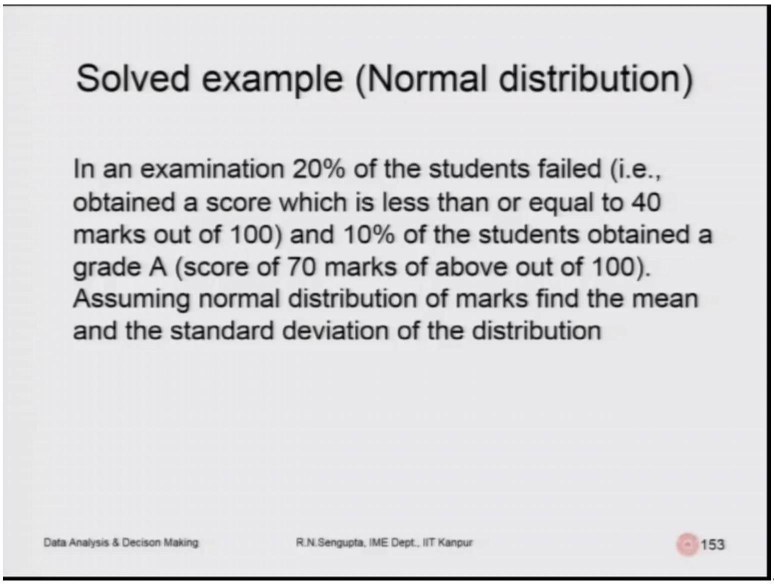
(Refer Slide Time: 08:13)



Now, in standard normal distribution, if you move plus 1 sigma minus 1 of sigma to the right on the left respectively, the overall point at which the graph changes. Now, what do I mean by graph changing is that d y, d x of that point is changing. So, if I continue finding out the slope from this point, at the mean or the mode or the median the d y d x is 0 that means, obviously the second derivative would be negative, because that is why you find out the maximum value.

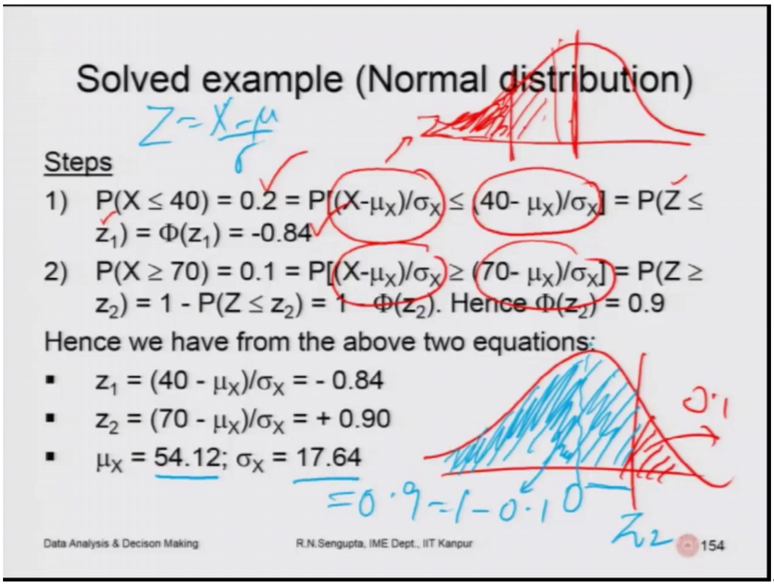
Now, there are two points at which the rate of change of the function changes from positive to negative or negative to positive that means, the second derivative is also 0, that those points are the point of inflection. So, the point of inflection for the normal distributions are plus sigma on to the right, and plus sigma on to the left. So, at this point, which are the black one, which I am now highlighting with red. So, at this point the rate of the change of the function changes from positive to negative, negative to positive.

(Refer Slide Time: 09:22)



So, let us consider a very simple example. In an examination 20 percent of the students failed that is obtained of score of obtained a score, which is less than or equal to 40 marks out of 100. And 10 percent of the students obtained a grade A that is score of 70 marks on a verb assuming normal distribution of the marks find the mean and the standard deviation of the variance of the distribution. So, we are considering the marks of the students are normally distributed.

(Refer Slide Time: 09:53)



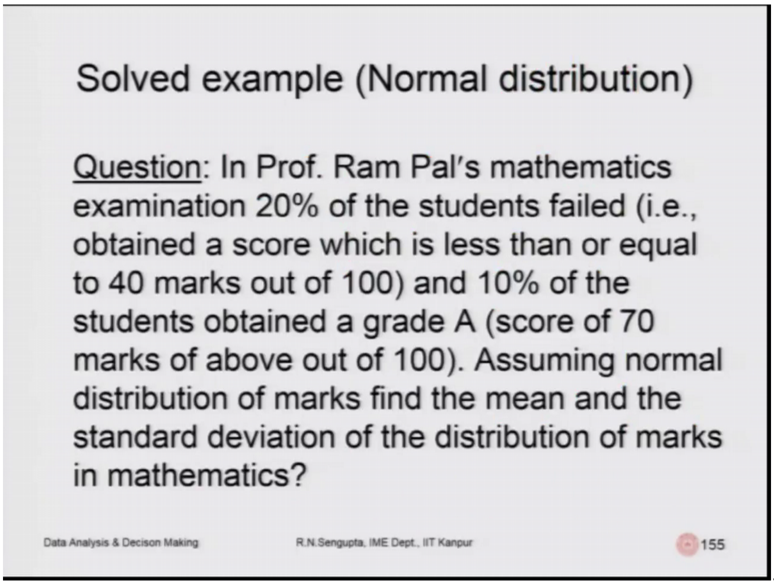
So, when we can consider probabilities of less than 40, it is 0.2. So, if I consider X has a mean value of mu, and a standard deviation sigma square, so obviously I do this transformation. So, this becomes Z. Corresponding value is 40 minus mu divide by sigma square. So, this is Z, and this is small z 1. So, in that case what I am doing technically, I am trying to find out the integration of all the values starting from minus infinity to that value of Z, which corresponds to 40. So, this is the value, which I find out.

And thus is given as when I at that is given as 0.2, so if I basically find out 0.2, because if you remember, till the mean values is 0.5. So, obviously the areas of till the value, where you calculate the value of 0.2, you would be on to the left. So, hence the negative value, so it comes out to be minus 0.84. For the case, when X is greater than 70, so that means on the right hand side, the values is given as 0.1. So, in this case greater than 70 onto the right is 0.1.

So, again we transform this becomes Z. This becomes Z 2. So, in that case if it is 0.1, which is the overall area if I add up here, use a different color, it will be easy. So, the area onto the left consider this is Z 2 would be 1 minus 0.1, which is equal to 0.9. So, if it is 0.9, so hence again from the table, I can find out Z 2. So, I found out Z 1 and Z 2. So, there are two equations, I find out Z 1 and Z 2. Now, correspondingly we know Z is equal to X minus mu by sigma, put Z 1, put Z 2, two equations, two unknowns solve then.

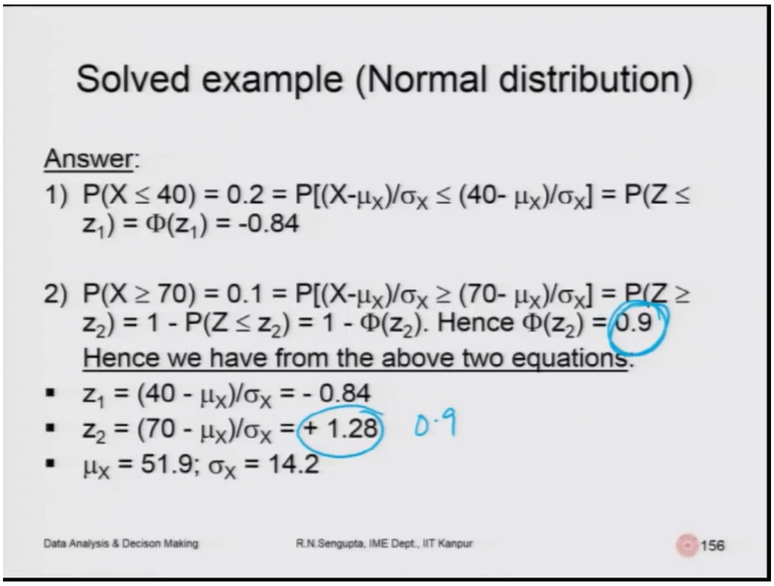
So, Z 1 comes out to be minus 0.84, and Z 2 comes out to be plus 0.9, because you are moving on to the right. So, this is the mean value which is 0. So, this is this is already this whole area is still 0 is 0.5. So, you have to add 0.4. So, 0.4 basically makes up such the case, you can find out the Z 1 and Z 2 value so. Based on that you find out the values of Z 1 and Z 2 and once you find out Z 1 and Z 2, you solve and find out that mu values is 54.12. And the variance, so square root of that gives me 17.64.

(Refer Slide Time: 12:50)



Consider a Professor Ram Pal's mathematics examination. So, this an example, 20 percent of the students failed the same thing. So, I am putting into the different words obtained a score, which is less than equal to 40, 10 percent obtains a grade a score of 70 and above. Assuming normal distribution, you solve it.

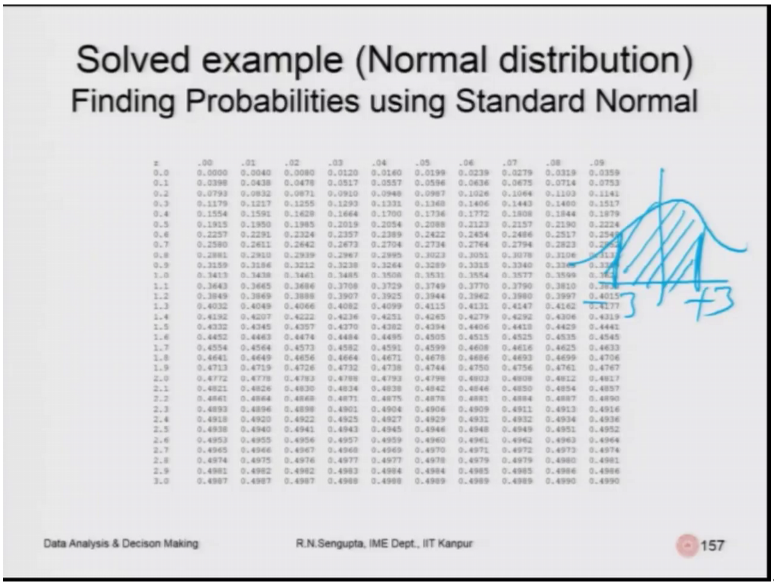
(Refer Slide Time: 13:09)



This is the same thing, so ok only one error. So, there the values is one point plus 1.28. This is 0 not 0.9. 0.9 was basically the overall probability. So, Z 1 is minus 0.84. Z 2 is

1.28, based on that you find out mu and sigma, which is 51 approx I am given the approx value, hence 14.

(Refer Slide Time: 13:35)

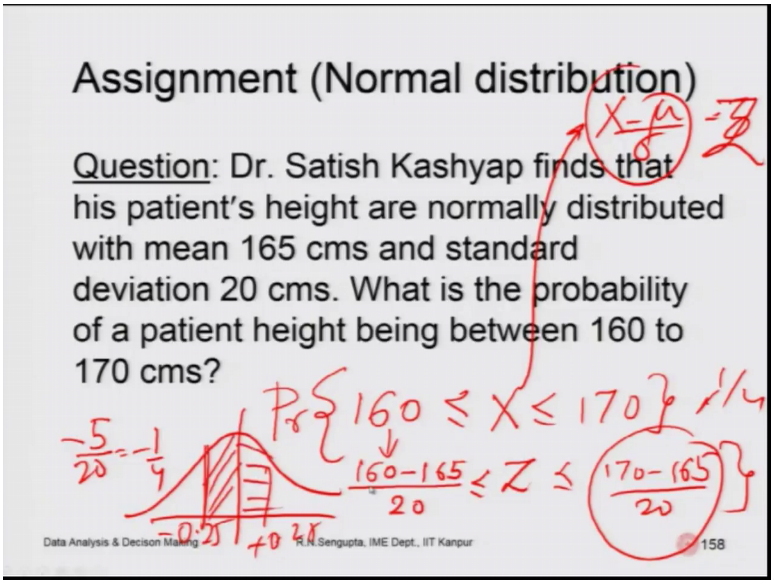


z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	[illegible]
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	[illegible]
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	[illegible]
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	[illegible]	[illegible]
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	[illegible]
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	[illegible]
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990

So, this is the solved for that case, you can use the standard normal deviate table. Only remember the standard normal deviate table is given on to the if it is only given for the half the distribution, you can easily used that for the full distribution. In many of the cases, the actual integration values based on, which the values are given are starting from one negative value to a positive value, such that you have to basically divided by 2 2 find out the overall area.

What I mean is this, so this is 0. So, it starts from negative value to positive value. And the tables are values are given accordingly. So, if it is given as say for example, minus 3 to plus 3 the overall area is given here somewhere in the table, you divide by 2 find out the (Refer Time: 14:25) area for the integration from minus 3 to 0 or from 0 to 3, what whatever you look. So, this 3 minus 3 and plus 3, I am taking as an example. So, it can be anything.

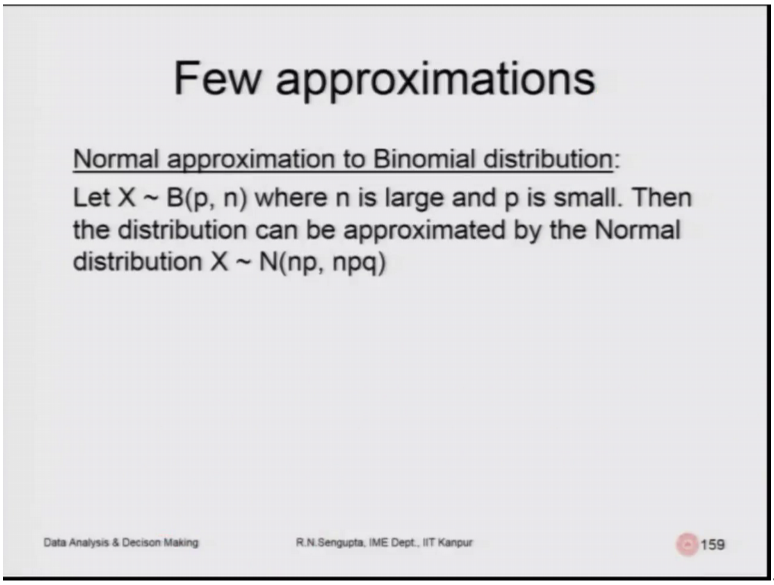
(Refer Slide Time: 14:38)



So, this is an assignment again not to be solved as an assignment in the examination just to be tried by you. Dr. Satish Kashyap finds that his patient's height are normally distributed with mean 165 and standard deviation of 20. What is the probability of a patient height being between 160 and 170? So, the mean values is given 165. So, I want to find out let me use the red color. So, I want to find out this, so what I do, I transform this into Z. So, it will be 160 minus 165 standard deviation 20 less than equal to so it become Z, because I have now transform X by, so this is equal to z, no sorry, so it is capital Z.

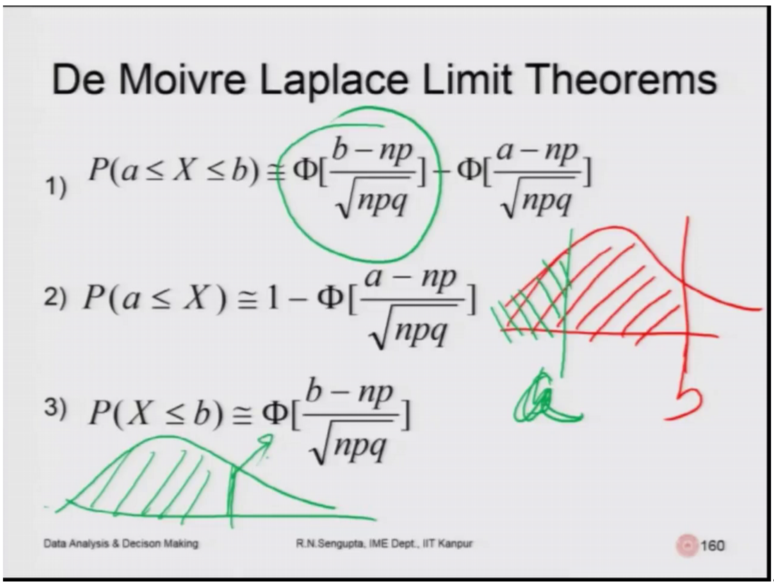
So, these are values here as well as the small Z, 170 minus 165 by 20. So, this becomes, so this value first is minus 5 by 20. This is minus 1, 4. This becomes 5 by 20 is basically 1, 4. So, at point minus 0.25 plus 0.25 so, you have the distribution from minus 0.25 to 0.25 you go, and try to find out the area.

(Refer Slide Time: 16:13)



So, normal approximation to binomial distribution let X B n p. So, binomial being parameters p and n, when is n large and p is small. Then the distribution can be approximated by the normal distribution given, where the mean value would now be the mean of the binomial distribution, which is N n q p. And the variance, which is for the manual cases is n p q will be utilized at the variance for the normal distribution, such that mean value and the variance for the binomial distribution become the mean value and the variance for the normal distributions.

(Refer Slide Time: 16:50)



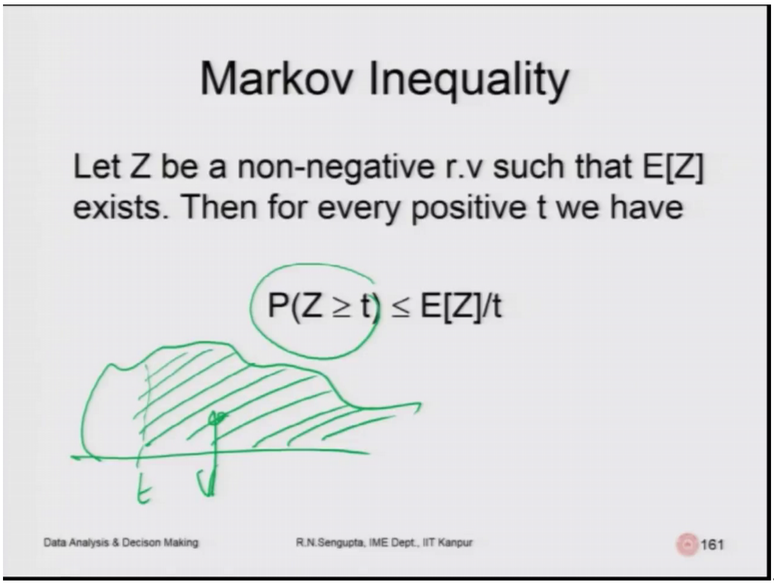
De Moivre Laplace limit theorem. So, in this case, so this is what what I just discuss for the binomial one. So, probability of X being between a and b both inclusive. So, so what I am trying to basically do is that, try to convert a discrete binomial distribution to the normal case, which is continuous that is the essence and how we replace that that is the main value, main cracks. .

Now, in that case when we sorry, so in that case and that case, when we converts X into Z, it will be X minus mu divide by sigma so, what is mu, mu is the corresponding mean value of the binomial distribution, which is n into p. And the variance is n p q. So, the square root of that will give me the standard deviation. So, it will capital sigma b minus n p divided by square root of n p q.

And for the case, when we find out for the b case, it will be because b is on to the right hand side. So, I have to add up much more values to get to do that, so it would be b minus n p again divide by square root of n p give. So, those formulas remains the same, only that we should remember the addition that means, all the probabilities I add up for b is from minus infinity to b. And for a it is minus infinity to this point in between that area will give me the overall area, which I want to find out.

If b tends to plus infinity, then obviously the capital phi with b becomes 1. If a tends to 0, so obviously you have to find out the corresponding values. So, if 8 oh sorry my mistake this is a, so, a tends to is 0, so obviously you add up all the values from minus infinity to b. So, this is capital phi as written.

(Refer Slide Time: 19:01)



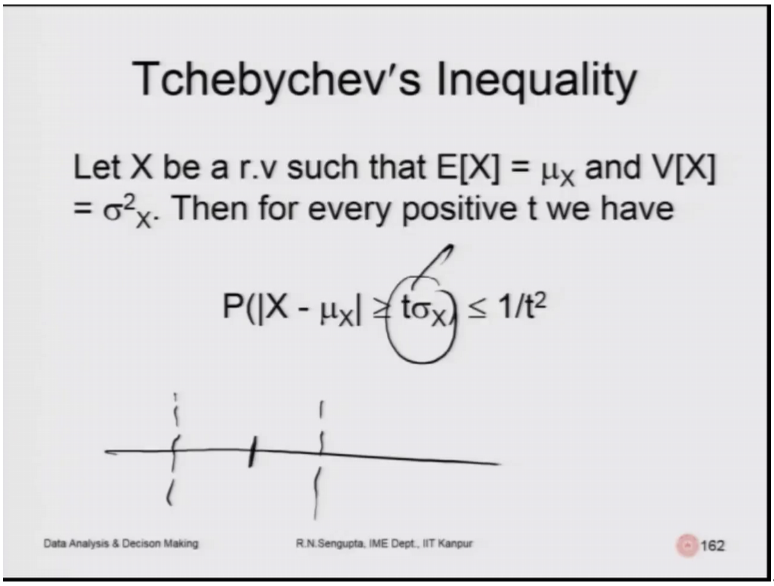
Will considered two inequalities. So, one will be this consider Z is a non-negative random variable. So, it does not mention anything about the variance, it only mention that, Z is Z is not the standard normal deviate remember. So, Z is a non-negative random variable such that the expected value exist. It does not mention anything about the variance. Then for every positive t, we will have the probability of Z being greater than a number t is less than equal to the expected value divide t.

So, what we are trying to do here, trying to find out some bound that means, corresponding to a distribution Z, which is non-negative. First moment exist, we are considering that the bound for that particular random variable to be greater than some fix value t. T is variable, but it can change, but we know beforehand is less than equal to the expected value. So, expected value this is the basically centre over gravity of the distribution.

Whatever the distribution is, so considered the distribution is and the centre gravity this. So, centre over of gravity divided by t. So, what we are trying to do is that, the probability is multiplied by the distance, distance from the for the case, so it is like this let me. So, consider the t value is here. So, the probability of Z value being greater than t, multiplied by t so that means, t is the distance on the X axis that would be given by the expected value.

So, what we try to basically portrait is that how number one we can portrait, the maximum value of the limits of the probability. And also we given a different expression for the expected value in terms of the probability, and the distance function. Distance I am just giving a nomenclature, it does not actually mean distance it basically a positive numbers.

(Refer Slide Time: 21:18)

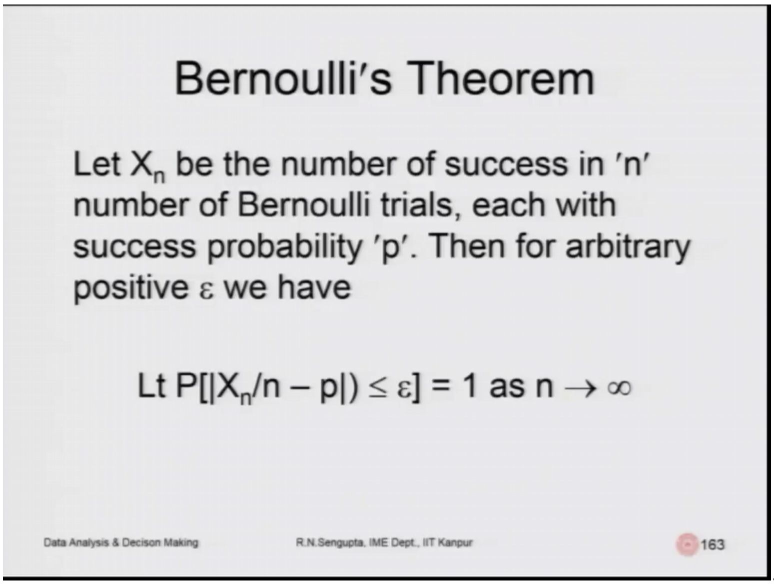


In the case of Tchebychev's inequality, let X be random variable. It does not mention anything whether is a positive or negative remember that. The first moment exists. So, mu X is nothing to do with the normal distribution is just a symbol being used for the mean value. Variance is sigma square again nothing to do with the standard normal of the normal distribution. Then for every positive t we have a bound. So, X minus mu mod of that is basically the fluctuation of X over the mean and under the mean, so that value would be greater than the value multiplied, any t value multiplied by the standard deviation.

And that probability will be always less than equal to 1 by t square. So, if you basically shrinked, that means make that overall mod value as small as possible. Then, obviously the corresponding probabilities would change. So, so remember that what we have actually, let me use the, this is the mean value, this the mod, plus and minus. So, I am trying to find out the overall probability that this is greater than t sigma.

So, if t is 1, so in that case probability value being greater than the standard deviation is less than equal to 1. So, obviously here also you are trying to put a bound based on which we will basically try to find out the probability between the limits. Limits means the mod value. So, consider t is 2 in that case the probability of the bound being less greater than equal to 2, sigma is less than half. So, so as you basically expand or contract the bounds the corresponding maximum are so called minimum values for the probabilities are given by the value t. t is also again not random, it is you can change, but it is a known value.

(Refer Slide Time: 23:12)



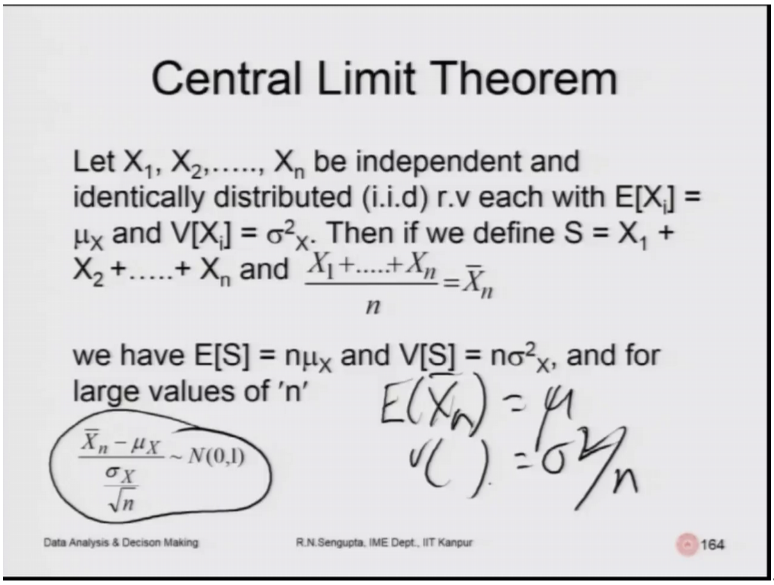
So, Bernoulli's theorem says that, if X n, so now X n is basically the number of successes, which we have so this n technically means the number of samples. Let us X n be the number of successes in n number of Bernoulli's trials, each with success probability p and an unsuccess or (Refer Time: 23:30) probability being 1 minus p or q. Then for arbitrary positive epsilon, so epsilon would basically depend on a n value.

We have as n tends to infinity that means, as n tends to in increases. Then the corresponding difference between the actual probability and the relative frequency in the long run start decreasing, and exact equals less than equal to some epsilon value, that is epsilon value can be very small. It can be 10 to the power minus 13, 10 to the power minus 14, 10 to the power minus 36 whatever it becomes smaller and smaller tends towards 0, such that that probability tends towards 1 as n tend to infinity.

So, let me give an example. We know, if you toss a coin probability of head and tail for a unbiased coin not the Sholay coin, which we see where both the faces are head. And unbiased coin has the probability of head as half, tail as half. Now, do think of this in this way, if you toss the coin, so the relative frequency is of the heads would be number of heads divided by the number of tosses you are doing. Relative frequency of tail would be number of tails divided by the by the number of tosses. Keep increasing the number of tosses.

So, we will get relative frequency of the head also and the tail also. In the long run as n increases, then the difference between the relative frequency of the head minus 0.5 would tend towards the difference would tend towards 0 with probability 1. Similar, in the relative frequency of the tail, number of tail divided by number of tosses minus the actual probability this is 0.5 that will tends towards 0 with probably 1, so that is the essence. So, it can be thought out for any of the examples, and it can be assumed accordingly.

(Refer Slide Time: 25:26)



So, we will consider the central limit theorem, until we utilize time and again. So, let X 1 to X n be independent and identically distributed i. i. d random variables. So, what do you have is that you are picking up any observation for any distribution. And you pick up the first one keep it. Do such n number of picking 1, 2, 3, 4 till n, do not worry about with replacement or without replacement, because the population is so large.

Now, note down the values, I can put them either in the box or you do not, because it is in infinite population. Again pick up, put them on the table from X 1 to X n note them down, either replace them or through them away, continue doing it. And do it infinite number of times. So, what will come out is that, if you know the distribution of that population, then it consider it is mean mu. Now, consider each of the picking, which you are doing, groups of picking you are doing, and only concentrated on the first picking. .

Technically, in the long run that place can be filled up by any one of those actual values, which are there in the box. If that is true, then if I want to find out the probability of the first picking, it is intuitively exactly equal to mean. Similarly, if I concentrate on the second one, two infinite such pickings only concentrate on the second one, the actual probability of the second picking is also mean values is also mu. So, in the long run if you continue infinite expected value of X i is mu, similarly in the variances of X i is also sigma square, because you are doing infinite one.

In that case, when I pick up the sum, so sum is basically X 1 till X n and I want to find out the mean of that sum. So, if I consider the sum, it is X 1 plus X 2 till X n, and if I want to find out the expected value of the sum, it will be expected value of X 1 plus expected value of X 2, dot, dot, dot till expected value of X n, because there are i.i.d because there are independent identical distributor, because they are infinite number of such observations.

Now, the expected value for each picking is mu. So, hence the expected value of S would be mu plus, mu plus, mu how many times, n number of times. So, the expected value S becomes n into mu. Now, I want to find out the expected value of S, which is the sum. So, what I do, I have found out the expected value of S, which is n into mu. The average value of actually for X 1 to X n is S by n. So, the expected values is n into mu divided by n. So, it again becomes mu that means, the expected value of the sample is exactly equal to the population value, which is mu. .

Now, if I consider from the case or the variance, now remember for the variance, I will come to that later on mode discussion the ways variances are also the pickings are also i.i.d. So, the covariance is the concept, which I am using for the first time will come to that later on would not exist. So, what you will have is basically only the principal diagonal in a n, n by n matrix. So, you will basically add up sigma square n number of times, but the actual formula for the variance is a square. So, n square root come, when you try to find out the mean value, would n square would come in the denominator. Hence, the actual value on the variance for the that average would be sigma square n into sigma square by n square, which will become sigma square by n.

And finally, if I put the actual distribution of X n, which is the mean values, so it will as I have mentioned, if the meanwhile you would be mu, the variance will be sigma square by n. So, When I convert that mean value considering a normal distribution, it will be what we know it is X minus sigma, which is the average value X divided by standard deviation X.

So, in this case sigma remains sigma. And the standard deviation is what the variance was sigma square by n, so the standard deviation will be sigma divided by square root n, so that would basically be a standard normal deviate as given. Because, we know the expected value of X n is the bar as the average is mu. And the variance is given by sigma square by n. So, finding out the square root would give you the corresponding thing. I will come to that those results later on more details. So, with this I will close the 11th class.

And wish you all the best and thank you very much for your attention. Have a nice day.