INDIAN INSTITUTE
OF
TECHNOLOGY
KHARAGPUR

NPTEL
National Programme
on
Technology Enhanced Learning

Probability Methods in Civil Engineering

Prof. Rajib Maity

Department of Civil Engineering
IIT Kharagpur

Lecture – 24

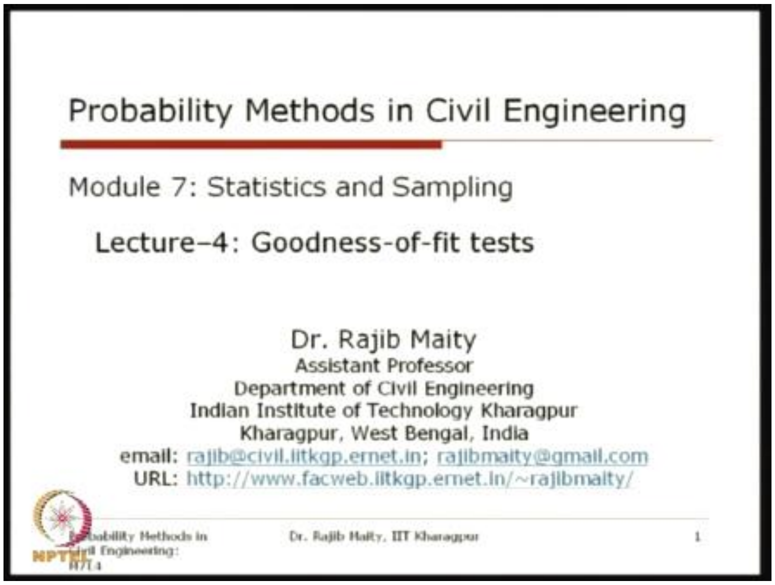
Topic

Goodness - of - fit tests

Hello and welcome to this lecture in this lecture we will learn some of the statistical methods to test the observed data, whether they are fitting to some particular distribution or not you know that in earlier lectures as well as in earlier modules also we have referred several time that while handling some problem we generally assume that the dataset that is following a particular distribution and we have seen now we will see that how we can test these things that whether the dataset is really following this distribution or not.

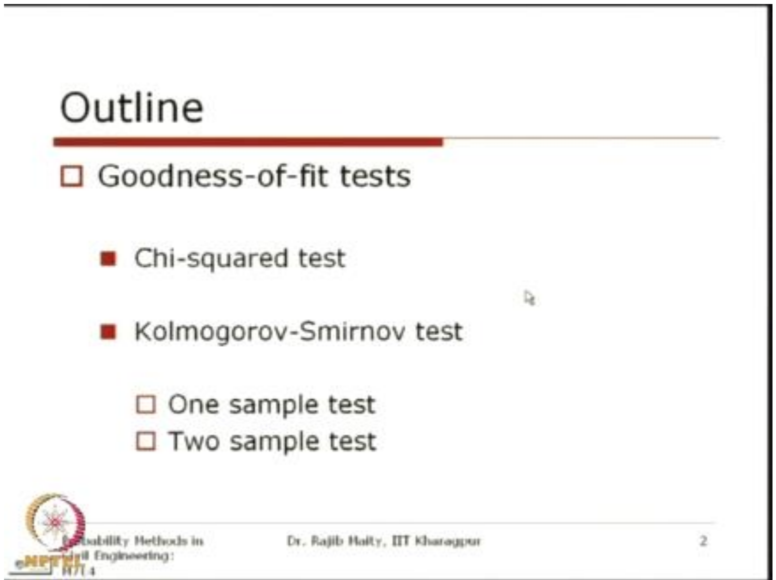
There are some of the statistical test that I am just going to tell in this in this lecture is that so to use that one and use the knowledge of this hypothesis testing will be testing it statistically that how the data is fitting to a particular module. So there are some tests and so we will just pick up some this popular test to see that how we can test.

(Refer Slide Time: 01:26)



This goodness of fit test this is test are known as the goodness of fit test. So, this is so that is our today’s lecture title is the Goodness of fittest.

(Refer Slide Time: 01:35)

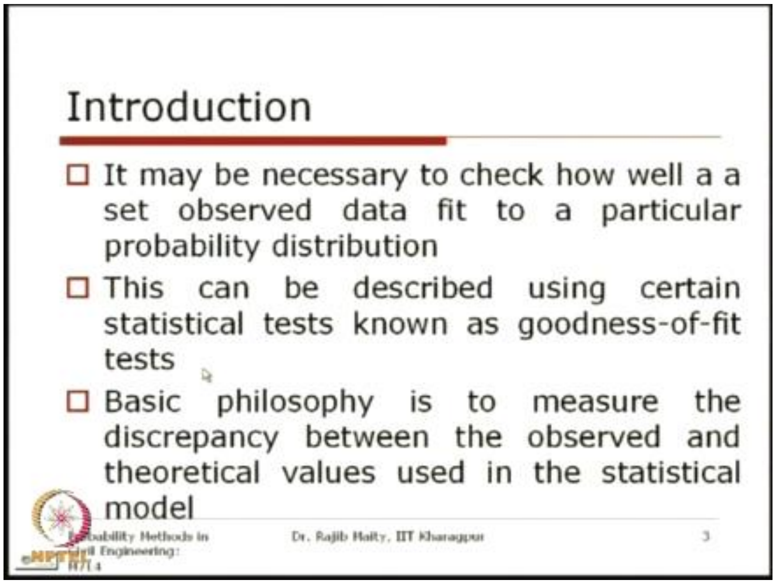


So, first we will just see that how this can work and after that we will take two test for this lecture that one is that chi square test this will be chi square, not d, d is mistake that is a chi square test and other one is the Kolmogorov Smirnov test it is also known as this KS test as an abbreviated form and this KS test again can be that one sample test and two sample test basically when we take that this one sample test then we generally test it for one sample that we take and we generally take that from whether that particular dataset follow a particular distribution or not.

So in this case one is our sample and other one is the standard distribution and when you go for the two sample test it is basically both are our sample data is there and we look for the answer whether both the samples are from the same population or not so when the same population means there are following the same distribution or not. So this is how we go for one sample test and this two sample test as you know that for this kind of statistical test we need some kind of significance level and some statistical significance level.

So whenever we generally conclude or we draw some decision from whatever the statistical test we do we have generally that decision is associated with statistical significance level so that is important so, what significance level that we are considering either before and that okay at this significance level whatever we are going to test is satisfied or not in a statistical sense that is what we will do well.

(Refer Slide Time: 03:42)

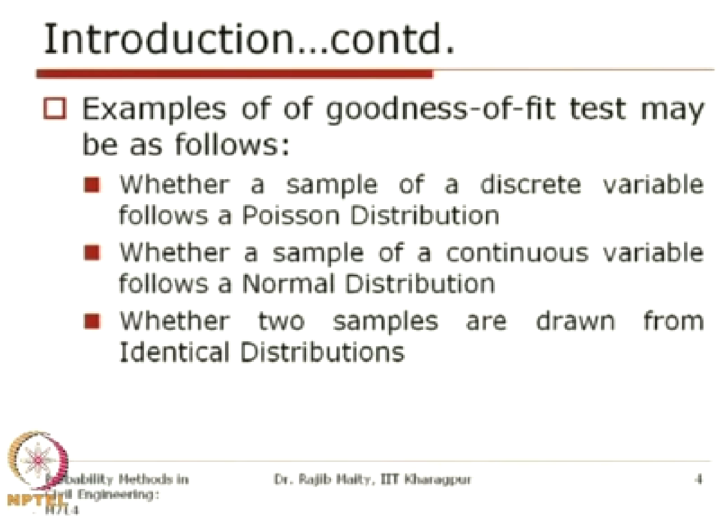


So now first we will go to this whatever I just mentioned so far is that it may be a necessary to check how well a set of observed data fit to a particular probability distribution and this one this can be described using the certain statistical test known as the goodness-of-fit test and as also, I just try to indicate that the basic philosophy is to measure the discrepancy between the observed and theoretical values used in the statistical model. So, this basically are the general thing for all the test that we going to describe now.

That and this one what we are describing is particularly for the one sample test suppose, when we are saying that I have a dataset and that follows a particular distribution so what is our interest is that whether there is any discrepancy between that whatever the observed distributions that we can see from this data and whatever the distribution that we are assuming to follow whether it is normal or log normal or even in the discrete side Poisson or so whether those distribution that theoretical values and this whatever the observed from this data are matching or not.

So that discrepancy we have to see and that discrepancy we have to decide which through some distribution we have to draw some inference in a statistical way.

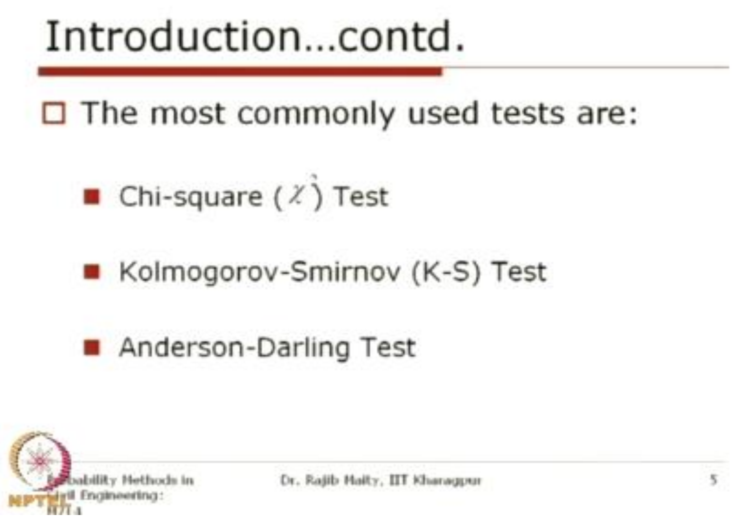
(Refer Slide Time: 05:16)



So, the example of this goodness - of - fit test may be as follows say thus that what type of question that we are looking for their answer say that whether a sample of a discrete variable follows a Poisson distribution or not. So these are just the example it is not that always I am looking for this Poisson distribution or so. So any distribution so I know that this random variable is a discrete variable so I have some hypothesize that whether that sample can follow a particular discrete distribution that we know already and those distribution we have discussed in earlier lectures, in earlier modules.

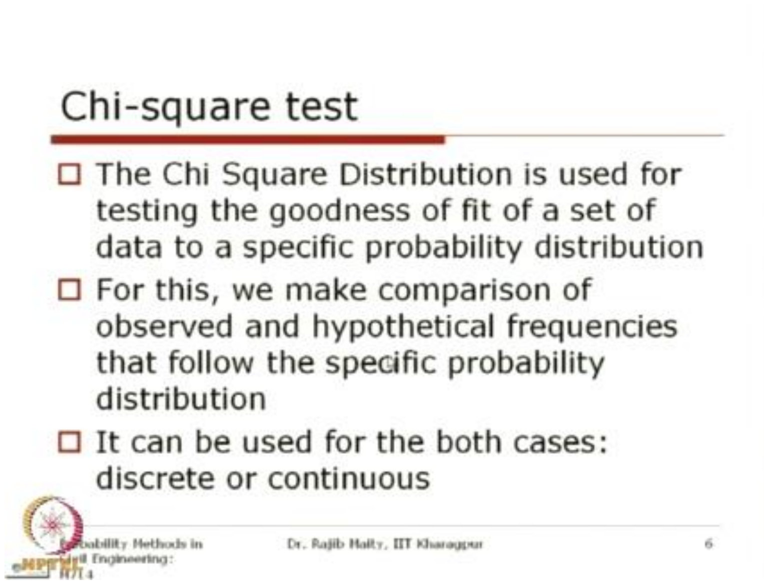
So, like that question I have a sample whether that sample follow a Poisson distribution or not. Similarly, whether a sample of a continuous variable follows a normal distribution or not, or gamma distribution or not, or lognormal distribution or not like that or incase of the two samples whether both the samples are drawn from the identical distribution or not so, when we are testing it, we will just take try to take that all this several types of this problems where we will try to cover almost all these different cases discrete variable case continuous variable case as well as the two samples case taking from two different sample we will take and whether we will test that whether they are from the identical distribution or not.

(Refer Slide Time: 06:53)



So, as we mentioned, there are some of these commonly used test are there these are standard test that we generally use for this goodness-of-fit test. The first one is this, chi square test and then, the Kolmogorov Smirnov test and there also other test which is a Anderson-darling test which is generally known to be a little improvement over this KS test. And we will take this one also, but may be for this lecture we will just consider this chi square test and KS test.

(Refer Slide Time: 07:33)



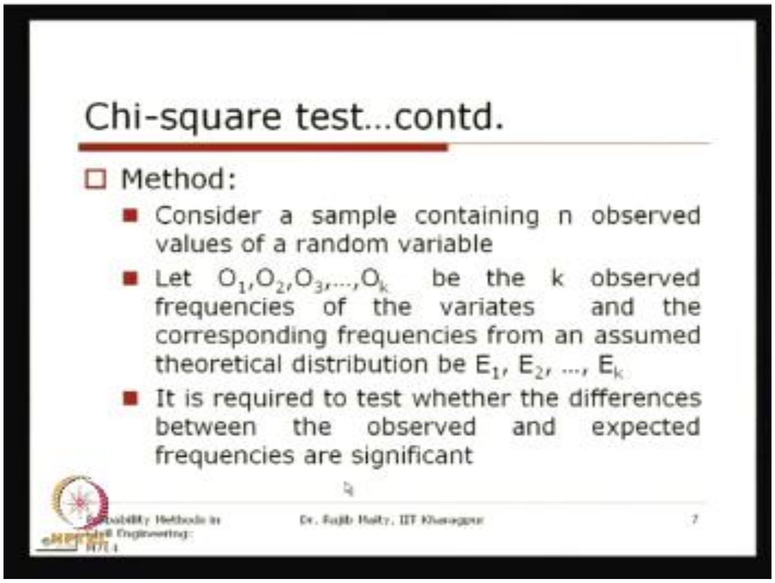
So now, this chi square test, the chi square distribution is used for testing the good ness of fit of a set of data to a specific probability distribution.

For this we make the comparison of the observed and hypothetical frequencies that follow the specific probability distribution so this is that what we have discussed for this basic philosophy and this is basically, if you see that overall approach is same for all this test so the comparison of this whatever you have observed from this data and whatever we are suppose to get from that hypothesize distribution whether they are matching or not so here basically this chi square test is based on their frequencies so we will find out the observed frequency and also see what is the hypothetical frequency.

And obviously as we are talking about these frequencies so we have to categorize the data into different beans and each bean what is the frequency, what is the observed frequency this observed frequency means whatever the data that we are having based on that what is the frequency that we can see, and we are hypothesize one distribution. So based on that distribution obviously, that distribution should have some parameters with that parameters what are the frequency that we can observed.

And this one can be used for the both the cases means, both for this discrete random variable as well as for the continuous random variable we can use this test.

(Refer Slide Time: 09:04)

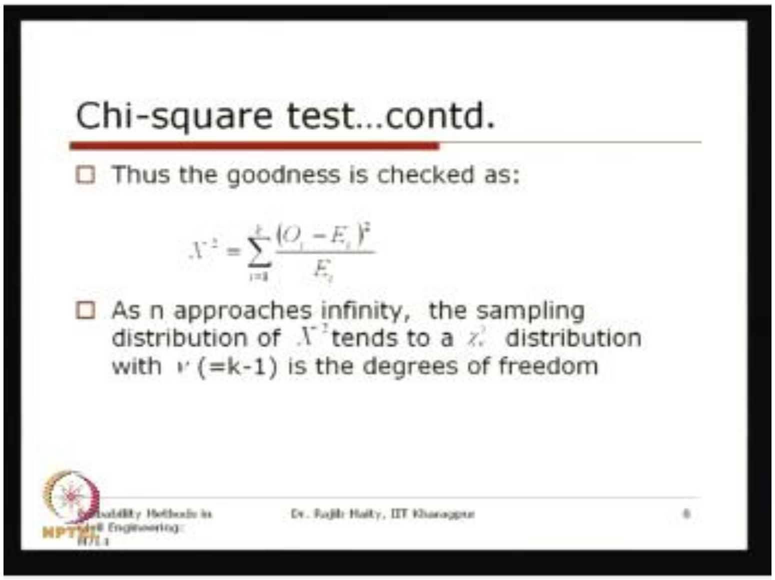


So what is done that may be we will discuss now in a step in different steps, first of all let us consider that a sample contains in observed values of the random variable, so whatever the data that we are having that is that n numbers of data that we are having. And now this notation that is O1,O2,O3……Ok be the K observed frequencies of the variates and the corresponding frequencies from the assumed theoretical distribution be E 1,E2,E3……EK.

So this is that observed that is this is from whatever that data that we are having is the O 1, O2,O 3,OK and these E1,E2,E3,EK, these are from that whatever the distribution that we are hypothesize that this data set may follow that particular distribution. So from that distribution whatever the frequency that we are getting is E's Ei's and whatever from the data that is Oi's.
So this it is required to test whether the difference between the observed and the expected frequencies are significant or not.

So this are the O1, O2, O3, Ok and these are from the hypothesize distribution, now they are difference whether this difference from the observed and to the theoretical is significant then we can say that whatever the data we are having is not following whatever the distribution that we have and on the other hand if they are almost same then we can say that yes that it is following that particular distribution.

(Refer Slide Time: 10:53)



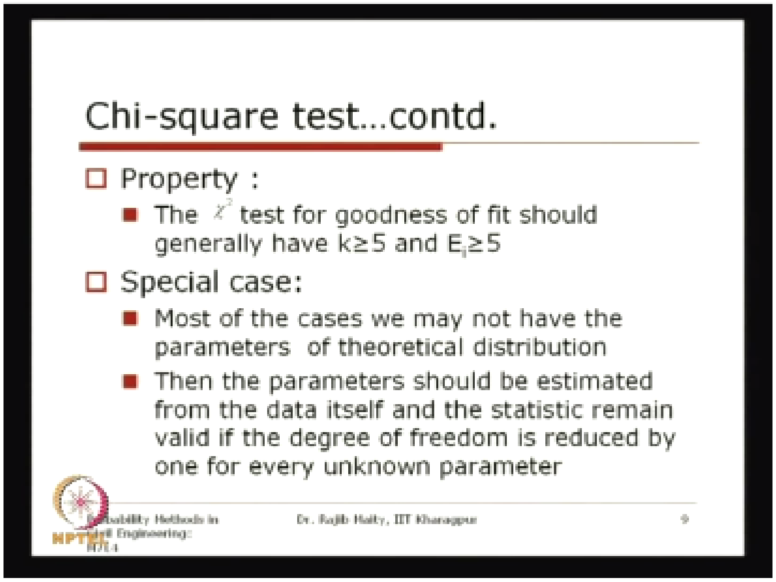
Now so how to do that one how to find out that discrepancy is through this parameter, we generally call it as a statistics and that statistics is obtained as that $(Oi – Ei)^2$ / Ei and sum of are all K' s means all this groups that is from 1 to K. Now this quantity this quantity is as I told that this is the statistics and denoted by here X^2 that we have denoted this one and it has been found that this X^2 or this statistics.

It follows a chi- square distribution, if the A tends to infinity that is that obviously the mathematical limit, for the large number of this sample data that is available this statistics follow a chi-square distribution and this chi-square distribution is having some degrees of freedom, here the degrees of freedom is k - 1.So this one show as you know that this particular statistics follows the chi-square distribution with k-1 degrees of freedom.

That means we know that what are its properties and based on that at what significance level I want to test whether this is means significantly this is significant or not that we have to test. One thing is I should mention here that sometimes it is required for example that when we are calculating this Ei it may require to estimate some of the parameter of that hypothesize distribution from the data itself.

If that parameter is not known and in that case as many parameters as we are required to estimate from this data then those many degrees of freedom will be lost. So if you are not estimating any parameter from that data available that time this the degree freedom is k-1 but if we estimate one parameter then this degrees of freedom will be k-2 if you are estimating 2 parameters then degrees of freedom will be k-2 sorry k - 3. So similarly so as many parameters you are estimating from the data those many more degrees of freedom will be lost.

(Refer Slide Time: 13:17)



So now this one, okay some more things some more properties these are not that means directly I cannot say that why these are required but for the better results that we report it for the better result this we should take care before we go ahead with the test. The first thing is that the K should be greater than equal to 5, so that number of a groups that we have divided that observed data it should be greater than equal to 5.

And that Ei is that is the frequencies that we got in each beam should be at least 5, so this Ei should be greater than equal to 5.Now the special case the most of the cases we may not have the parameters of the theoretical distribution and then the parameters should be estimated from the data itself and the statistics remain valid if the degree of freedom is reduced by one for every unknown parameter.

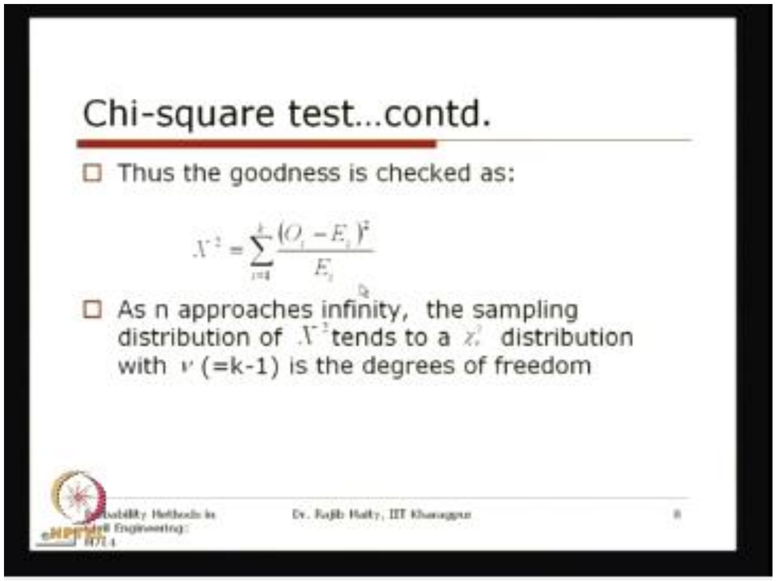
So that is what I just explained, so if you need to estimate the parameter of the hypothesize distribution from the data then you need to then you need to allow that much degrees of freedom will be lost. Now assuming that the distribution follows as this one that I told there will be square here as you have seen earlier this statistics this observed minus this hypothesize square divided by Ei and their summation.

So if this one is less than this value what is shown is C(1αμ) and this μ as it is already explained that this is the degrees of freedom, this C(1 – αμ) that is C(1-μα) is a value of approximate that this chi-square distribution with μ degrees of freedom at the cumulative probability 1-α so now this α is that significance level that we are talking about, the assumed theoretical distribution is an acceptable model at the significant level of this alpha.

So this is what that I was mention initially that all these test should be associated with some significant level, generally this significance level are kept around say 0.01 or sorry 0.01or 0.05. So at this cumulative probability 1 –α, so if it is that 5 percent significance level if we say that means that cumulative probability at which we are testing it is at the 95 %. So if the observed statistics is less than is less than that cumulative probability of that distribution.

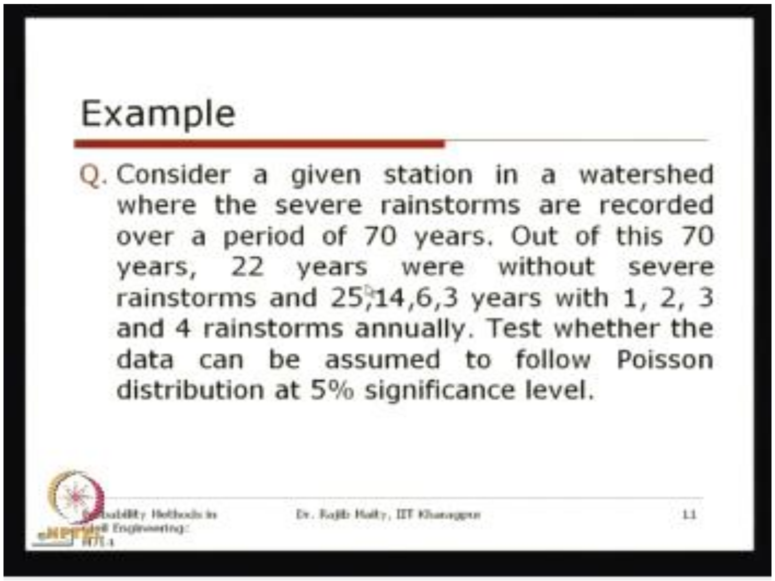
Here it is the chi-square distribution of this 95% if the alpha is 0.05 significance level and then we can say that yes that whatever we have hypothesized may not be rejected. So yes as I told that this should be the square.

(Refer Slide Time: 16:36)



As we have seen it here also this statistics this y minus Ei power square.

(Refer Slide Time: 16:41)

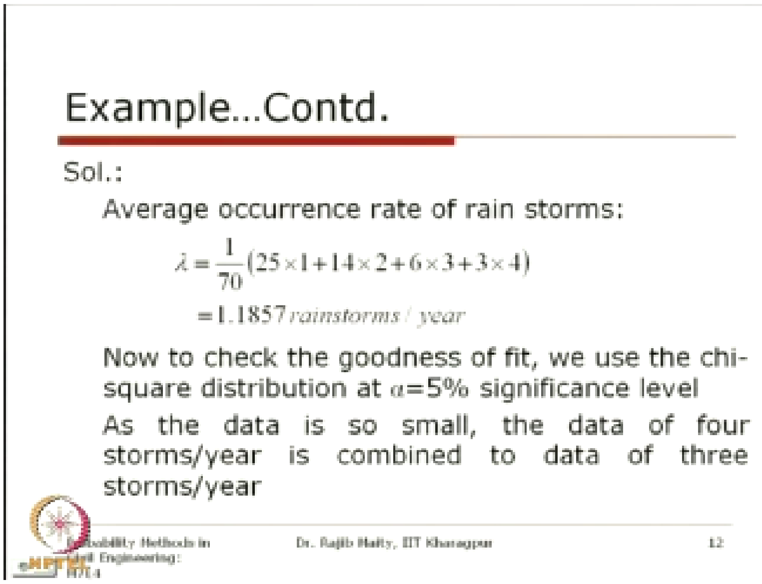


Now we will take up one example and this example we have taken for the discrete case and because this chi-square distribution can be used for the is for the discrete parameters. So we have it can be used for the continuous as well as discrete but the other than next one what we are going to cover is the K S test that is for the continuous distribution, so here we are taking a discrete example.

Consider a given station in a watershed, where the severe rainstorms are recorded over a period of 70 years, last70 years we have recorded the how many rainstorms are there in a particular year and out of these 70 years, 22 years were without severe rainstorms, so there are in 22 years there are no severe rainstorms, so number of rainstorms is 0 and 25 years,14 years, 6 years and 3 years are with 1 rainstorm, 2rainstorm, 3 rainstorm and 4 rainstorms respectively.

So 25years we have observed there is 1 rainstorm, 14years we have seen there are 3 rainstorms, 16, 3 rainstorm, and 3 years there are 4 rainstorms are there. Test whether the data can be assumed to follow a poisson distribution at 5% significance level. So here as you can see that we are hypothesizing that whether the data that is given that whatever we have recorded whether it is following the Poisson distribution or not and the significance level is given as 5%.

(Refer Slide Time: 18:19)



So here now you see so the Poisson distribution you know there is some λ there is one parameter that we have discussed earlier modulus that parameter that lambda is the mean rate of occurrence so rate so average occurrence rate of this rainstorm so if you want to calculate then we have seen that there are 22 years where there is no rainstorms so 22 multiplied by 0 basically the first thing then 25 years1 rainstorm 14 years2 rainstorm 6 years 3 rainstorm and 3 years 4 rainstorms are there are total so in this way out of these70 years what are how many total rainstorms that we have seen that divided by70 gives you that average rainstorm that can occur in a year.

So 1.1857 rainstorms you can see in a year so λ here is that 1.1857 and the t that we are - λ t - that is the total quantity that t is here the per year so one year how much rainstorms has occurred so this is the quantity 1.1857 now remember that we have estimated this one from the data itself so this was not supplied sometimes what happens these data could have been supplied directly whether the test whether the data is following the Poisson distribution with the λ is equals to say1.3 or 1.1 like that if that is given to us that means we are not estimating that one from the data.

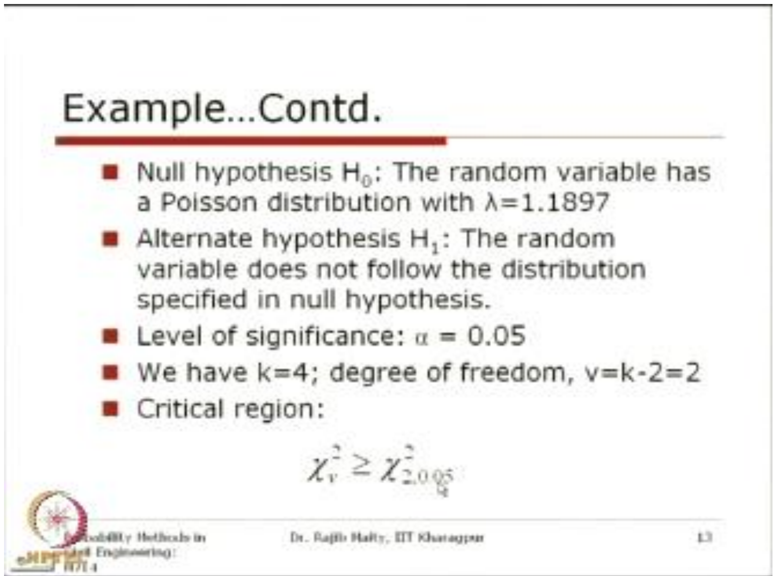
So they are the degrees of freedom whatever I told that it is k - 1 will be there but here we have already estimated one parameter so now the degrees of freedom will be k – 2 so now to the check the goodness- of- fit we use the chi - square distribution at α = 5% significance level as the data is so small the data for the4 storms per year is combined with the 3 storms per year so this one is generally done but so there is some discussion is required we can now we have to think that what distribution you have to hypothesize so it is the Poisson distribution that you have hypothesize and what is the support for that distribution now the Poisson distribution.

That support that we are looking for is basically starting from the 0 to 0 1 2 the discrete values and goes up to ∞ now the data that is obviously is given to us is that 1 rainstorm 2 rainstorm 3 rainstorm up to the 4 rainstorm per year but when we are hypothesizing that this is the this is the Poisson distribution of obviously I cannot cuttle at any particular point so here what is the general practice is that for the higher side where data is becoming very small we can combine those things to take care two things one is that so I can declare that yes more than equal to this value is having this frequency.

And the second thing is that we will also be testing whether that each group is having that minimum requirement of this frequency is available or not because this kind of asymptotic distribution that is it is going towards + ∞ so this type of it should be open bounded but the way the problem is given it is just as a close boundary at 4 rainstorm per year so here if we just consider that 2rainstorms are combined together that is the 3rainstormper year and 4rainstorm per year case then we can say that will greater than equal to 3 rainstorm per year.

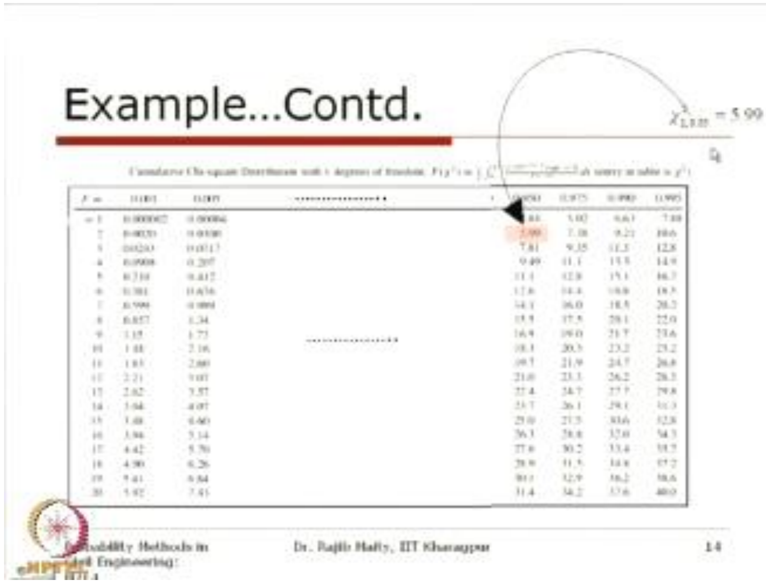
So that remains that positive side remains unbounded and also this will help to check whether that at least greater than 5 that is for the better result that we have mention earlier so here also what we are doing is that the 4storms per year and the 3 rainstorms per year are combined together.

(Refer Slide Time: 22: 25)



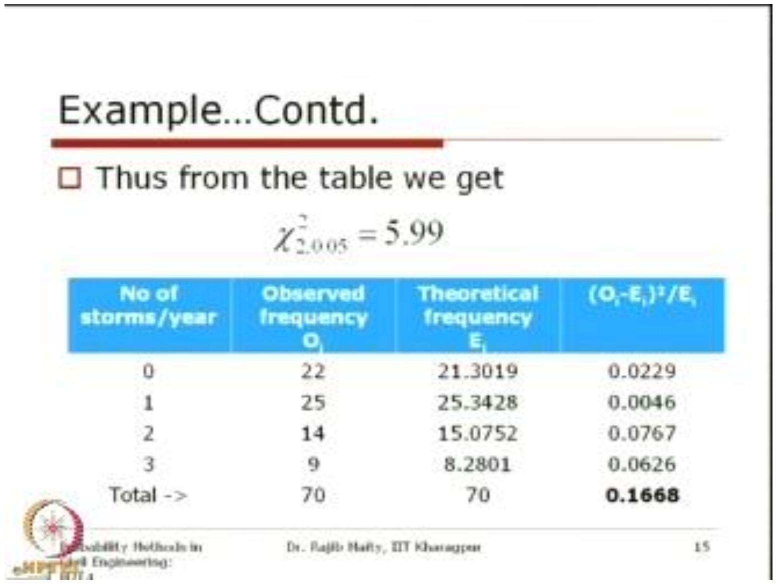
So here the null hypothesis is the random variable has a Poisson distribution with λ equals to 1.1897 alternate hypothesis is the random variable does not follow the distribution specified in null hypothesis what is specified here significance level is 0.05 at 5% significance level mentioned and we have the k = 4 so the degrees of freedom here are the k - 2 so k - 1 is from there and one parameter we have estimated so it is k – 2 the degree of the freedom is 2 so that critical region here is this chi square distribution that is chi square distribution with 2 degrees of freedom at α equals to 0.05.

(Refer Slide Time: 23: 11)



Now if you see so this is that chi square distribution table basically you know that we have discussed this distribution earlier and this standard values are listed in any standard text book so here if you see that this point at this .95 these that cumulative probability are given here and for these degrees of freedom 2 this value is .5.99 so these value is our critical value that we have to test against.

(Refer Slide Time: 23: 44)

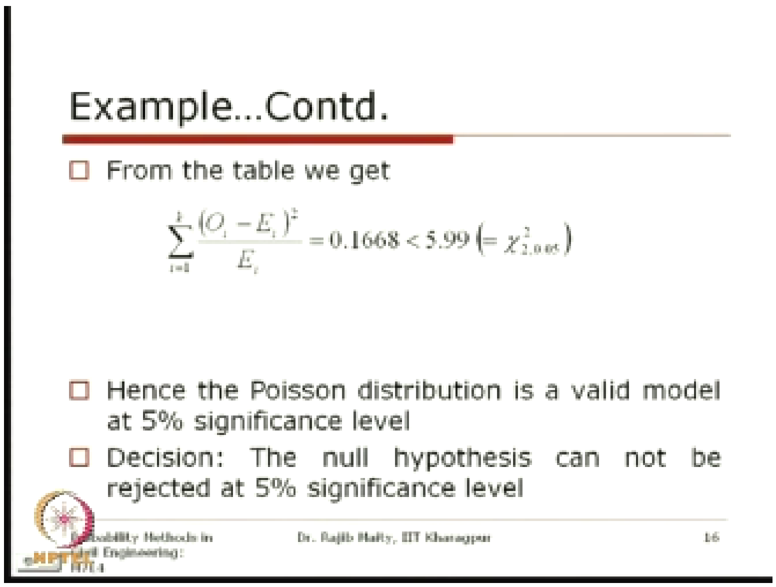


So this is that 5.99 that we got from this table and these are the beans here now so 0rainstormper year 1 rainstorm per year 2 rainstorm per year and this is the greater than equalto3 rainstorm per year so now as it is given in this data that there are 22 years where this0 rainstorms are there and the 25 year are there is that 1 rainstorm 14years 2 rainstorm and 6 years 3 rainstorm and that 3 years4 rainstorms are there so we have combined it here to get the 9 years data where it is greater than equal to4 rainstorm per year.

So we have avoided that one in one occurrence it is becoming 3 which is less than 5 so that is also avoided and that right side is kept open and this theoretical frequencies that is these values what we are getting is that this from this Poisson distribution that λ t here equals to this λ equals to that1.1897 and t = 1 so1 year so this λ t value we will get and from this Poisson distribution if you just put that x = 0 x = 1 x = 2 and x = 3 with that value of lambda t equals to that value then we will get this values which are the so are the theoretical frequencies for this Poisson distribution.

So now so these are so these values we are getting when we are putting this x equals to 0 and this value you are putting when you are putting x equals to 1 from this Poisson distribution so this is the theoretical frequencies and these are the observed frequencies now what we have to do we have to get this statistics that is y - Ei whole square divided by Ei so if we do this one then we get these values and we have to sum it up this one so this is the summation from is that0.1668.

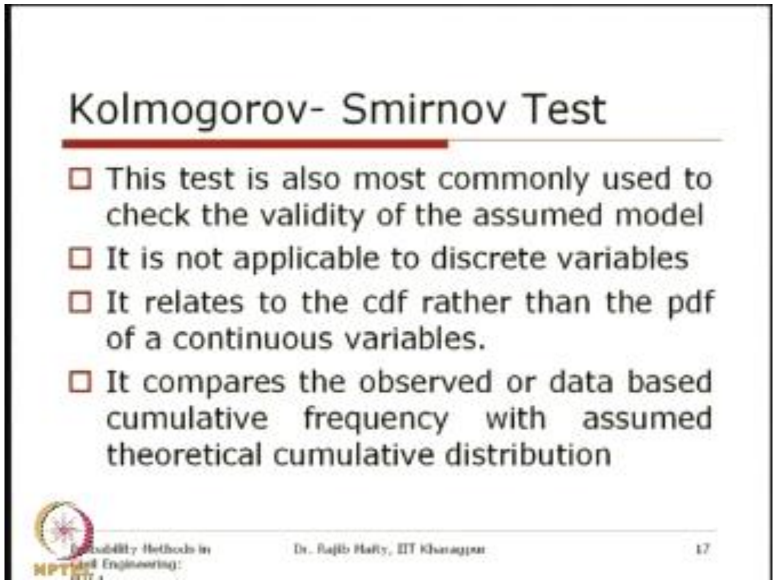
(Refer Slide Time: 25:58)



So these one this is our statistics which is equals to 0.1668 that we have seen from the table so that so which is now is the less than 5.99 that we have seen from the table so this is less than this critical value so hence the Poisson distribution is a valid model at a 5% significance level so the decision is that for this test is that the null hypothesis cannot be rejected at 5% significance level so there could be some other words that we can express okay the null hypothesis is not rejected and all so what I feel is that this should be the proper decision proper probabilistic or the proper statistical inference should be written.

As this one that null hypothesis is accepted is not the right thing to declare so rather the complete thing that we can declare is that the null hypothesis cannot be rejected at 5% significance level because the result that we have got it depends on what significance level that we have adopted for so that is what we have to mention that at this significance level the null hypothesis cannot be rejected.

(Refer Slide Time: 27:17)



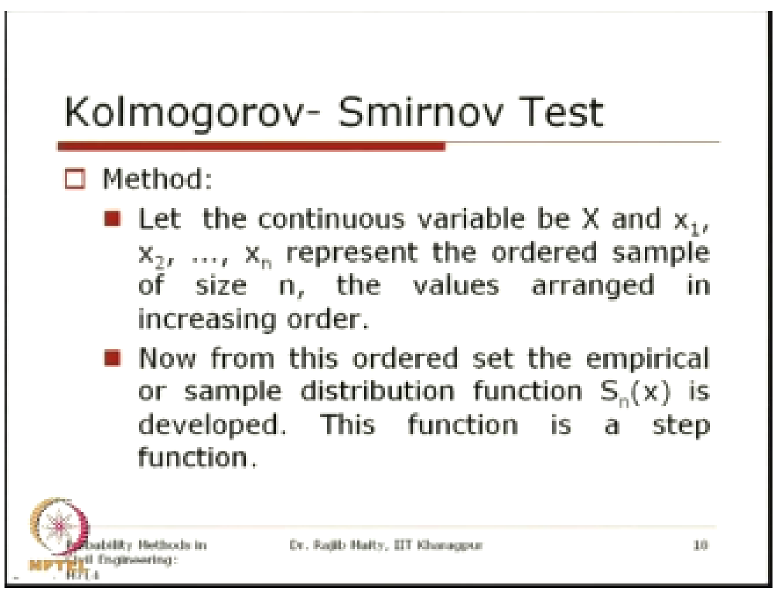
Okay so next we will go to that our second test which is the kolmogorov-Smirnov test in brief we generally mention here K S test this test is also most commonly used to check the validity of the assumed model and it is not applicable for the discrete variables so mostly that continuous random variable if the data set is available with that we generally get this one it relates to the cdf rather than the pdf to a continuous variable so here earlier what we have seen that the frequency that when we are talking about in this chi square test that is basically a pdf that we are talking about here this K S test is generally based on this what is there cdf, that is cumulative distribution function, and it compares the observed or data based cumulative frequency with the assumed theoretical cumulative distribution. So data based cumulative distribution, sorry this should be distribution with the assumed theoretical cumulative distribution.

So there are some kinds of discrepancies or short comings I should say, in this chi square test is that, we need to define that we need to first of all get that beans that we have to first categorize the dataset, the full dataset that is, okay this is my range and in the discrete it is okay, fine that whatever the example that we have seen. Now for the as I told this chi square distribution is also applicable for the continuous distribution.

Now, if we take some continuous distribution in case of this chi square test, what we have to do, we have to first categorize the data into different beans and each bean we have to see the frequency and also we have to check that whether each bean is having so, the number of beans should not be less than5, as well as each bean should have minimum that minimum frequency should be 5, for the better results that we are mention.

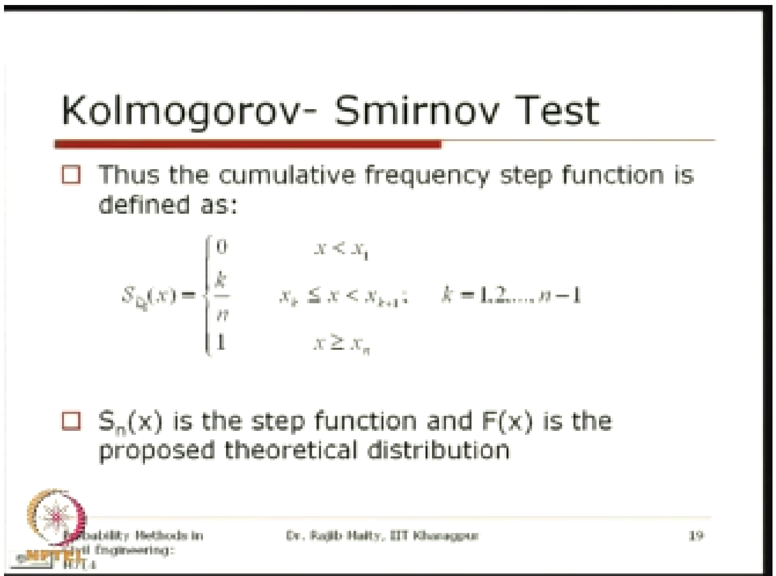
So, these two things are not there in this K S test, so it is directly getting the, it is directly access the, it is results from the cdf itself directly, so there is no need to categorize the data into beans. And so, it is avoiding those requirements of these minimum 5 beans and that each bean should have that minimum frequency of 5,so which is not there in this K S test.

(Refer Slide Time: 29:48)



Now, so if suppose that there is a random variable X and we have some dataset of this x_1,x_2,x_3 up to x_n, so represent the ordered sample of size n. So, whatever the data that we are having from this actual observation, we can first of all we can make it in an arranged in an increasing order, and that increasing order if that increasing order is that x_1, x_2, x_n if that is available. Now, from this ordered set the empirical or sample distribution function $S_n(x)$ is developed and this function is basically a step function. So, this is from this data so how from this sample how to get that this cumulative distribution function is as follows.

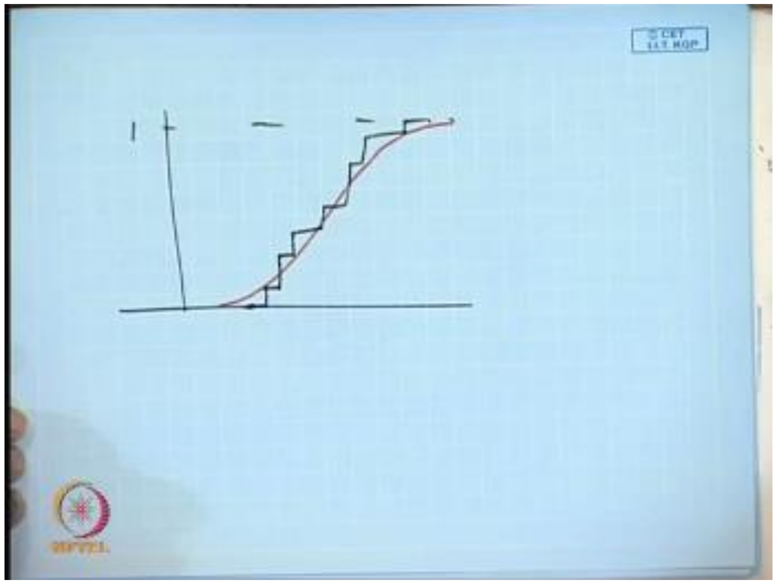
(Refer Slide Time: 30:37)



That is for, when this $S_n(x)=0$, in case, when x is less than x_1, $S_n(x)=k/n$ when this x is in between k to k+1 and k can vary from 1,2,3 up to n-1. So, basically for all this $x_1$to x_n for this thing we are defining these value of this cumulative distribution is k/n. And for x>n = 1. So, this is basically what we are getting is the whatever the representative, cumulative distribution from directly from the data and each point it is changing, this value is changing, so it will look like a step function.

So, obviously as you can see this essence should will start from a value 0, it will start from a value from 0 and at some point it will go.

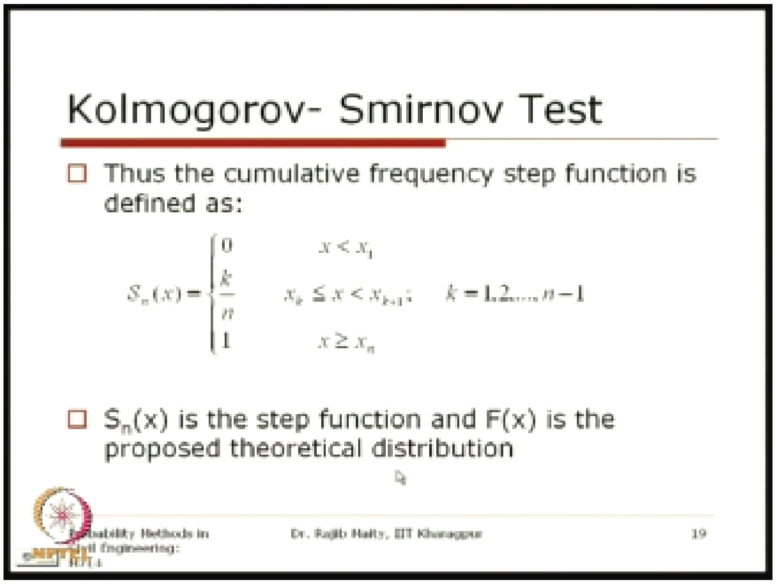
(Refer Slide Time: 31:40)



It will increase and again from the next one, it should go and like this, it will go somewhere it will go flat, and in this way what will happen it will ultimately attain the value where it is 1.Now, this is the distribution basically we got it from this data. Now, the hypothesized distribution, suppose that I want to match that normal distribution, so that normal distribution with the parameters of course whether it is supplied or obtained from this sample data is that with that data, I can also plot what should be the shape of this that particular distribution, say that normal distribution if I say.

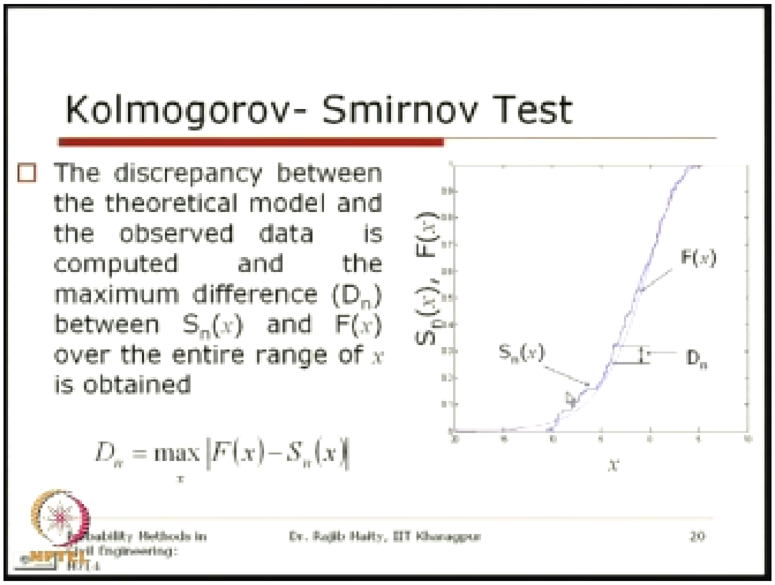
Now if this, whatever we have observed the black line here, if this is very close to this red one which is the theoretical distribution obviously, then the data is from that particular distribution. So, this discrepancy, now again, keeping the same approach same for all that test that I mention at the beginning of this class, is that here also we have seen that what is the maximum difference between this two distribution that cdf, so one is the theoretical and the other one is the database. So, that discrepancy we have to assess through some statistical test, this is what.

(Refer Slide Time: 33:03)



That is why we have got this S_n (x) so, this S_n (x) is the step function and this F(x) is the proposed theoretical distribution that what I have drawn in the reading just now.

(Refer Slide Time: 33:14)

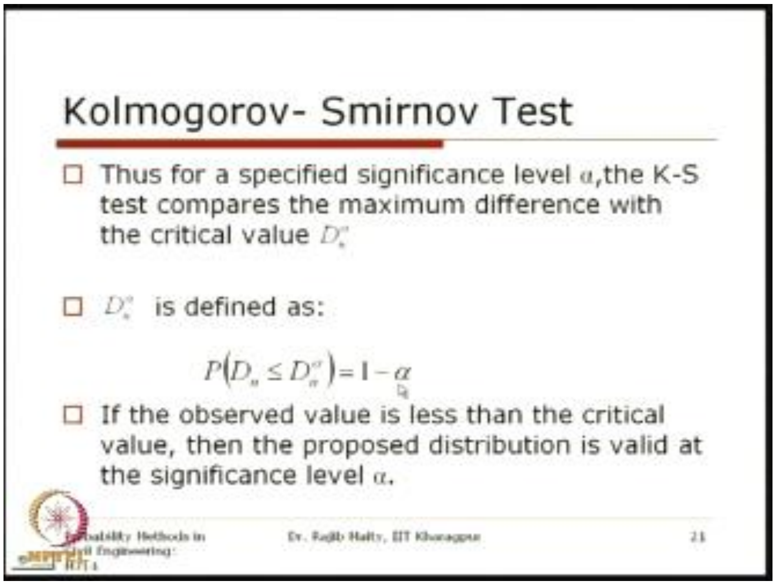


So now, here if we see the discrepancies between the theoretical model and the observed data is computed, and the maximum difference the D_n between the $S_n(x)$ and the F(x) is over the entire range of x is obtained, which is denoted as D_n, which is the maximum for all x,$F(x)-S_n(x)$ obvious the absolute value. So, now you can see here, for a particular point here as you can see so from the blue one that I have drawn, this is also a step function here the blue line, and this pink one is that your, is that theoretical distribution.

Now, the difference between any point to that theoretical distribution is your that value is the difference between two things. Now, what we have to pickup from this two information is that what is the maximum difference so, at each point there will be some difference between this blue line and this pink line, so that difference at each and every point I have to find out and we have to select, we have to pick up the maximum one, so what is the maximum difference.

This blue line obviously, as we are coming left towards this, this is 0, and from where it is ending towards the right to that one it is equal to 1. And so over this entire range, we have to pick up what is the maximum difference.

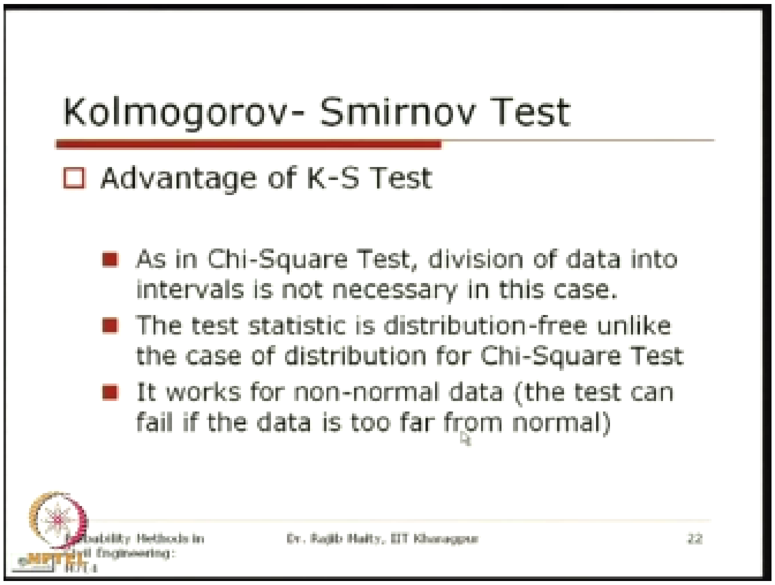
(Refer Slide Time: 34:40)



So, thus for a specified significance level at α that K-S test compares the maximum difference with the critical value D_n^α. Now, what is this D_n^α this D_n^α is defined as the probability that D_n less than equals to D_n^α=1-α, again this α here is that significance level that we have mentioned at the beginning of this class. So, if the observed value is less than the critical value, then the proposed distribution is valid at the significance level α. So, we have to check that whether whatever the maximum difference that we get, and whatever the critical value.

So, this probability, that is that observed D_n less than equals to that critical value whether, it is equal to this1-α or not.

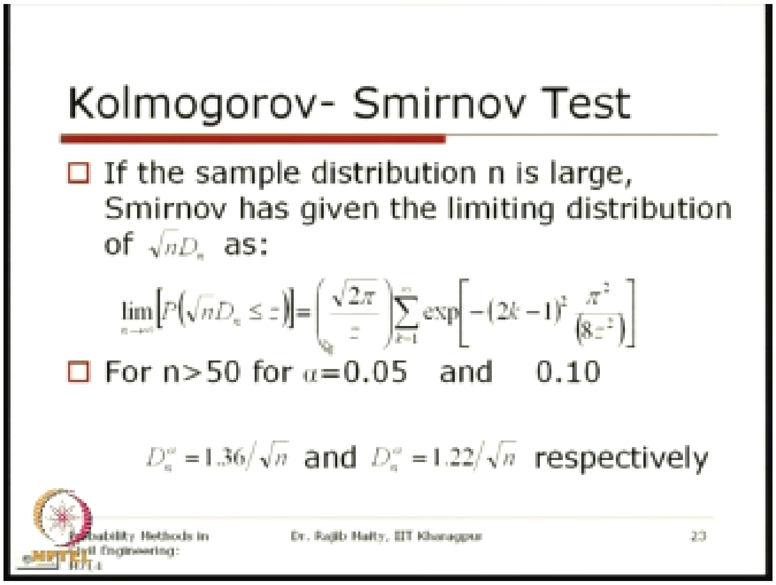
(Refer Slide Time: 35:39)



Now, the advantage of this K-S test as in the chi square test, division of the data into interval is not necessary in this case, so I think this things I was just mentioning while at this starting of this K-S test. So, we so these intervals are not necessary here, because we are just observing each and every data point. The test statistics is distribution free unlike in the case of the distribution of the chi square test. It works for this non-normal data, however the test can fail if the data is too far from this normality.

So, there is no such restriction that with the data should be approximate normal or so, but it is better to get the better result again that data should be somewhere near normality that is why the popularity of this K-S test is we have seen, it is very frequently, it is applied to test whether the data follow the normal distribution or not that is where the maximum application of this K-S test has been found.

(Refer Slide Time: 36:42)



Now, the sample distribution, if the sample distribution that n is large, not the distribution sample size, if the sample size n is large, Smirnov has given the limiting distribution of $\sqrt{n}$ multiplied by this D_n ,so this is that D_n that we have defined the maximum difference that multiplied by $\sqrt{n}$, these quantity follow a distribution like this that limit n tends to ∞, probability of this quantity $\sqrt{n}$ multiplied by that D_n the statistic less than equals to z is equals to $\sqrt{2\pi}/z$ multiplied by $\sum(k-1)$ to ∞ $\exp[-(2k-1)^2(\pi^2/8z^2)]$.

So, this is what it is giving is that for this n tends to ∞ means, when this n is very large that time what we can get is that, how this D_n is varying is that through this, that we have to find out suppose that, suppose now so it depends on this what is the significance level that we have fixed, suppose that this significance level if it is 0.05 that means, this probability is equals to $1-\alpha$ that means this 0.95.Now, if we solve this right hand part with equal to that, what is that,0.95,then we will get that is z becomes 1.36 if you see this quantity here inside this exponential term, that is,$(2k -1)^2 \times \pi^2 /8z^2$.

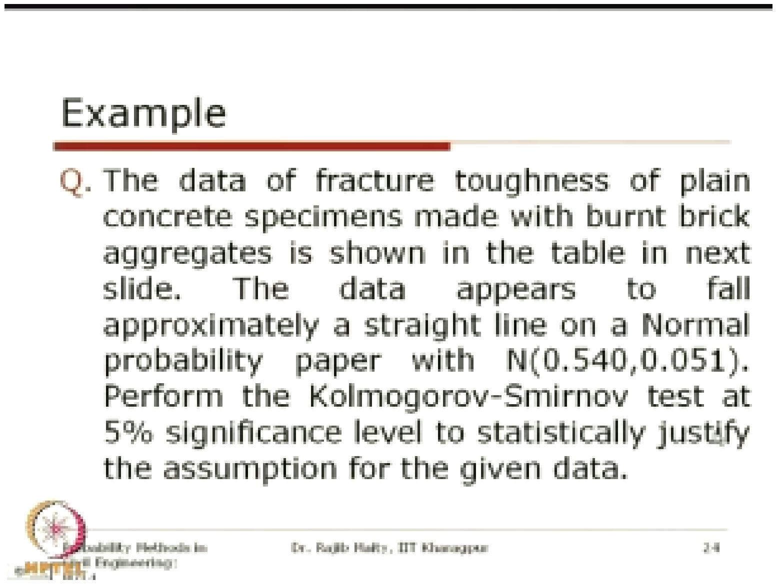
So, this term basically is changing that is from this k = 1 to in infinity, now k if k is 1,you can see that this is just minus of this quantity plus exponential of, if k becomes2 then, it becomes 9, so -9,so exponential -9 times of this one and if k becomes 3then it is 25 times of this one, so basically if you just do just a hand calculation also you will see that, if you just consider the first term itself that is k =1 only and remaining if you just ignore it then, also you will see for this, this is almost very closely matching with this one may be, it is just varying after third decimal or so.

So, for this n >50, some times in some text book can refer to that if is n >35 itself, and for this α = 0.05 that means, this right hand side is equated with this 0.95 and if we just calculate what should be the value of this z, if we just consider only one value of k= 1 then, you will see that this z becomes very close to this 1.36.So,for this significance level α = 0.05 that critical value is 1.36 /√n, so that means this Z becomes 1.36 and this is that 1.√n.

So this critical value this should be so our observed statistics that is D_n should be less than this particular value to declare that, at that significance level, we can accept that particular hypothesis or that null hypothesis cannot be rejected. And similarly, if you put that say α = 0.1; that means, this right hand side if we equate to… if with that 0.9 then, we will see that this quantity, this z is becoming1.22,

So the critical value is 1.22 /√n and for these lower values of n, this is also available in the standard table from that distribution and we can refer to those tables.

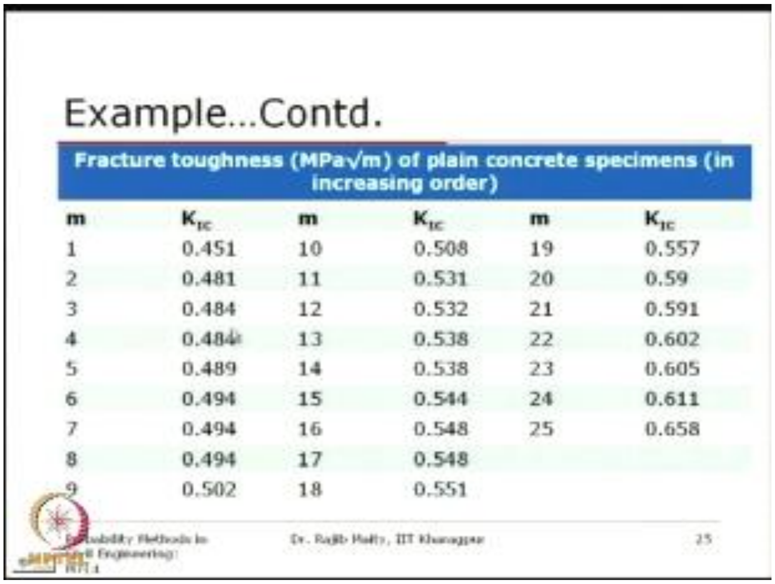
(Refer Slide Time: 40:41)



To get these critical values so, we will take up one example to just to discuss all these things, and here we have taken one example of this continuous random variable. And as I mentioned that this is mostly used when we are considering that when, whether the data set is following normal distribution or not. So, that data of the fracture toughness of the plain concrete specimen made with the burnt brick aggregate is shown in the table in the next slide.

That data appears to fall approximately a straight line on a normal probability paper, that if it falls approximate normal on a normal probability paper, there is a possibility of it may follow a normal distribution, and that the parameters are $\mu=0.54$ and the $\sigma = 0.051$.To perform the Kolmogorov - Smirnov test at5 % significance level to statistically justify the assumption of the assumption for the given data.

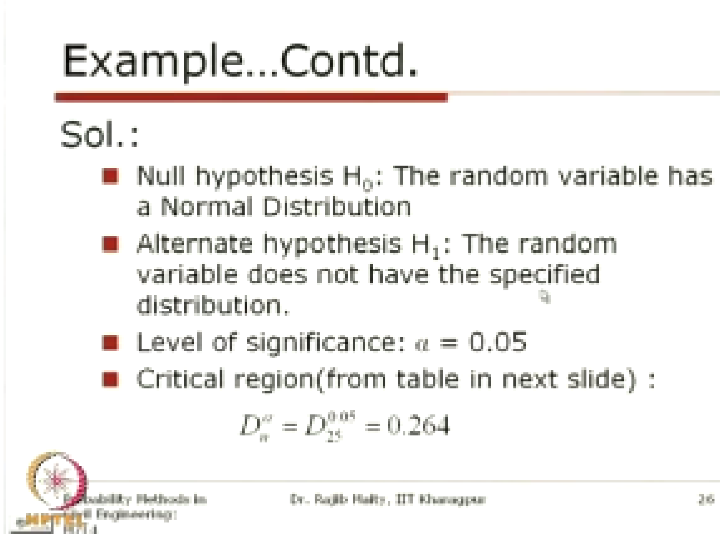
(Refer Slide Time: 41:48)



So data is supplied here that is the fractures toughness, which is having a unit of MPa scales square root of meter of the plain concrete specimen, and this is already arranged in an increasing order. So, you can see that, so that there are total 25 samples are there1to up to this 25and this one that KIC which is the notation for this fracture toughness is arranged in an increasing order, so from0.451 to 0.658.

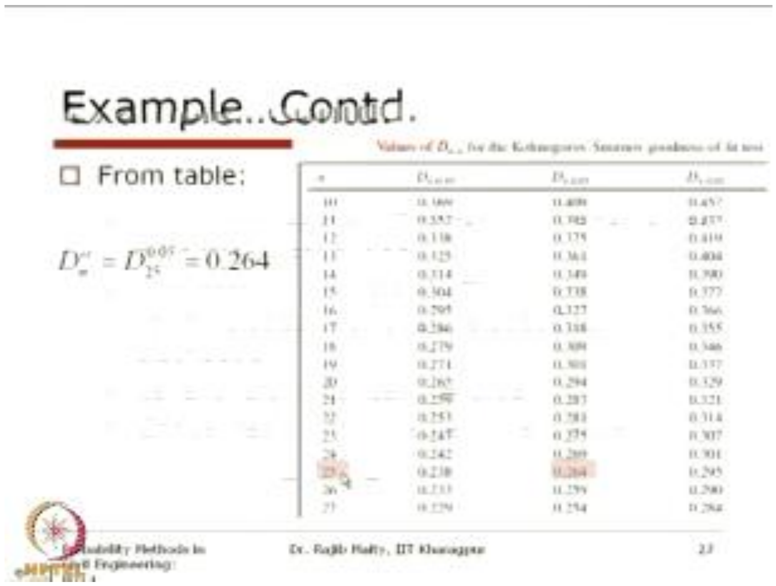
So we have to test that whether this dataset is following the normal distribution or not, and normal distribution having the parameters 0.54 and 0.051.

(Refer Slide Time: 42:31)



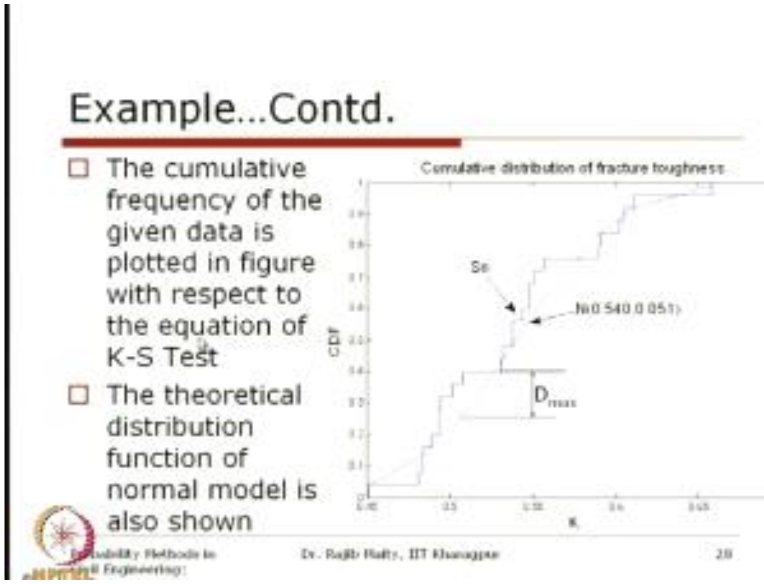
So, over null hypothesis here is the random variable, has a normal distribution with those parameters of course. Alternative hypothesis, the random variable does not have the specified distribution in this null hypothesis, level of significance is 0.05.And the critical region from the table just show, that is, here what the number of data is25 that is available and significance level is 0.05.

(Refer Slide Time: 42:55)



Now, if you see this table here, that is, these are the values of this K S test goodness–of-fittest. This is that for different n is listed in the first column 10, 11, 12 like this, and the second one is that, is you’re α 0.05. So, this kind of table is available to any standard text book and here, if you just see this value is highlighted for this n = 25 and the α=0.05, the value is 0.264,so from this table we have just picked up these the critical value of that test.

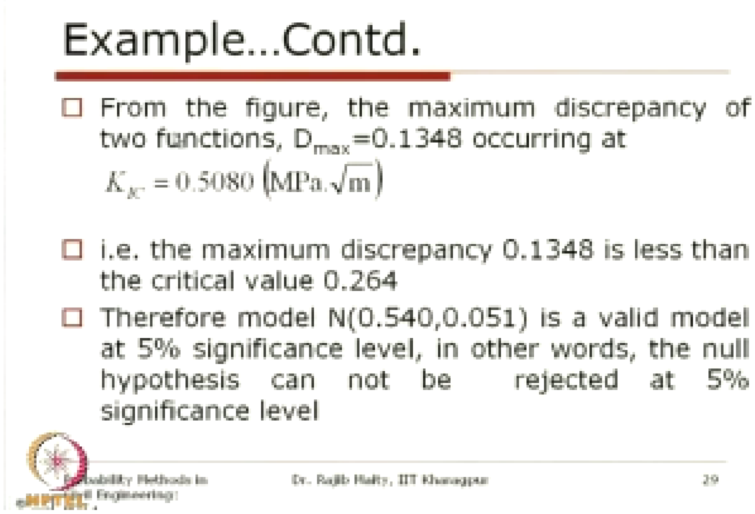
(Refer Slide Time: 43:31)



Now, if we just do this one, do these same methods that we have explained; now whatever the data that we are having, we have plotted it for it is distribution with this step function as shown in the blue line here. And the theoretical distribution for this normal distribution that cdf of this normal distribution with parameter0.54 and 0.051 is shown in the line magenta, now whatever the maximum difference between these two that we have to pick up.

So, the cumulative frequency of the given data is plotted in this figure with respect to the equation of this K S test and the theoretical distribution function of the normal model is also shown, what is shown here.

(Refer Slide Time: 44:16)

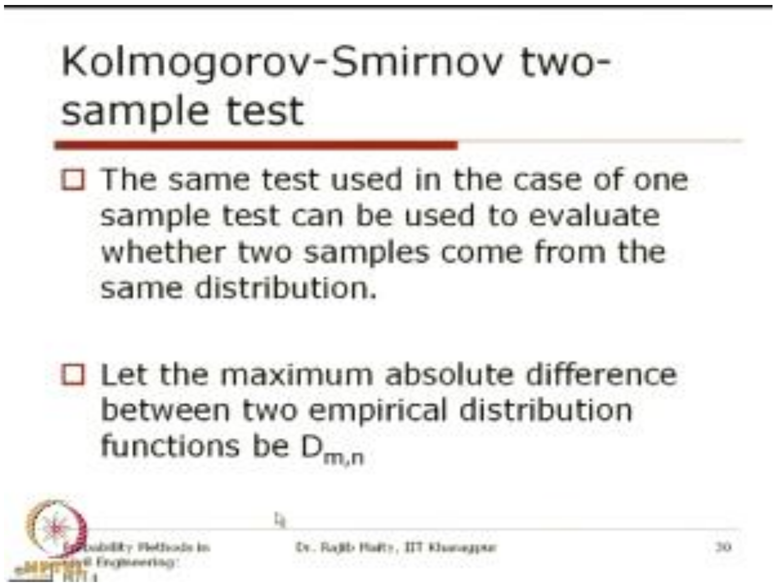


From the figure and of course, you can you can check it in this calculation also, the maximum discrepancy of the two functions is D_{max} = 0.1348which is occurring at KIC = 0.508, so at this value the D_{max} is0.1348.So, this is the only value that we can that we have to pick up from this comparison, this is what that K S test and sometimes from this point onwards, may be that further improvement, we will look for that, but that is the later part.

But here from this graph, only one information that we are picking up is, what is the maximum difference, just one particular value we have to pick up and that value is 0.1348.The maximum discrepancy is 0.1348, now we can test that it is less than that critical value that we have seen that also which is the critical valueis0.264 that we have seen from the table. So, what we can see here, so this model is a normal distribution with parameter 0.54 and 0.051is a valid model at 5 percent significance level.

In other words or I should say that this should be used to declare that the null hypothesis cannot be rejected at 5 % significance level.

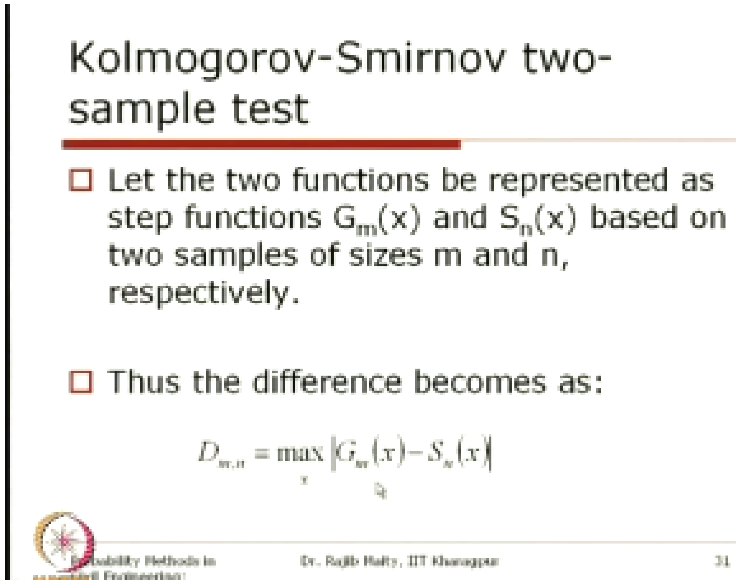
(Refer Slide Time: 45:40)



Well, now we will go to that two sample test, keeping the basic philosophy, again here is the same thing; one is that, so in the two samples; in one sample test what we have done is that one the data that we have obtained and other one is that sum of some standard distribution that we already know. So, in the example we have seen that one normal distribution that we have used. Now, the two sample test means that we are not using any standard distribution, the two samples are there, two samples both will have their own that event that observed cumulative distribution function and we have to pick up the difference between two.

So, in the one sample test basically, that is one observed data and other one is some standard known theoretical distribution. And in two sample test, both are the observed data, and we are plotting, and what values we have picking up are exactly the same thing. So, the same test used in the case of the one sample test can be used to evaluate whether the two samples come from the same distribution or not. Let the maximum absolute difference between two empirical distributions functions be Dmn.

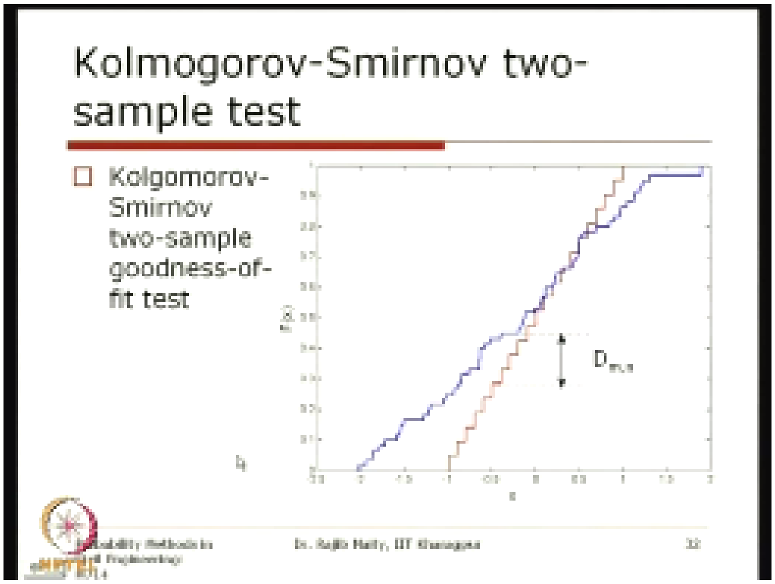
(Refer Slide Time: 46:57)



Now, let the two functions be represented as the step function G_m (X) and S_n (X) based on the two samples of size m and n respectively, so that two samples are there, one sample is having the m data other one is having n data, so this should be flexible. It is not that both the samples may not have the same size of this data. And what we have to go what we have to get is that from here, is that we have to find out what is the G_m (X) and what is the S_n (X), and following the same equation that we have shown in this one sample test using K S test.

So, thus here, the difference will be that maximum difference that we have to get, which is the absolute value of this difference between this G_m (X) and that S_n (X) and that one we have just pick up only single value again similar to the one sample test, which is denoted as that $D_{m\,n}$ this m is the size of this one sample, other one is the size of the other one.

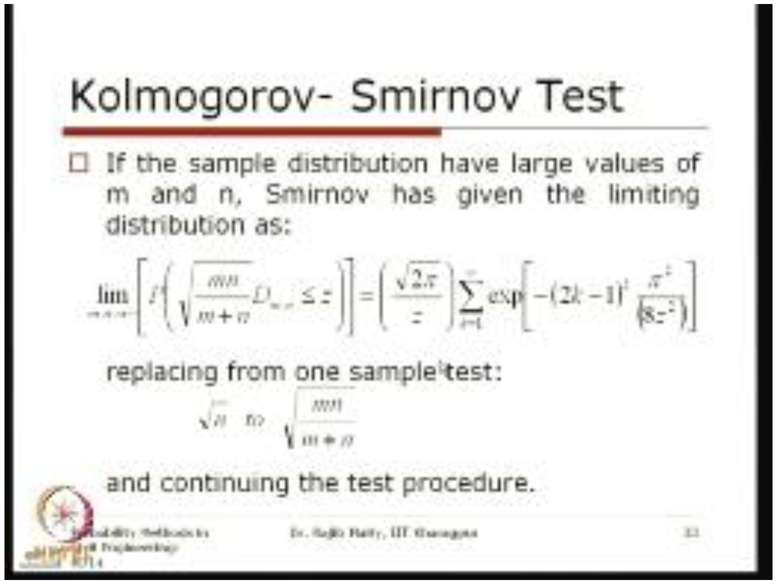
(Refer Slide Time: 48:00)



Now, this is one such typical example, how it will look like, the blue one is for the one sample and the red one is for another one. So, now we have to find out, so basically the red is again approaching towards this all are 0 value here, and here also red is going all are one- values are one. So, now each and every point we can find out what is the difference there, so at this point may be around point-0.4 or so, the difference is shown here between this point and this one, this is Dmn for this particular value.

So, for all such differences we have to pick up the maximum one, so this KS test for the two sample goodness of it as looks like this and

(Refer Slide Time: 48:52)

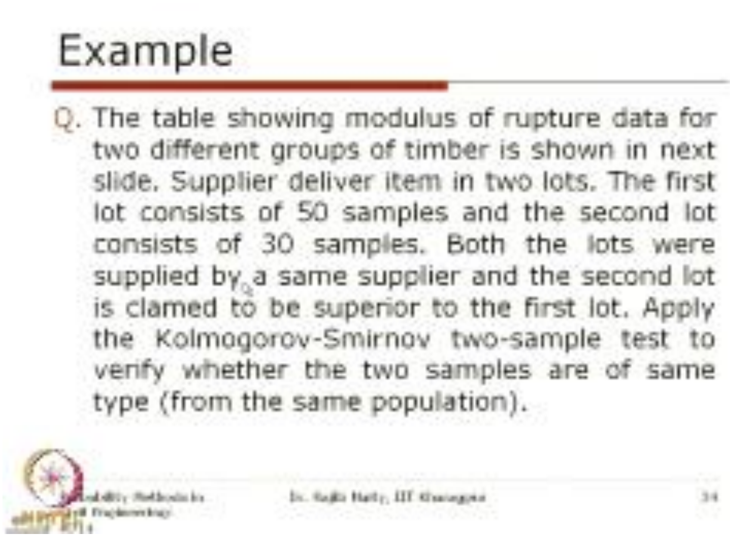


We have to pick up the maximum one. Now again, the sample distribution have the large values of the m and n, if this if this values are large enough that is m and n, the Smirnov has given the limiting distribution as this. The square root of m n / m + n, which is the data size for the two samples, multiplied / this D m n less than equals to z is equals to square root 2 π / z summation of k equals to 1 to ∞, exponential of - 2k- 1 square multiplied / π square /8z square. So, basically what happens from the single sample, it was here it was square root n and for the two samples test, it is replaced / this square root of m n / m + n.

So, basically the sample size in the one sample test we have used it for this n and for the representative sample size for the two sample test is this quantity m n / m + n. So, if we just change this one, so if we once, we are having this two samples, so what is the representative data length we have to calculate first. And the remaining thing is same, whatever we have discussed for the one sample test, that is, if we take that the significance level is 0.05 say, then this quantity be equate with this 0.95 then, it will again come that same value which is 1.36.

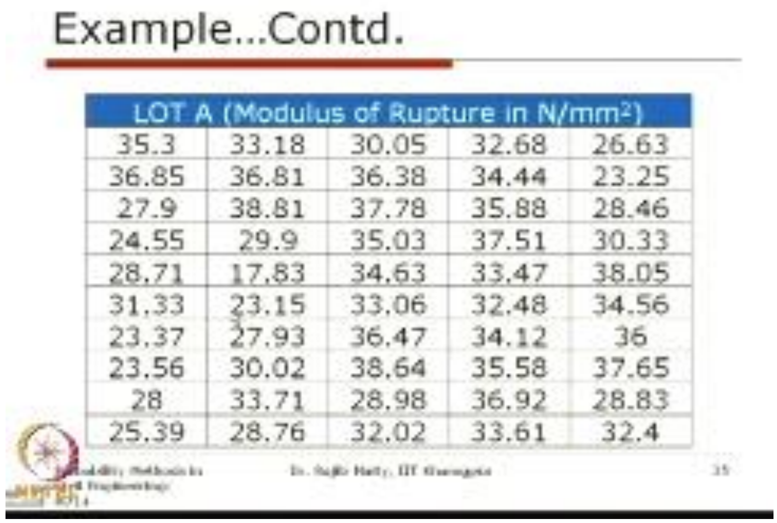
So now, the critical value that is that D m n should be less than equals to 1.36 / the square root of this full quantity, earlier it was square root of minnow it is square root of this m n / m + n that is the difference.

(Refer Slide Time: 50:39)



So, we will take one example here, the table showing the modulus of rupture data for two different groups of timber is shown in the next slide. Supplier delivered items in two lots. The first lot consists of the 50 samples and the second lot consists of 30samples.Boththe lots were supplied / the same supplier and the second lot is claimed to be superior to the first lot. Apply the K S test, the Kolmogorov-Smirnov test, two sample test, to verify whether the two samples are from the of the same type or not, that is, whether in the statistical sense, we should say that whether both the samples are from the same population or not. So, this is what we have to decide and this50samples and 30samples for all these timbers what is the modulus of rapture is shown in this table.

(Refer Slide Time: 51:36)

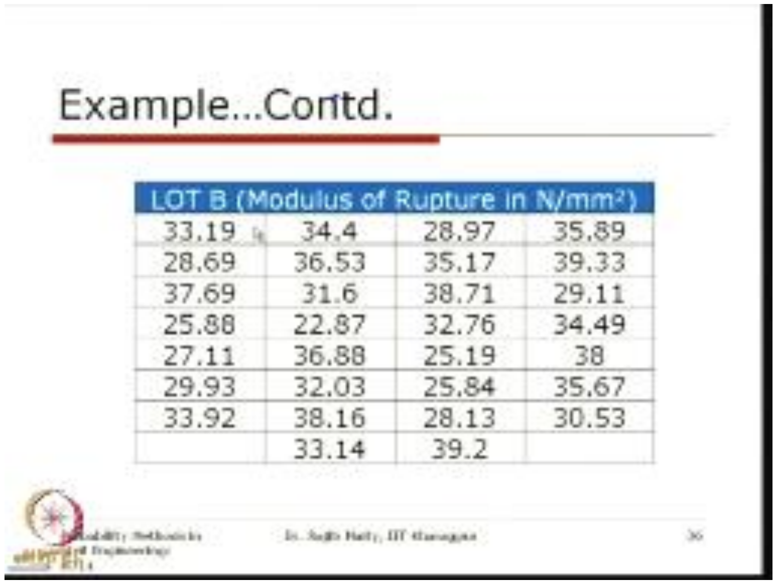


Example...Contd.

LOT A (Modulus of Rupture in N/mm²)				
35.3	33.18	30.05	32.68	26.63
36.85	36.81	36.38	34.44	23.25
27.9	38.81	37.78	35.88	28.46
24.55	29.9	35.03	37.51	30.33
28.71	17.83	34.63	33.47	38.05
31.33	23.15	33.06	32.48	34.56
23.37	27.93	36.47	34.12	36
23.56	30.02	38.64	35.58	37.65
28	33.71	28.98	36.92	28.83
25.39	28.76	32.02	33.61	32.4

This is the first lot that is lot A, this modulus of rapture is given in Newton per millimeter square. So, this is35.3and like this, you can see that this is 5 / 10 columns, so all these data refers to the modulus of rapture for the 50 samples supplied in lot A.

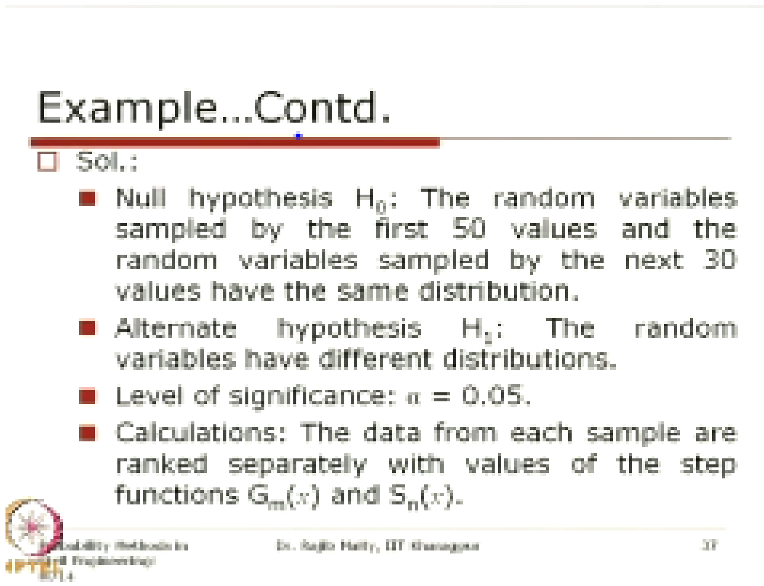
(Refer Slide Time: 52:03)



LOT B (Modulus of Rupture in N/mm²)			
33.19	34.4	28.97	35.89
28.69	36.53	35.17	39.33
37.69	31.6	38.71	29.11
25.88	22.87	32.76	34.49
27.11	36.88	25.19	38
29.93	32.03	25.84	35.67
33.92	38.16	28.13	30.53
	33.14	39.2	

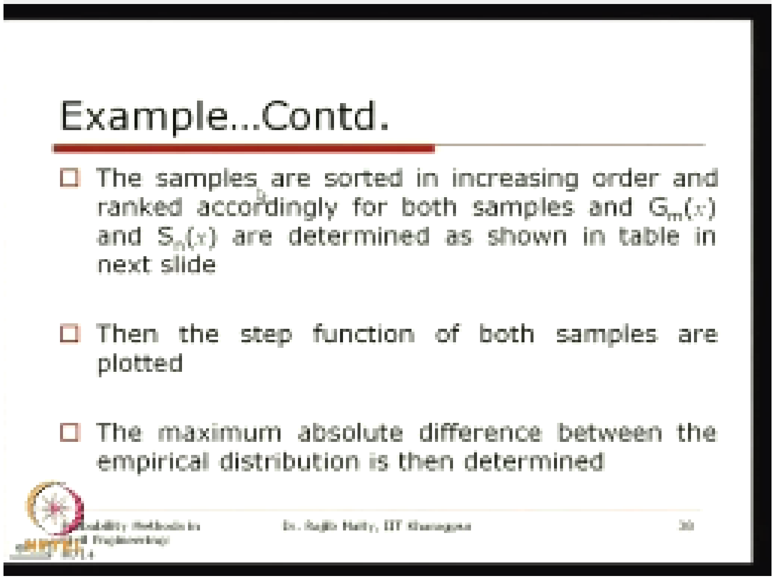
Similarly, in the lot B, these are30 data point of this modulus of rapture in newton per millimeter square which is supplied in the lot B. Now, we have to test whether both this dataset is the same is from the same population or not.

(Refer Slide Time: 52:20)



So, the null hypothesis is that the random variables sampled / thefirst50 values and the random variables sampled / the next 30 values, have the same distribution or not. And so and the alternative hypothesis are the random variables have the different distributions, so whatever we have hypothesized in the null hypothesis is not valid. Level of significance here is0.05 and the calculation that we have to do is that the data from the each sample are ranked separately, so we have to make it, we have to sort it. In the last example, it was already sorted and that data was supplied, but here we have to sort it, first we have to give the rank and from there we have to calculate their respective cumulative distribution, so both are that step functions.

(Refer Slide Time: 53:13)



The samples are sorted in an increasing order and the ranked accordingly for both the samples G m X and this SnX are determined as shown in this table in the next slide. And then, the step functions of the both samples are plotted, the maximum absolute difference between the empirical distribution is then determined. So, the basic steps what we have seen in the single sample and this two samples are same.

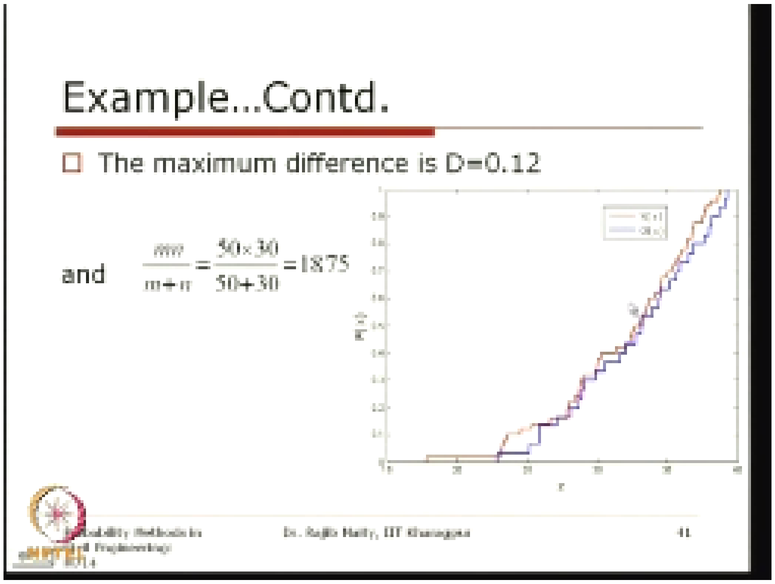
(Refer Slide Time: 53:40)

Example...Contd.

Rank	x at lot A	MR at A	Rank	x at lot A	MR at A
1	0.02	17.83	26	0.52	33.06
2	0.04	23.15	27	0.54	33.18
3	0.06	23.25	28	0.56	33.47
4	0.08	23.37	29	0.58	33.61
...	...	...	...	...	...
22	0.44	32.02	47	0.94	37.78
23	0.46	32.4	48	0.96	38.05
24	0.48	32.48	49	0.98	38.64
25	0.5	32.68	50	1	38.81

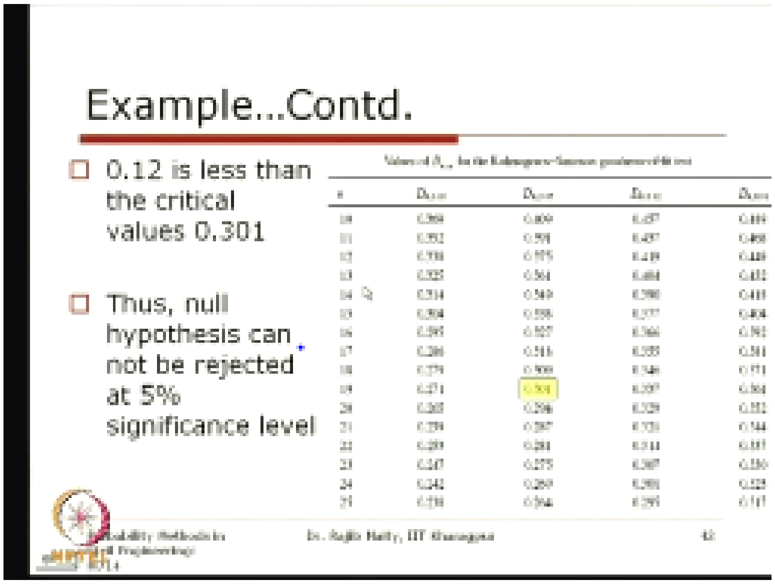
So, this is for the first lot, and you can see this rank 1, 2,this is shortened this is able is shortened just to accommodate in a single slide that the rank is 1,2,3,4 and this continuing up to 22,23,24,25,26,27 and going up to 50.Now, this is that value of that k / n, that is, that rank m / n, so it is 0.02,0.04, and0.006 like that from this0 it will go on, and it will come to that 1. So, this is for this lot A, and similarly, this is for the lot B, and if we plot this one,

(Refer Slide Time: 54:25)



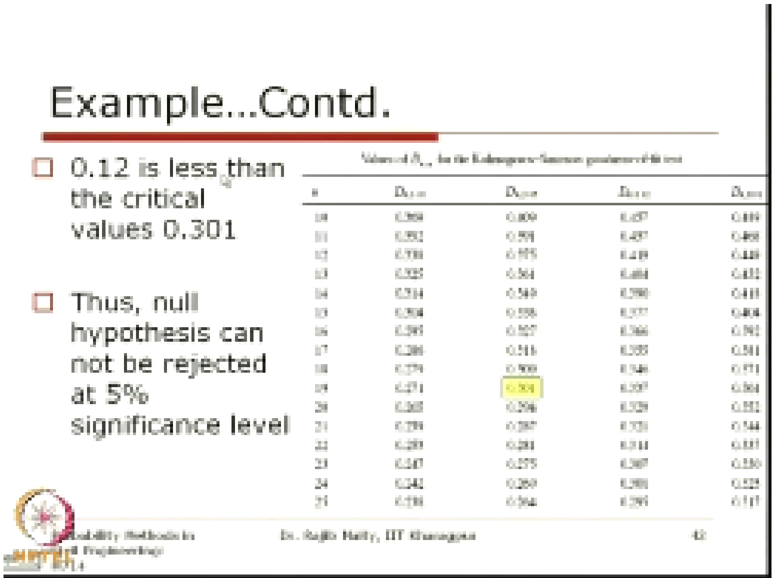
It looks like this two plots are there, one is that shown by red and the other one is that blue. Now, the maximum difference we can observe that the difference and we can get it and this maximum difference is found to be 0.12 and as I told that now, there are two samples, one is that m is the50 data samples and n is the 30, so if we just get it becomes at 18.75, so we have to see what is the critical value against this sample size of 18.75.

(Refer Slide Time: 55:03)



And here if you see, that this sees approximately we have taken is 19 and obviously that for the proper value. We can go for this linear interpolation between 18 and 19,but here we have just taken this 0.031 as against this n equals to 19 and this D n 0.05, which is this significance level - at 5 percent significance level. So, this so 0.12 that is what the maximum difference that we got here.

(Refer Slide Time: 55:03)



Is now less than critical values of this point critical value not s so critical value of 0.301 which I have seen from this table, so thus the null hypothesis cannot be rejected at 5 percent significance level. So, that means, both the samples are basically from the same population, so what the supplier has claimed that the second sample is superior than the first one is not validate at least at this5 percent significance level.

So, in this lecture, we have discussed two statistical test to test whether now, what type of particular distribution if particular sample is following through two statistical test, one is the chi square test and second one is the KS test. Generally, the chi square test can be applied for both that discrete and continuous random variable and this KS test is for the continuous random variable, but most of this application, whether the dataset is following the normal distribution or not that we test, and also we test that whether the two samples that we are having whether, they are following the same distribution or not, that is what we can test using this K S test. So, we will take up some other test in the next lecture; thank you.

Probability Methods in Civil Engineering

End of lecture 38

Next:"Regression Analyses and Correlation"

In lecture 39

NPTEL Video Recording Team
NPTEL Web Editing Team
Technical Superintendents
Computer Technicians

A IIT Kharagpur Production
www.nptel.iitm.ac.in