Probability Methods in Civil Engineering

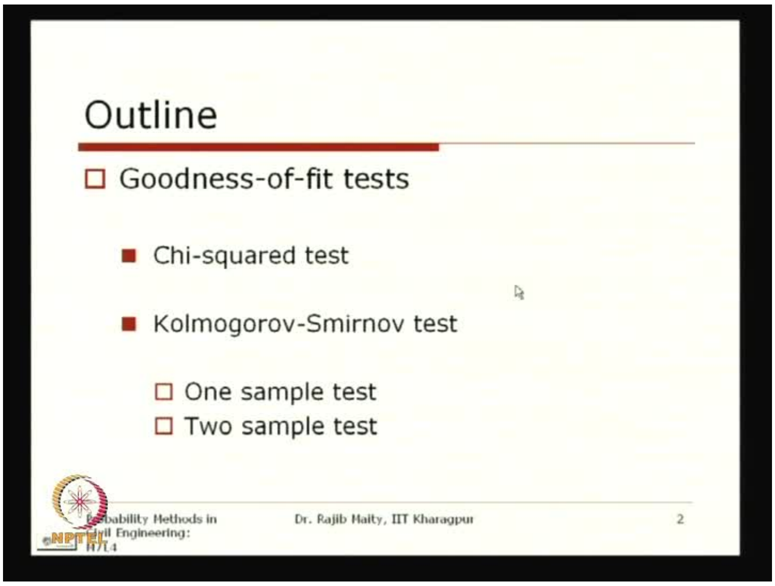
Prof.Dr.RajibMaity
Department of Civil Engineering
Indian Institute of Technology, Kharagpur

Lecture No.# 38
Goodness - of – fit tests

Helloandwelcometothislecture;inthislecture,we will learn some of the statisticalmethods to testthe observed data,whether they are fitting to some particular distribution or not.You know that in earlier lectures, as well as in earliermodules also, we havereferred several time thatwhile handling some problem,wegenerally assume that the dataset that is followinga particular distribution andwe have seennow,we will see that howwe can testthese thingsthat whetherthe dataset is reallyfollowing this distribution or not.

There are some of thestatistical test that I am just going totell in this in this lecture is that,so to use that one and use the knowledge of this hypothesis testing will betestingits statistically that howthe data is fitting to a particularmodule.So, there are some tests and so we will just pick up some thispopulartest to see that how we cantest thisgoodness of fit test, these test are known as the goodness of fittest,so, this is,sothat is our today’s lecture title is theGoodness of fittest.

(Refer Slide Time: 01:36)



Outline

□ Goodness-of-fit tests

- Chi-squared test
- Kolmogorov-Smirnov test

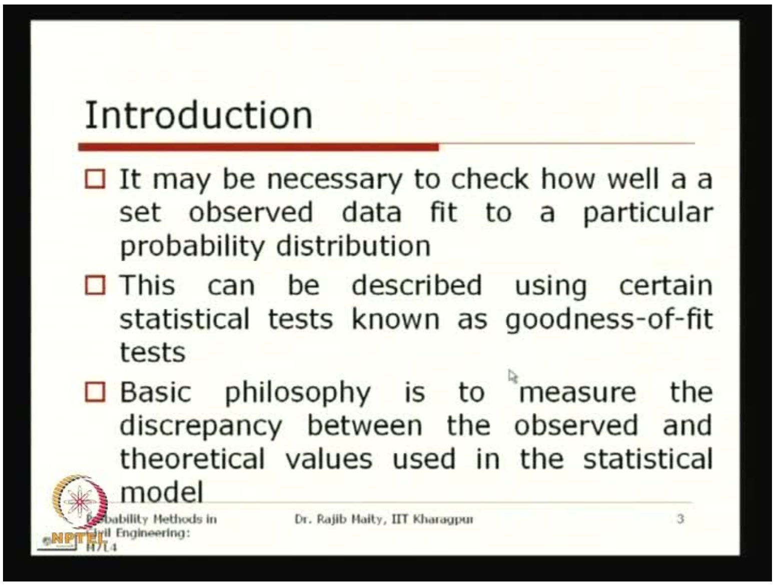
  □ One sample test
  □ Two sample test

Probability Methods in Civil Engineering: H7L4
Dr. Rajib Maity, IIT Kharagpur
2

So,first we will justsee that howthis canwork, and after that we will taketwo test for this lecturethat 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 thekolmogorov smirnovtest,it is also known as this K S testas anabbreviated form.And this K S test again can be that one sample test and two sample test.Basically,when we take thatthis one sample test,then we generally test it forone sample that we take and we generally take that fromwhether that particular dataset follow aparticular 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 andwe look for the answer whetherboth the samples are from the same population or not,same population means,there are following the same distribution or not. So, this ishow we go for one sample test and thistwo sample test.As you know that for this kind of statistical test,weneed some kind ofsignificance level andsome statisticalsignificance level. So, whenever we generally conclude or we draw somedecisionfrom whatever the statistical test we do, we have generally that decision is associated withstatistical significance level,so that is important.

(Refer Slide Time: 03:42)



Introduction

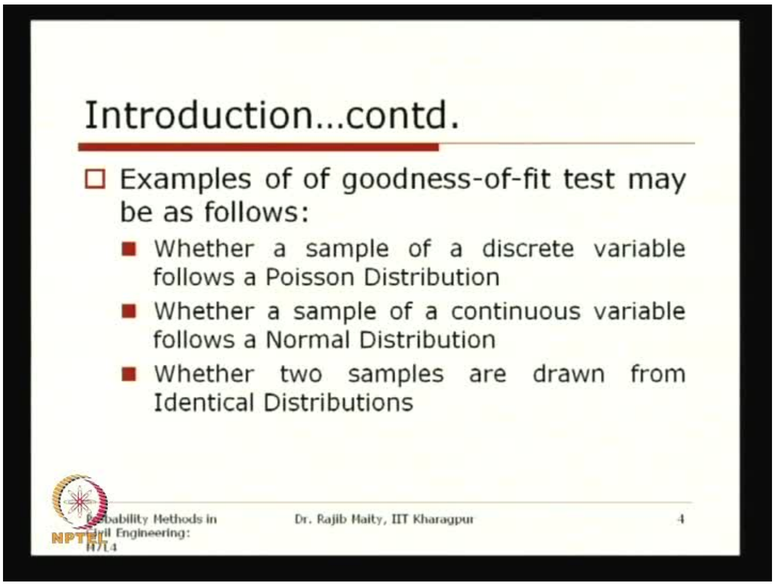
- □ It may be necessary to check how well a a set observed data fit to a particular probability distribution
- □ This can be described using certain statistical tests known as goodness-of-fit tests
- □ Basic philosophy is to measure the discrepancy between the observed and theoretical values used in the statistical model

Probability Methods in Civil Engineering: M7L4 Dr. Rajib Maity, IIT Kharagpur 3

So, what significance level that we are considering eitherbeforehand thatok,at thissignificance level,whatever we are going to test is satisfied or notin a statistical sense that is what we will do.Well, so now, first,we will goto this, whatever I justmentionedso far is that, it may be a necessary to check howwell a set of observed datafit to aparticular probability distribution, and this one, this can be describedusing the certainstatistical test known as the goodness-of-fit test.Andas also, I just try to indicate that the basic philosophy is to measure the discrepancy between the observed and theoretical values used in thestatistical model.So, this basically are the generalthing forallthetestthat 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, sowhat is our interest is that, whetherthere is any discrepancy between that whatever theobserveddistributions that we can see from this data.And whatever thedistribution 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,thatdiscrepancy we have to see, and that discrepancywe have to decide whichthrough some distribution, we have todraw someinference in a statistical way.

(Refer Slide Time: 05:15)



Introduction...contd.

□ Examples of of goodness-of-fit test may be as follows:

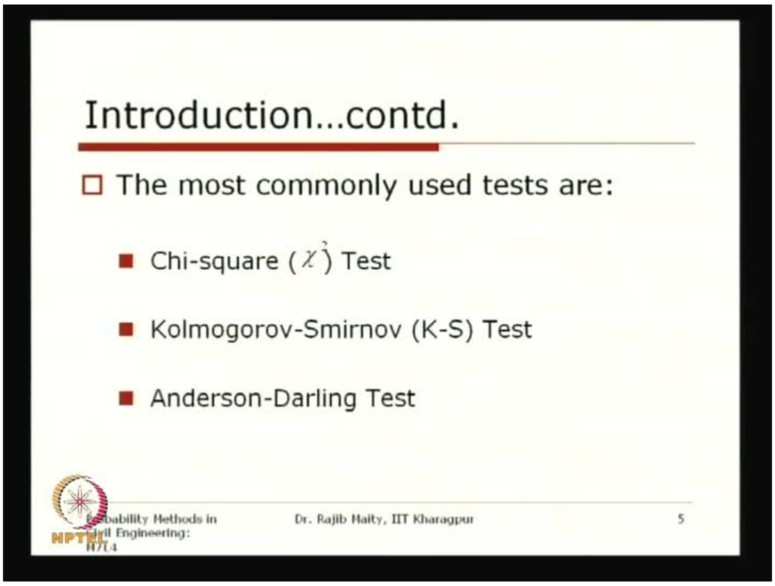
- Whether a sample of a discrete variable follows a Poisson Distribution
- Whether a sample of a continuous variable follows a Normal Distribution
- Whether two samples are drawn from Identical Distributions

Probability Methods in Civil Engineering: M7L4 Dr. Rajib Maity, IIT Kharagpur 4

So,the example of thisgoodness–of-fit test may be as follows,say,thus that what type of question that we are looking for theiranswer, say that, whether a sample of a discrete variablefollows 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, soI know that this randomvariable is a discrete variable,so I have somehypothesize that whether thatsample can follow a particular discrete distribution thatwe know already, and those distribution we have discussed inearlier 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 log normal distribution or not like that,orin case 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 ofthis problems, where we will try to cover almost all thesedifferent cases discrete variable case,continuous variable case, as well as thetwo samples case taking from two different sample we will take and whetherwe willtest that whether they are from the identical distribution or not.

(Refer Slide Time: 06:52)



Introduction...contd.

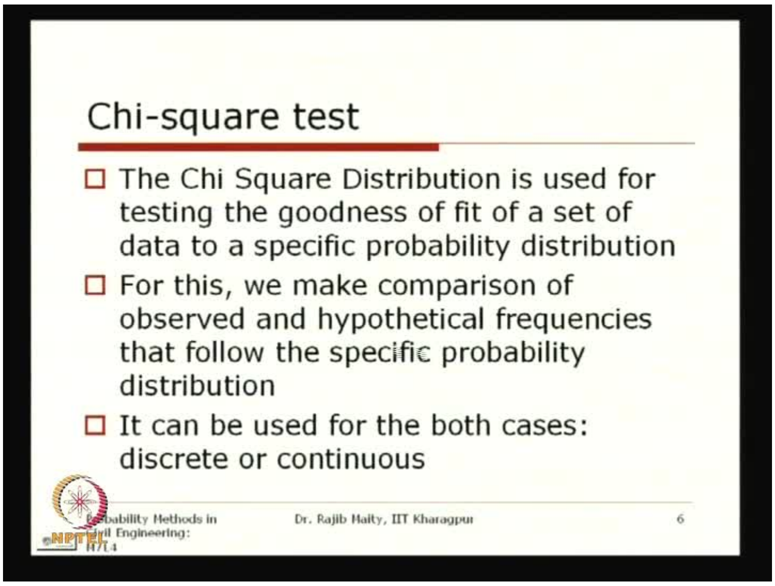
□ The most commonly used tests are:

- Chi-square (χ^2) Test
- Kolmogorov-Smirnov (K-S) Test
- Anderson-Darling Test

Probability Methods in Civil Engineering: M7L4 Dr. Rajib Maity, IIT Kharagpur 5

So, aswe mentioned, there aresome of thesecommonly used test,are therethese are standard test that we generally use forthis goodness-of-fit test.The first one is this, chi square test and then, the kolmogorov smirnov test, there alsoother test which is a Anderson-darling test which is generally known to be a littleimprovement over this K S test.And we will take this one also, but may be for this lecture we will justconsider thischi square test and K S test.

(Refer Slide Time: 07:32)



Chi-square test

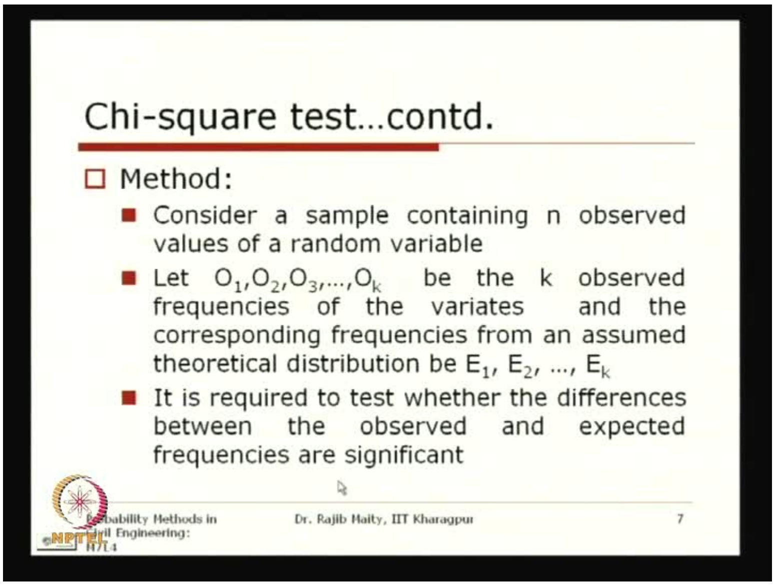
- □ The Chi Square Distribution is used for testing the goodness of fit of a set of data to a specific probability distribution
- □ For this, we make comparison of observed and hypothetical frequencies that follow the specific probability distribution
- □ It can be used for the both cases: discrete or continuous

Probability Methods in Civil Engineering: M7L4 Dr. Rajib Maity, IIT Kharagpur 6

Now, this chi square test, the chi square distribution is used for testing the goodness 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 basicphilosophy.And this is basically, if you see thatoverallapproach is same for all this test,so the comparison of thiswhatever 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 frequencyand obviously, as we are talking about these frequencies, so we have tocategorize the data into different beansand each bean what is thefrequency, 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 onedistribution.So, based on that distribution obviously, that distribution should have some parameters, with thatparameters, what are thefrequency that we can observed.And this one can be used for the both the cases; means, both for thisdiscrete random variable as well as for the continuous randomvariable, we can use this test.

(Refer Slide Time: 09:04)



Chi-square test...contd.

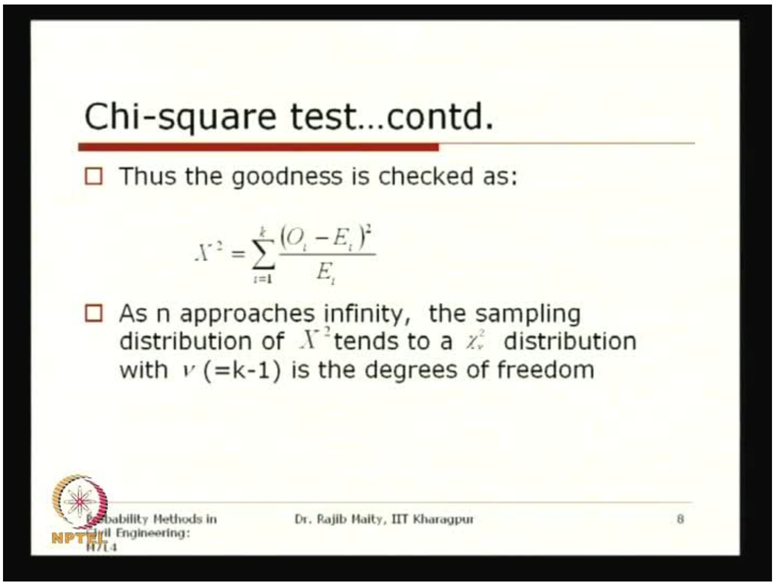
- □ Method:
  - Consider a sample containing n observed values of a random variable
  - Let $O_1, O_2, O_3, \ldots, O_k$ be the k observed frequencies of the variates and the corresponding frequencies from an assumed theoretical distribution be $E_1, E_2, \ldots, E_k$
  - It is required to test whether the differences between the observed and expected frequencies are significant

Probability Methods in Civil Engineering: NPTEL Dr. Rajib Maity, IIT Kharagpur 7

So, what isdone that may be we will discuss nowin a step in different steps,first of all let us consider that a samplecontains n 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 thisnotation, that is,O1,O2,O3up to OK, be the K observed frequencies of the variates and the corresponding frequencies from the assumed theoretical distribution be E 1,E 2,E 3 up to EK.So, this is that observed, that is,this is from whatever that data that we are having is the O 1, O 2,O 3,OKand these E1,E2,E3,EK, these are from that whatever the distribution,that we are hypothesize that this dataset may follow that particular distribution.So, from that distribution whatever the frequency that we are gettingis,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 datawe 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 particulardistribution.

(Refer Slide Time: 10:53)



Chi-square test...contd.

□ Thus the goodness is checked as:

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

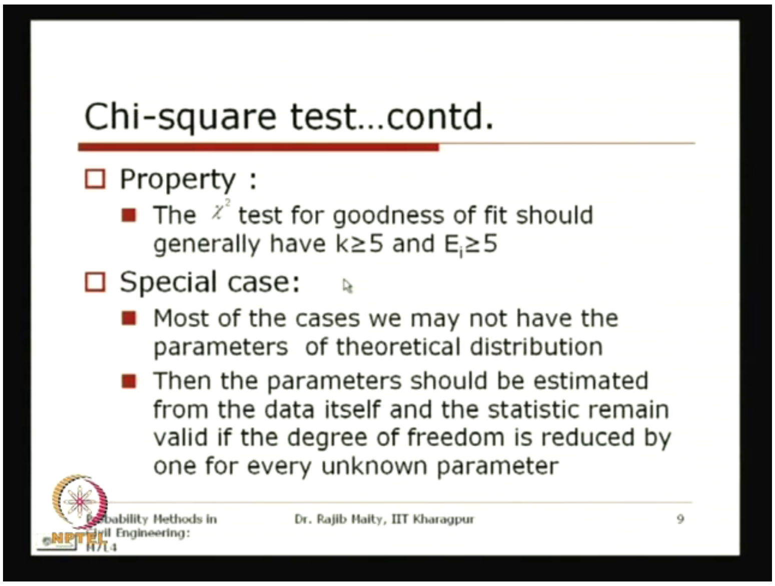
□ As n approaches infinity, the sampling distribution of X^2 tends to a χ^2_ν distribution with ν (=k-1) is the degrees of freedom

Probability Methods in Civil Engineering: M7L4 Dr. Rajib Maity, IIT Kharagpur 8

Now, so,how to do that one, how tofind out that discrepancy is through thisparameter, we generally call itas a statistics and that statistics is obtained as that Oiminus Eiwhole square divided by E iand sum of are allKSmeans, all thisgroups 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 s square that we have denoted this one, and it has been found that this X s square or this statistics,it followsa chi square distribution.If n tends to infinity, that is,thatthat mathematical limitfor the large number of thissample data, that is available, this statistics follow a chi square distribution and this chi square distribution is having somedegrees of freedom,here the degrees of freedom is k minus 1.

So, this one show, aswe know that this particular statistics follows the chi square distribution with k minus 1 degrees of freedom; that means, we know that what areitsproperties, and based on that at what significance level I want to test whether,this ismeanssignificantly,this is significant or notthat we have to test.One thing is, I should mention here that sometimes it is required,for example that when we are calculating this E i,it may require to estimate some of the parameterof 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 datathen,thosemany degrees of freedom will be lost.So,if you are not estimating any parameter from that data available, that time this the degree of freedom is k minus 1, but if we estimate one parameter then, this degrees of freedom will be k minus 2, if you are estimating 2 parametersthen degrees of freedom will be k minus 3.

(Refer Slide Time: 13:16)



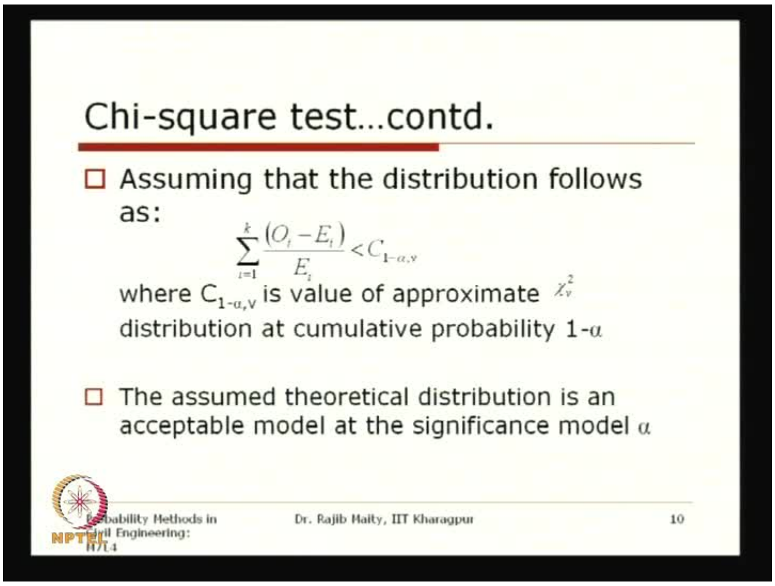
Chi-square test...contd.

- □ Property :
  - The χ^2 test for goodness of fit should generally have k≥5 and E_i≥5
- □ Special case:
  - Most of the cases we may not have the parameters of theoretical distribution
  - Then the parameters should be estimated from the data itself and the statistic remain valid if the degree of freedom is reduced by one for every unknown parameter

Probability Methods in Civil Engineering: NPTEL Dr. Rajib Maity, IIT Kharagpur 9

So,similarly, so as many parameters, you are estimating from the data those many more degrees of freedomwill be lost. So, now this onesome morethings, some moreproperties, these are notthat,means, directly I cannotsay that why these are required,but for the betterresults that is within quote, that for the betterresult, 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 havedivided that observed data it should be greater than equal to5 and that Eiis that is thefrequencies that we got in each beam should be at least5, so this Eishould be greater than equal to 5.

(Refer Slide Time: 14:36)



Chi-square test...contd.

□ Assuming that the distribution follows as:

$$\sum_{i=1}^{k}\frac{(O_i - E_i)}{E_i} < C_{1-\alpha,v}$$

where $C_{1-\alpha,v}$ is value of approximate χ^2_v distribution at cumulative probability 1-α

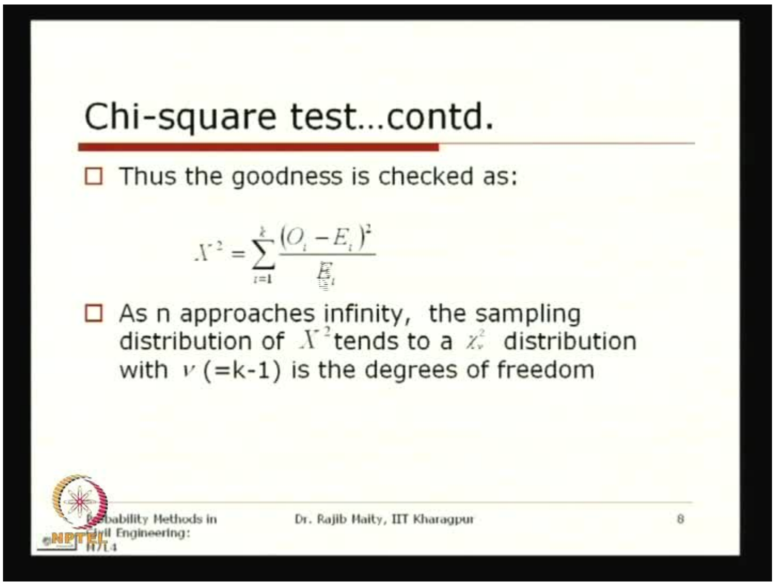
□ The assumed theoretical distribution is an acceptable model at the significance model α

Probability Methods in Civil Engineering: M7L4 Dr. Rajib Maity, IIT Kharagpur 10

Now, the special case,the most of thecases 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 tothen you need to allowthat much degrees of freedom will be lost. Now,assuming that the distribution follows as this onethat I told,there will be square here, as you have seen earlier this statistics,this observed minus, this hypothesize square divided by Eiand their summation.So,if this one is less than this value, what is shown is,C1 minus alphamuandthis mu as it is already explained that, this is the degrees of freedom,this C1 minus alpha mu, that is C1, minus alpha is a value of approximate that this chi square distribution with mu degrees of freedom at the cumulative probability1 minus alpha.

Now, this alpha is that significance levelthat we are talking about, the assumed theoretical distribution is anacceptable 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 keptaround, say, 0.01 or,sorry, 0.01or 0.05. So,at this cumulative probability one minus this alpha, so if it is that 5 percent significance level if we say; thatmeans that cumulative probability at which we are testing it is at the95 percent.So, if the observed statistics is less thanis less than that cumulativeprobability of that distribution,here it is the chi square distributionof this95 percent if the alpha is 0.05significance leveland then, we can say, yes, thatwhatever we havehypothesized may not be rejected.

(Refer Slide Time: 16:34)



Chi-square test...contd.

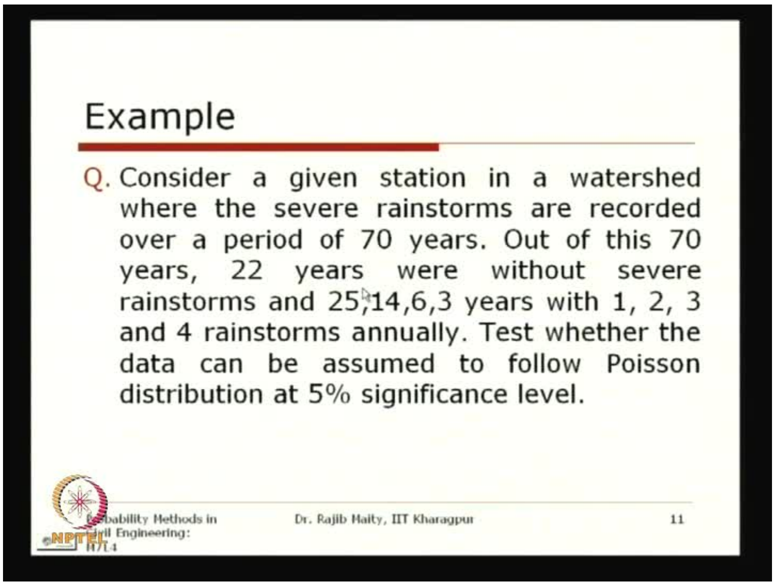
□ Thus the goodness is checked as:

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

□ As n approaches infinity, the sampling distribution of X^2 tends to a χ^2_ν distribution with ν (=k-1) is the degrees of freedom

Probability Methods in Civil Engineering: Dr. Rajib Maity, IIT Kharagpur 8

(Refer Slide Time: 16:40)



Example

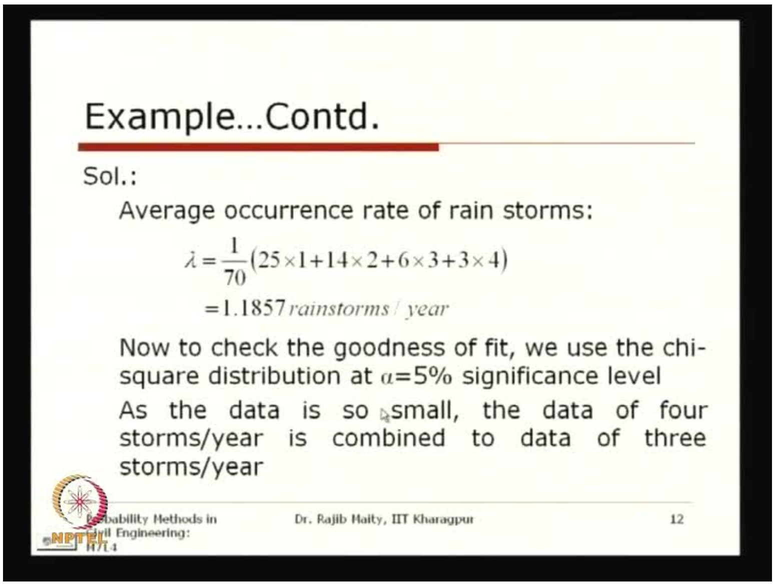
Q. Consider a given station in a watershed where the severe rainstorms are recorded over a period of 70 years. Out of this 70 years, 22 years were without severe rainstorms and 25,14,6,3 years with 1, 2, 3 and 4 rainstorms annually. Test whether the data can be assumed to follow Poisson distribution at 5% significance level.

Probability Methods in Civil Engineering: Dr. Rajib Maity, IIT Kharagpur 11

So, yes, as I told that this should be thesquare, as we have seen it here also, thisstatistics,this y minus Eipower square. Now, we will take up one example and this example, we have taken for the discrete case and because,thischi squaredistribution can be used for the is for the discreteparameters. So,we have it can be used for the continuous as well as discrete, butthe other than next one what we are going to cover is, the K S test that isfor 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 70years,last70 years, we have recorded the how many rainstorms are there in a particular year.And out of these70 years,22 years were without severe rainstorms,so there are in 22years there are nosevere rainstorms, so number rainstorms is 0.And25 years,14 years, 6 years and 3 years are with 1 rainstorm,2 rainstorm,3 rainstorm and 4 rainstormsrespectively.So,25 years we have observed there is 1 rainstorm, 14years we have seen there are3 rainstorms,16 3 rainstorm, and 3years, there are4 rainstorms are there. Test whether the data can be assumed to followa poisson distribution at 5 percent significance level.So, here as you can see that we are hypothesizingthat whether the data that is giventhatwhatever we have recorded, whether it is following the Poisson distribution ornot, and the significance level is given as5 percent.

(Refer Slide Time: 18:19)



Example...Contd.

Sol.:

Average occurrence rate of rain storms:

$$\lambda = \frac{1}{70}(25\times 1+14\times 2+6\times 3+3\times 4)$$

$$=1.1857\, rainstorms / year$$

Now to check the goodness of fit, we use the chi-square distribution at α=5% significance level

As the data is so small, the data of four storms/year is combined to data of three storms/year

Probability Methods in Civil Engineering: M7L4 Dr. Rajib Maity, IIT Kharagpur 12

So, here now you see,so the Poisson distribution you know there is somelambda, there is one parameter that we have discussed earliermodulus, thatparameter that lambda is the mean rate of occurrence. So, rate so average occurrence rate of this rainstorm,so if you want to calculatethen we have seen that there are22 years where there is norainstorms. So,22 multiplied by 0,basically, the first thing,then 25 years 1 rainstorm, 14 years2 rainstorm,6 years3 rainstorm and 3 years 4 rainstorms are there are total. So, in this way out of these70years what are how many total rainstorms that we have seenthat divided by70 gives youthat average rainstorm that can occur in ayear,so 1.1857 rainstormsyou can see in a year.So, lambda here is that 1.1857 and the tthat we are- lambda t - that is the totalquantity that t is here the per year,so one year how muchrainstorms has occurred, so this is the quantity 1.1857.

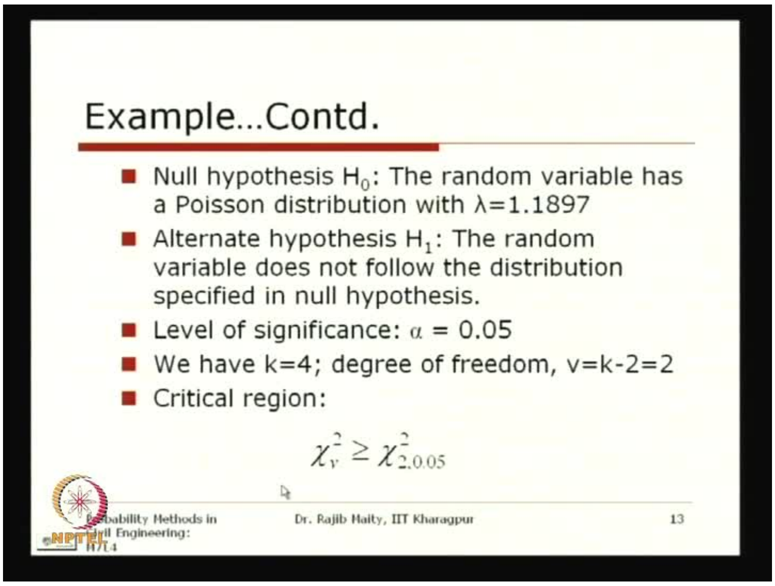
Now remember that, we have estimated this one from the data itself,so this was not supplied.Sometimes what happens?these data could have been supplied directly, whether the test,whether the data is following the Poisson distribution with the lambda is equals to, say,1.3 or 1.1 like that.If that is given to us that means, we are not estimating that one from the data, so there the degrees of freedom whatever I told that, it is k minus 1 will be there;but here, we have already estimated one parameter, sonow the degrees of freedom will bek minus 2.

So, now to the check, thegoodness- of-fit, we use the chisquare distribution atalpha equals to 5 percent significance level, as the data is so small, the data for the4 storms per year is combined with the3 storms per year. So,this one is generally done, but so there issome discussion is required.We canNow, we have to think that what distribution you have to hypothesize,so it is the Poisson distribution that you have hypothesize andwhat is the support for that distribution. Now, the Poisson distribution that support that we are looking for is basically starting from the 0 to0,1,2 the discrete values and goes up to infinity.

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 thethis isthe Poisson distribution of obviously, I cannotscuttle at any particular point.So, here what is the general practice is that for the higher side where data is becoming very small,wecan 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 thatwe will also betesting whether that each group is having that minimum requirement of thisfrequency is available or not.

Because this kind ofasymptotic distribution, thatis, it is goingtowards plus infinity, so this type of it should be open bounded,but the way the problem is given it is just as a close boundary at 4rainstormper year. So, here if we just considerthat2rainstorms are combined together, that is, the 3rainstorm per year and4rainstorm per year case, then we can say that will greater than equal to3rainstorm per year. So, that remains that, positive side remains unbounded and also, this will help tocheck whether thatat least greater than 5, that is, for the better result that we havemention earlier.

(Refer Slide Time: 22:25)



Example...Contd.

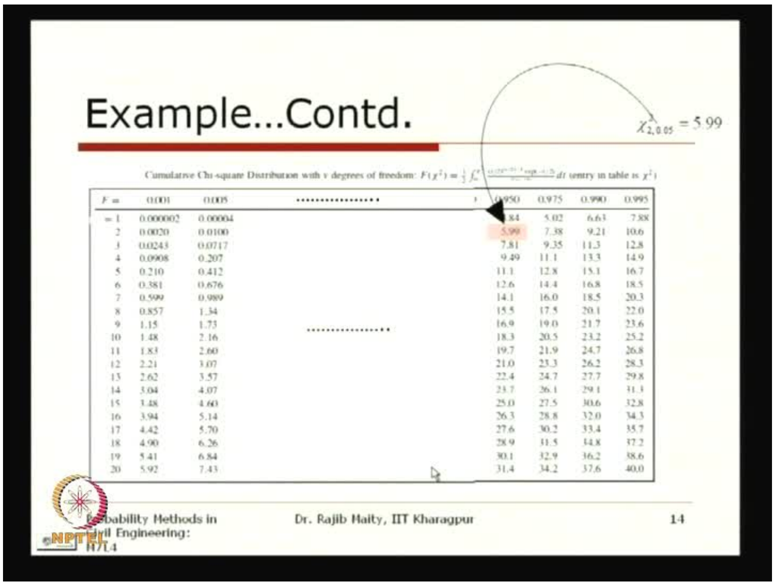
- Null hypothesis H_0: The random variable has a Poisson distribution with λ=1.1897
- Alternate hypothesis H_1: The random variable does not follow the distribution specified in null hypothesis.
- Level of significance: α = 0.05
- We have k=4; degree of freedom, v=k-2=2
- Critical region:

$$\chi_v^2 \ge \chi^2_{2,0.05}$$

Probability Methods in Civil Engineering: Dr. Rajib Maity, IIT Kharagpur 13

So, here also what we are doing is that the 4storms per year and the 3rainstorms per year are combined together. So, here the null hypothesis is the random variable has a poisson distribution with lambda equals to 1.1897;alternatehypothesis is the random variable does not follow the distribution specified in null hypothesis, what is specified here?Significance level is 0.05 at 5 percent significance level mentioned.And we have thek equals to 4, so the degrees of freedom here are the k minus 2, sok minus 1 is from there and one parameter we have estimated,so it isk minus 2, the degree of the freedom is 2. So, that criticalregion hereis thischi square distribution, that is, chi square distribution with2 degrees of freedom at alpha equals to0.05.

(Refer Slide Time: 23:11)



Example...Contd.

$\chi^2_{2,0.05} = 5.99$

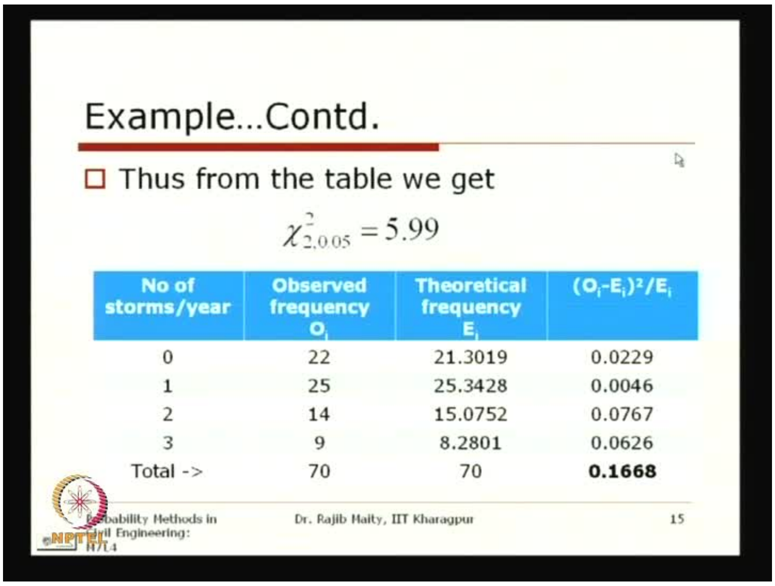
Cumulative Chi-square Distribution with ν degrees of freedom: $F(\chi^2) = \frac{1}{2}\int_0^{\chi^2} \frac{(t/2)^{(\nu-2)/2}\exp(-t/2)}{\Gamma(\nu/2)} dt$ (entry in table is χ^2)

F =	0.001	0.005	....	0.950	0.975	0.990	0.995
ν = 1	0.000002	0.00004		3.84	5.02	6.63	7.88
2	0.0020	0.0100		5.99	7.38	9.21	10.6
3	0.0243	0.0717		7.81	9.35	11.3	12.8
4	0.0908	0.207		9.49	11.1	13.3	14.9
5	0.210	0.412		11.1	12.8	15.1	16.7
6	0.381	0.676		12.6	14.4	16.8	18.5
7	0.599	0.989		14.1	16.0	18.5	20.3
8	0.857	1.34		15.5	17.5	20.1	22.0
9	1.15	1.73	....	16.9	19.0	21.7	23.6
10	1.48	2.16		18.3	20.5	23.2	25.2
11	1.83	2.60		19.7	21.9	24.7	26.8
12	2.21	3.07		21.0	23.3	26.2	28.3
13	2.62	3.57		22.4	24.7	27.7	29.8
14	3.04	4.07		23.7	26.1	29.1	31.3
15	3.48	4.60		25.0	27.5	30.6	32.8
16	3.94	5.14		26.3	28.8	32.0	34.3
17	4.42	5.70		27.6	30.2	33.4	35.7
18	4.90	6.26		28.9	31.5	34.8	37.2
19	5.41	6.84		30.1	32.9	36.2	38.6
20	5.92	7.43		31.4	34.2	37.6	40.0

Probability Methods in Civil Engineering: M7L4 Dr. Rajib Maity, IIT Kharagpur 14

Now, if you see,so this is that chi square distribution table, basically,you know that we have discussed this distributionearlier and this standard values are listed in any standard text book. So, here if you see that this point at this 0.95 these that cumulative probability are given here and for these degrees offreedom 2,this value is 0.5, 0.99,so these value is our critical value, that we have to test against.

(Refer Slide Time: 23:44)



Example...Contd.

□ Thus from the table we get

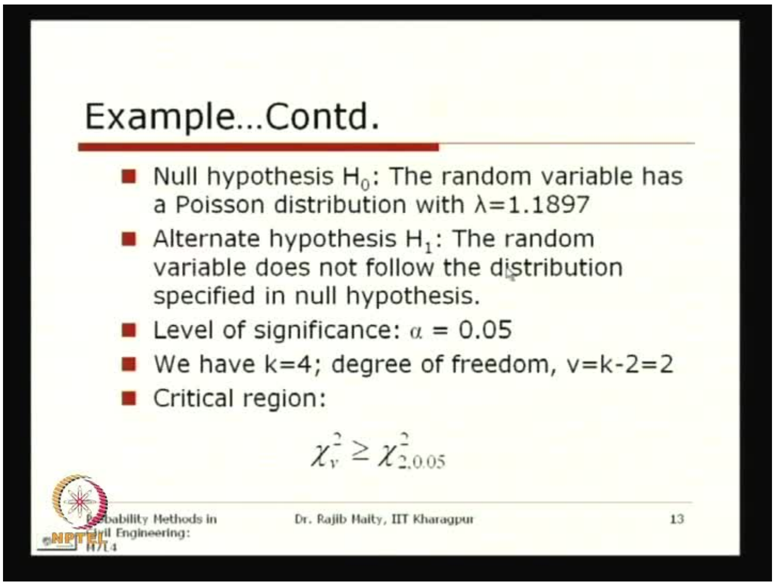
$$\chi^2_{2,0.05} = 5.99$$

No of storms/year	Observed frequency O_i	Theoretical frequency E_i	$(O_i-E_i)^2/E_i$
0	22	21.3019	0.0229
1	25	25.3428	0.0046
2	14	15.0752	0.0767
3	9	8.2801	0.0626
Total ->	70	70	**0.1668**

Probability Methods in Civil Engineering: M7L4 Dr. Rajib Maity, IIT Kharagpur 15

So, thisis that 5.99 that we got from thistable andthese are thetheseare the beanshere now,so0rainstorm per year,1 rainstorm per year,2 rainstorm per year and this is the greater than equal to3 rainstormper year.So, now,as it is given in this data, that there are 22 years where this0 rainstorms are there and the 25 yearare thereis that 1 rainstorm,14years 2 rainstorm and 6 years 3 rainstorm and that3 years4 rainstorms are there, sowe have combined it here to get the 9 years data, where it is greater than equal to4 rainstorm per year.

(Refer Slide Time: 24:57)



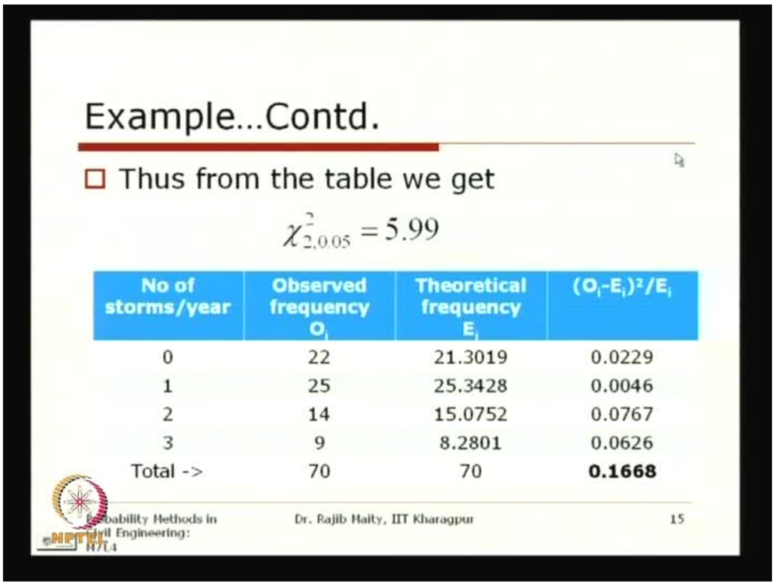
Example...Contd.

- Null hypothesis H_0: The random variable has a Poisson distribution with λ=1.1897
- Alternate hypothesis H_1: The random variable does not follow the distribution specified in null hypothesis.
- Level of significance: α = 0.05
- We have k=4; degree of freedom, v=k-2=2
- Critical region:

$$\chi_v^2 \geq \chi_{2,0.05}^2$$

Probability Methods in Civil Engineering: NPTEL Dr. Rajib Maity, IIT Kharagpur 13

(Refer Slide Time: 23:44)



Example...Contd.

□ Thus from the table we get

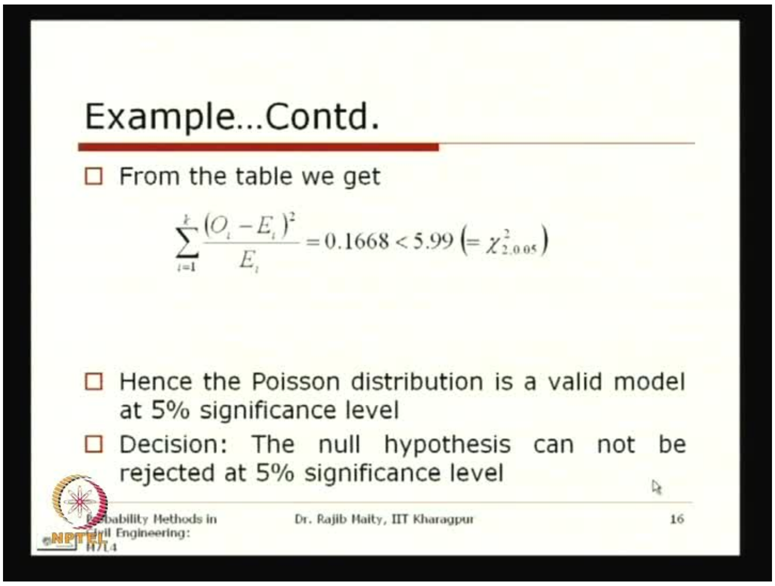
$$\chi_{2,0.05}^2 = 5.99$$

No of storms/year	Observed frequency O_i	Theoretical frequency E_i	$(O_i-E_i)^2/E_i$
0	22	21.3019	0.0229
1	25	25.3428	0.0046
2	14	15.0752	0.0767
3	9	8.2801	0.0626
Total ->	70	70	**0.1668**

Probability Methods in Civil Engineering: NPTEL Dr. Rajib Maity, IIT Kharagpur 15

So, we have avoided that one, in one occurrence it is becoming 3which is less than 5, so that is also avoided and that right side is kept open.And this theoretical frequencies, that is,thesevalues what we are getting is that this from this Poisson distribution that lambda t here equals to this lambda equals to that 1.1897 and t equals to 1,so1 year, so this lambda t value we will get,and from this poisson distribution if you just putthat x equals to 0, x equals to 1, x equals to 2 and x equals to 3 with that value of lambda t equals to that value, then we will get thisvalues,which are the, so are the theoreticalfrequencies for this Poisson distribution.

(Refer Slide Time: 25:58)



Example...Contd.

- □ From the table we get

$$\sum_{i=1}^{k}\frac{(O_i-E_i)^2}{E_i}=0.1668<5.99\left(=\chi^2_{2,0.05}\right)$$

- □ Hence the Poisson distribution is a valid model at 5% significance level
- □ Decision: The null hypothesis can not be rejected at 5% significance level

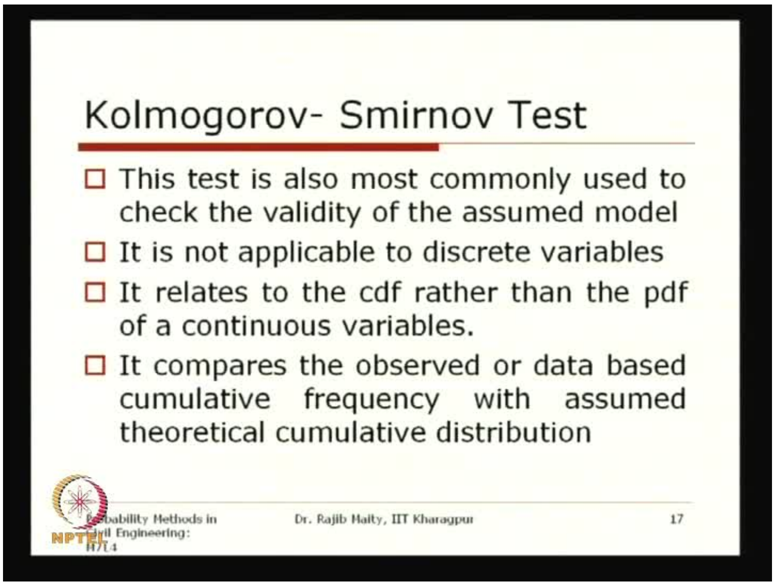
Probability Methods in Civil Engineering: IIT/L4 Dr. Rajib Maity, IIT Kharagpur 16

Now, so these areso 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 thisstatistics, that is, y minus Eiwhole square divided by Ei,so if we do this one then we get these values and we have to sum it upthis one, so this is the summation from is that0.1668.

So,theseonethis isourstatistics which is equals to0.1668 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 this critical value, so hence thePoisson distribution is a valid model at a5 percent significance level.So, the decision isthat for this test is that the null hypothesis cannot be rejected at 5 percent significance level.

So,there could be some other words that we can express, ok,the null hypothesis is not rejected and all,so what I feel is that this should be the proper decision, properprobabilistic or the proper statistical inference should be written as this one thatnull hypothesis is acceptedis not theright thing to declare.So,rather the complete thing that we can declare isthat the null hypothesis cannot be rejected at 5 percent significance level because, the result that we have got it depends on what significance level that we have optedfor. So, that is what we have to mention that at this significance level the null hypothesis cannot be rejected.

(Refer Slide Time: 27:19)



Kolmogorov- Smirnov Test

- □ This test is also most commonly used to check the validity of the assumed model
- □ It is not applicable to discrete variables
- □ It relates to the cdf rather than the pdf of a continuous variables.
- □ It compares the observed or data based cumulative frequency with assumed theoretical cumulative distribution

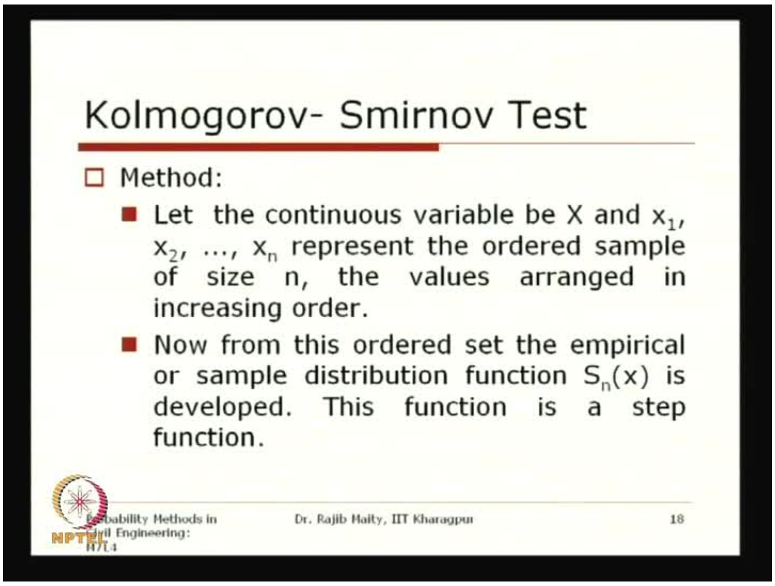
Probability Methods in Civil Engineering: IIT/L4 Dr. Rajib Maity, IIT Kharagpur 17

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 alsomost 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 dataset is availablewith thatwe generally get this one. It relates to the cdf rather than the pdf to a continuousvariable,so here earlier what we have seen that the frequency that when we are talking about in this chi square test that is basicallya pdf that we are talking about. Here, this K S test is generally based on this what is therecdf, that is cumulative distribution function, and it compares the observedor data based cumulative frequency with the assumed theoretical cumulative distribution, sodata based cumulative distribution, sorry, this should be distribution with the assumed theoretical cumulative distribution.

So,there are some kinds ofdiscrepancies orshortcomings I should say, in this chi square test is that,we need to define that,we need to first of allget thatthat beans thatwe have to first categorize the dataset, the full dataset that is, ok,this is my range.And in the discrete it isfinethat whatever the example that we have seen,nowfor the as I told this chi square distribution is also applicable for the continuous distribution.

Now, ifwe take some continuous distribution in case of this chi square test, what we have to do, we have to firstwe 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 than 5,as well as each bean should have minimumthatminimum 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 directlygetting it is directlytheaccess it is results from the cdf itselfdirectly,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: 29:49)



Kolmogorov- Smirnov Test

□ Method:

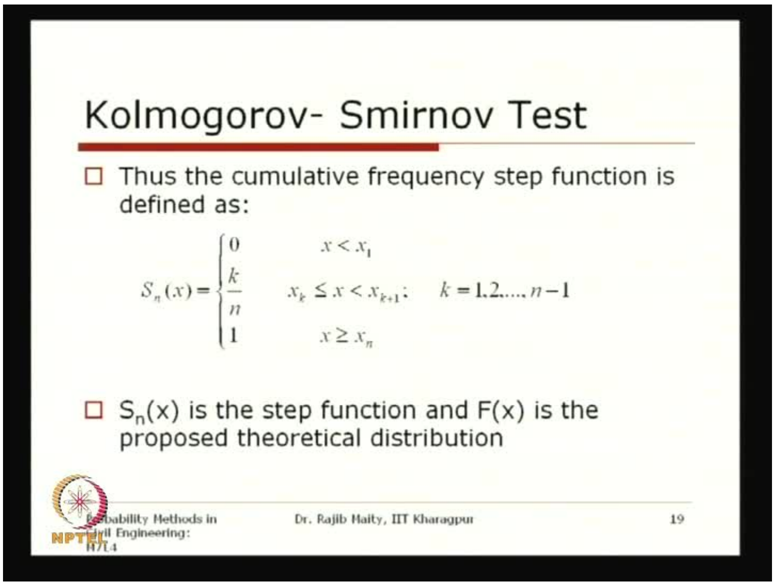
- Let the continuous variable be X and x_1, x_2, …, x_n represent the ordered sample of size n, the values arranged in increasing order.
- Now from this ordered set the empirical or sample distribution function $S_n(x)$ is developed. This function is a step function.

Probability Methods in Civil Engineering: Dr. Rajib Maity, IIT Kharagpur 18

Now,so, if suppose, that there is a random variable x and we havesome dataset of this X1,X2,X3 up to X n,so representthe 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 X1,X2,X nif that isavailable.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 from this sample how to get thatthis cumulative distribution function is as follows.

(Refer Slide Time: 30:37)



Kolmogorov- Smirnov Test

- □ Thus the cumulative frequency step function is defined as:

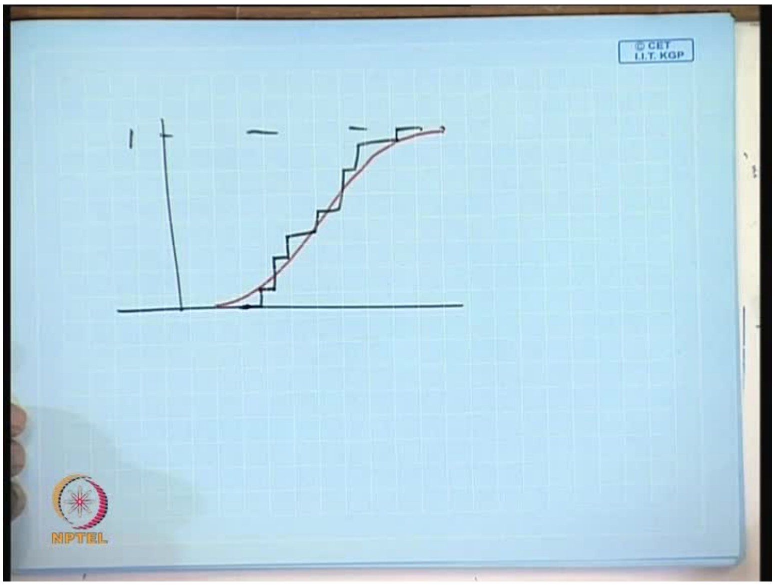
$$S_n(x) = \begin{cases} 0 & x < x_1 \\ \dfrac{k}{n} & x_k \le x < x_{k+1}; \quad k = 1,2,\dots,n-1 \\ 1 & x \ge x_n \end{cases}$$

- □ $S_n(x)$ is the step function and F(x) is the proposed theoretical distribution

Probability Methods in Civil Engineering: NPTEL Dr. Rajib Maity, IIT Kharagpur 19

That is for, when this S n X is equals to 0,in case, when X is less than X1,S n X is equals to K by n, when this X is in between k to k plus 1 and k can vary from 1,2,3 up to n minus 1. So, basically for all this X1 to X n for this thing we are definingthese values,the value of thiscumulativedistribution is k by n. And forXgreater than n is equal to 1. So, this is basically what we aregetting is, whatever the representative, cumulative distribution directly from the data and each point it is changing, this value is changing, so it will look like a step function.

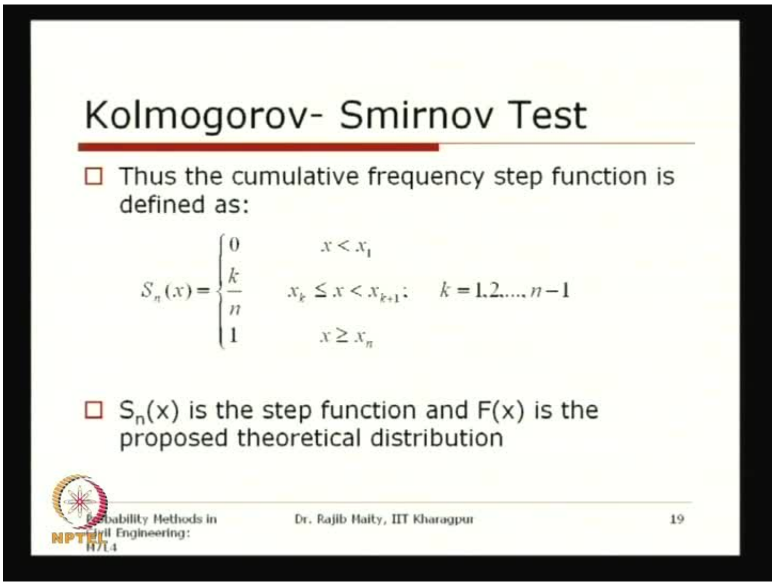
(Refer Slide Time: 31:40)



So, obviously as you can see this essence should will start from a value0,it will start from a value from 0 andat some point it will go, it will increase,and again from the next one, it should go andlike 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 thedistribution basically we got it from this data.Now, thehypothesized distribution, supposethat I want to match thatthat normaldistribution,so that normal distribution with the parameters of course, whether it is supplied or obtained from thissample data is that with that data, I can also plotwhat should be theshape of thisthat particular distribution, say,that normal distribution if I say.

(Refer Slide Time: 30:03)



Kolmogorov- Smirnov Test

- □ Thus the cumulative frequency step function is defined as:

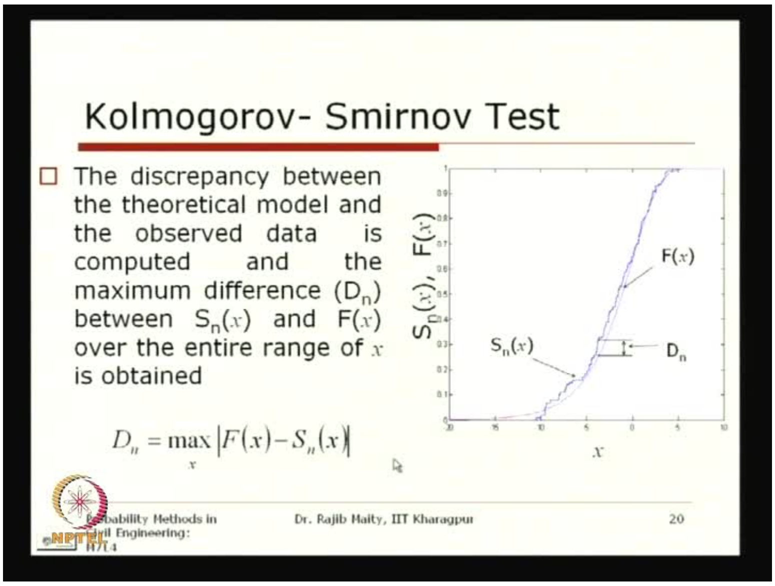
$$S_n(x) = \begin{cases} 0 & x < x_1 \\ \dfrac{k}{n} & x_k \le x < x_{k+1}; \quad k = 1,2,\ldots,n-1 \\ 1 & x \ge x_n \end{cases}$$

- □ $S_n(x)$ is the step function and F(x) is the proposed theoretical distribution

Probability Methods in Civil Engineering: NPTEL   Dr. Rajib Maity, IIT Kharagpur   19

Now, if whateverwe have observed the black line here, if this is very close to this red onewhich is the theoretical distribution obviously,then the data is from that particular distribution. So, this discrepancy, now again, keeping the sameapproach samefor 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 thatcdf, so one is the theoretical and the other one is theother one is thedatabase.So, that discrepancy we have to assess through somestatistical test,this is what, that is why we have got this S n X.

(Refer Slide Time: 33:14)



Kolmogorov- Smirnov Test

- □ The discrepancy between the theoretical model and the observed data is computed and the maximum difference (D_n) between $S_n(x)$ and F(x) over the entire range of x is obtained

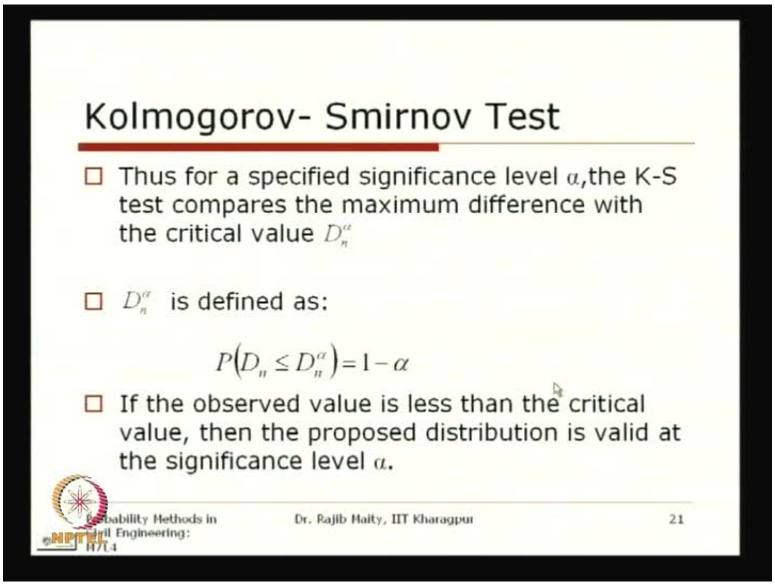
$$D_n = \max_x \left| F(x) - S_n(x) \right|$$

Probability Methods in Civil Engineering: NPTEL   Dr. Rajib Maity, IIT Kharagpur   20

So, this S n X is the step function and this FX is the proposed theoretical distribution that what Ihave drawn in the reading, just now.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 FX is over the entire range of X is obtained, which is which isdenoted as D n, which is the maximum for all X,FX minus S n X obvious theabsolute 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 thispink one is that youris that theoretical distribution. Now, the difference between any point to thattheoretical distribution is your that value is the difference betweentwo things.Now, what we have to pick up 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 thisand this pink line,sothat difference at each and every pointhave to find out and we have to select, we have to pick up the maximum one,so what is the maximumdifference.This blue line obviously,as we are coming leaps 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 maximumdifference.

(Refer Slide Time: 34:40)



Kolmogorov- Smirnov Test

□ Thus for a specified significance level α,the K-S test compares the maximum difference with the critical value D_n^α

□ D_n^α is defined as:

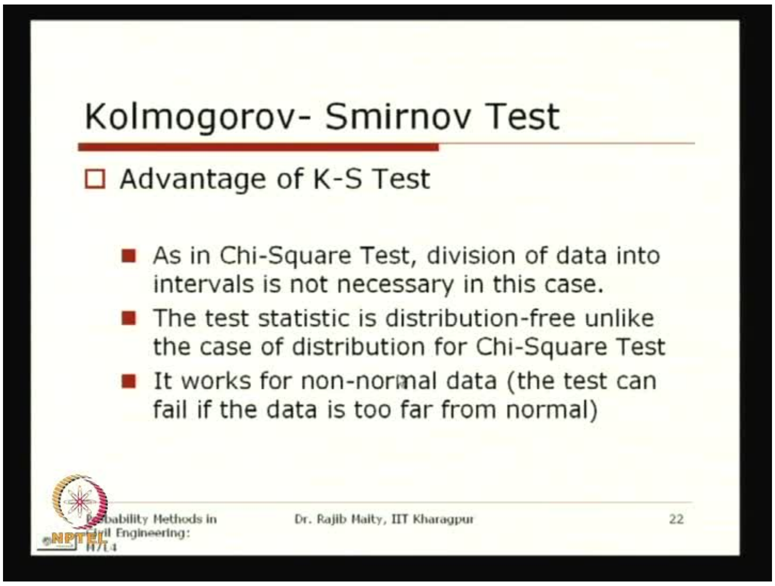
$$P\left(D_n \le D_n^\alpha\right) = 1-\alpha$$

□ If the observed value is less than the critical value, then the proposed distribution is valid at the significance level α.

Probability Methods in Civil Engineering: Dr. Rajib Maity, IIT Kharagpur 21

So, thus for aspecified significance level at alpha that K S test compares the maximum difference with the critical value D n alpha.Now, what is this D n alpha?ThisD n alpha is defined as the probability that D n less thanequals to D n alpha is equals to 1 minus alpha,again this alpha here is that significance level that we have mentioned at thebeginning of this class.So, if the observed value is less than the critical value, then the proposed distribution is valid at the significance level alpha.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 tothat critical value whether, it is equal to this1 minus alpha or not.

(Refer Slide Time: 35:39)



Kolmogorov- Smirnov Test

□ Advantage of K-S Test

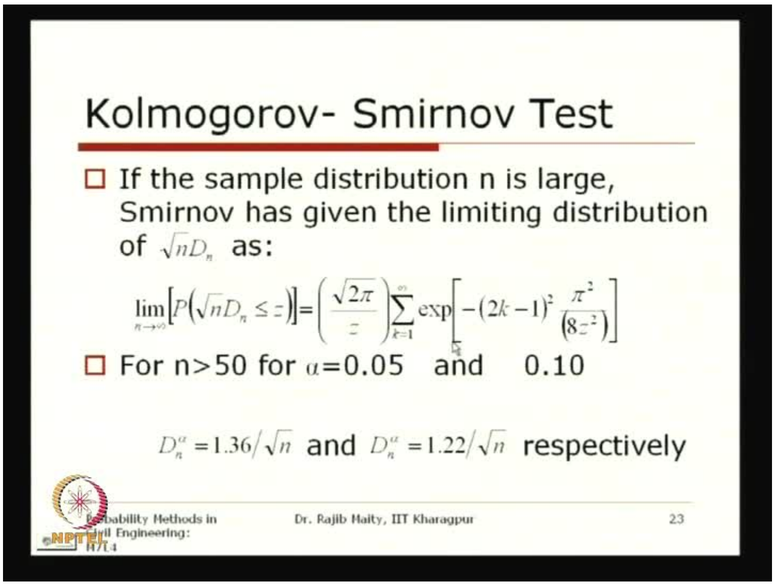
- As in Chi-Square Test, division of data into intervals is not necessary in this case.
- The test statistic is distribution-free unlike the case of distribution for Chi-Square Test
- It works for non-normal data (the test can fail if the data is too far from normal)

Probability Methods in Civil Engineering: NPTEL Dr. Rajib Maity, IIT Kharagpur 22

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,wesothese intervals arenot necessaryhere, because we are justobserving thisat eachand 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 log normal data,however, the test can fail if the data is too far from thisnormality.

So,there is no such restrictionthat with the data should be approximate normal or so, but it is better to get the better result again that data should be somewherenearnormality 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: 36:43)



Kolmogorov- Smirnov Test

- □ If the sample distribution n is large, Smirnov has given the limiting distribution of $\sqrt{n}D_n$ as:

$$\lim_{n\to\infty}\left[P\left(\sqrt{n}D_n \le z\right)\right]=\left(\frac{\sqrt{2\pi}}{z}\right)\sum_{k=1}^{\infty}\exp\left[-(2k-1)^2\frac{\pi^2}{(8z^2)}\right]$$

- □ For n>50 for α=0.05 and 0.10

$D_n^\alpha = 1.36/\sqrt{n}$ and $D_n^\alpha = 1.22/\sqrt{n}$ respectively

Probability Methods in Civil Engineering: M7L4 Dr. Rajib Maity, IIT Kharagpur 23

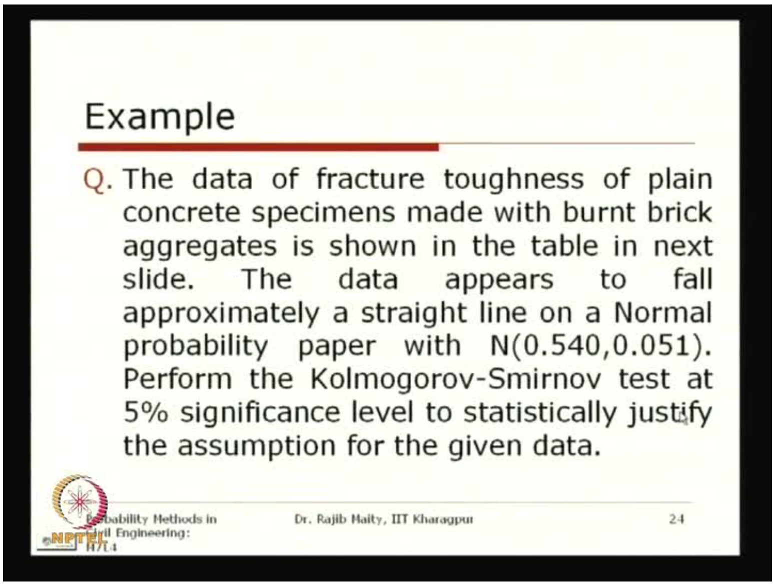
Now, the sample distribution,if the sample distribution that n is large, not the distribution sample size, if the sample size n is large,Smirnov has given the limiting distribution of square root n multiplied by this D n,so this is that D n that we have defined the maximum difference that multiplied by square root n, these quantity follow a distribution like this that limit n tends to infinity, probability of this quantity square root n multiplied by that D n,the statistic less than equals to z is equals to square root 2 pi by z multiplied by summation of k equals to 1 to infinity, exponential minus 2k minus 1 square multiplied by pi square by 8z square.

So, this is what it is giving is that for this n tends to infinity means, when this n isvery largethat time what we can get is that, how this D n is varying is thatthrough this, thatwe 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, thisprobability is equals to 1 minus alpha that means this 0.95.Now, if we solve this right hand part with equal to that, what isthat,0.95,then we will get that is z becomes 1.36,if you see this quantity hereinside this exponential term, that is,2k minus 1 whole square multiplied by pi square by 8z square.

So, this term basically is changing that is from this k equals to 1 to ininfinity, now k ifk is 1,you can see that this is just minus of this quantity plus exponential of,ifk becomes 2 then, it becomes 9, so minus 9,so exponential minus 9 times of this one and if k becomes3 then 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 equals to one only and remaining if you justignore it then, also you will see for this,this is almost very closelymatching with this onemay be, it is justvarying afterthird decimal or so.

So, for this n greater than 50, sometimes in some textbookcan refer to thatif is n is greater than 35 itself, and for this alpha equals to 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 consideronly one value of k equals to 1 then,you will see that this z becomesvery close to this 1.36.So,for this significance level alpha equals to 0.05 that critical value is 1.36 divided by square root of n,so that means this Z becomes 1.36 and this is that 1.22square root of n.So,this critical value this should be, so our observed statistics that is D n should be less than thisparticular value todeclare that,at that significancelevel, we can acceptthat particularhypothesis or that null hypothesis cannot be rejected.

(Refer Slide Time: 40:40)



Example

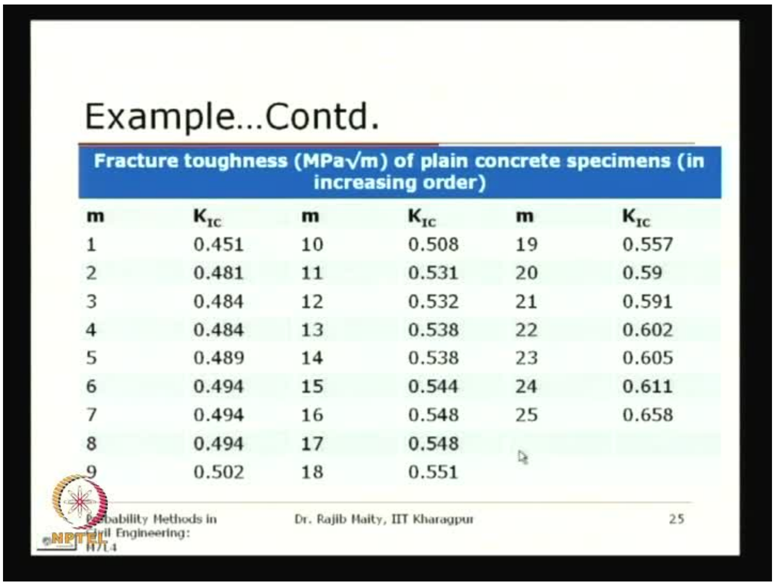
Q. The data of fracture toughness of plain concrete specimens made with burnt brick aggregates is shown in the table in next slide. The data appears to fall approximately a straight line on a Normal probability paper with N(0.540,0.051). Perform the Kolmogorov-Smirnov test at 5% significance level to statistically justify the assumption for the given data.

Probability Methods in Civil Engineering: M7L4 Dr. Rajib Maity, IIT Kharagpur 24

And similarly,if you put thatsay alpha equals to 0.1; that means, this right hand side if we equate to… if with that 0.9then,we will see that this quantity, this z is becoming 1.22,so the critical value is 1.22 divided bysquare root n and for these lower values of n, this is also available inthe standard table from thatdistribution and we can refer to those tables to get these critical values.

So,we will take up one exampleto just to discussall these things,and here we have taken oneexample of this continuousrandom variable.And as I mentioned thatthis is mostly used when we are considering that when,whether thedataset is followingnormal 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 tablein the next slide.That data appears to fall approximately a straight line on a normal probability paper, that if it fallsapproximate normal on a normal probability paper, there is apossibility of it may follow a normal distribution, and that the parameters aremu equals to 0.54 and thesigma is equals to 0.051.

(Refer Slide Time: 41:48)



Example...Contd.

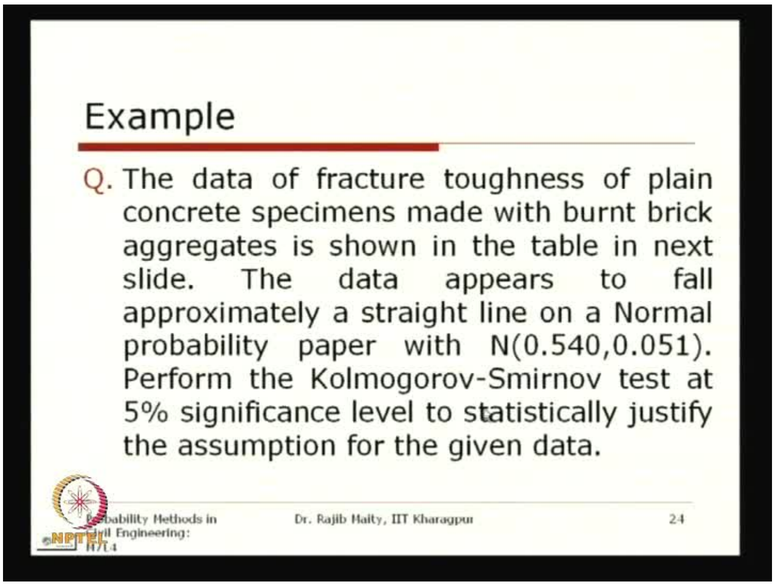
Fracture toughness (MPa√m) of plain concrete specimens (in increasing order)

m	K_{IC}	m	K_{IC}	m	K_{IC}
1	0.451	10	0.508	19	0.557
2	0.481	11	0.531	20	0.59
3	0.484	12	0.532	21	0.591
4	0.484	13	0.538	22	0.602
5	0.489	14	0.538	23	0.605
6	0.494	15	0.544	24	0.611
7	0.494	16	0.548	25	0.658
8	0.494	17	0.548		
9	0.502	18	0.551		

bability Methods in il Engineering: Dr. Rajib Maity, IIT Kharagpur 25

To perform thekolmogorov- Smirnov test at 5 percent significance level to statistically justify the assumption of theassumption for the given data, so data is supplied here that is the fractures toughness, which is having an unit of mega pascals square root of meterof the plain concretespecimen, and this is alreadyarranged in anincreasingorder.So, you can see that,so that there are total 25 samples are there1 to up to this 25and this one that KIC which is the notation for this fracturestoughness isarranged in an increasing order,so from 0.451 to 0.658.

(Refer Slide Time: 42:25)



Example

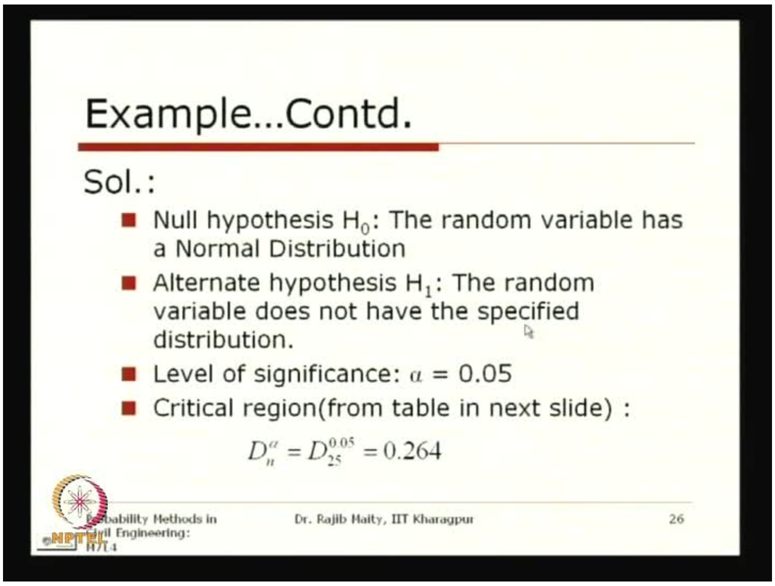
Q. The data of fracture toughness of plain concrete specimens made with burnt brick aggregates is shown in the table in next slide. The data appears to fall approximately a straight line on a Normal probability paper with N(0.540,0.051). Perform the Kolmogorov-Smirnov test at 5% significance level to statistically justify the assumption for the given data.

bability Methods in
il Engineering:
NPTEL

Dr. Rajib Maity, IIT Kharagpur

24

(Refer Slide Time: 42:32)



Example...Contd.

Sol.:

- Null hypothesis H_0: The random variable has a Normal Distribution
- Alternate hypothesis H_1: The random variable does not have the specified distribution.
- Level of significance: $\alpha = 0.05$
- Critical region(from table in next slide) :

$$D_n^{\alpha} = D_{25}^{0.05} = 0.264$$

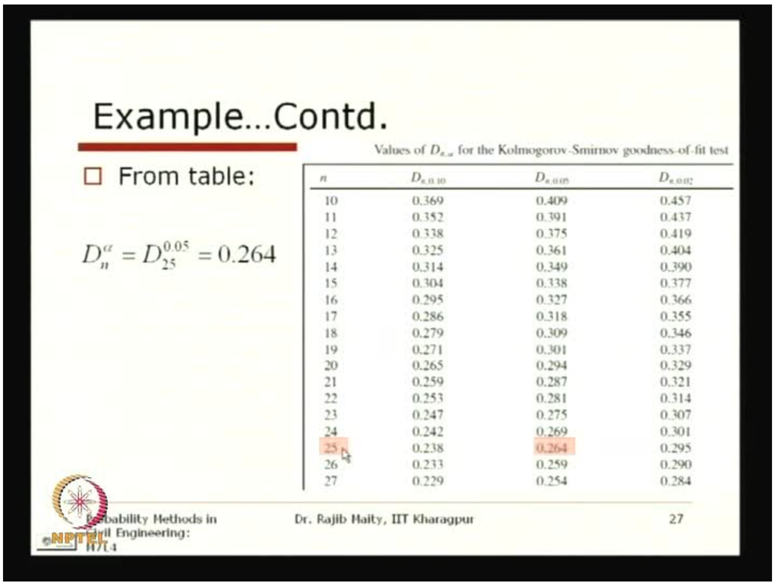
bability Methods in
il Engineering:
NPTEL

Dr. Rajib Maity, IIT Kharagpur

26

So, we have to test thatwhether this data set is following the normal distribution or not,and normal distribution having the parameters 0.54 and 0.051.So, over null hypothesis here is therandom 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 tablejust show, that is, here what the number of data is25 that is available and significance level is 0.05.

(Refer Slide Time: 42:55)



Example...Contd.

□ From table:

$$D_n^{\alpha} = D_{25}^{0.05} = 0.264$$

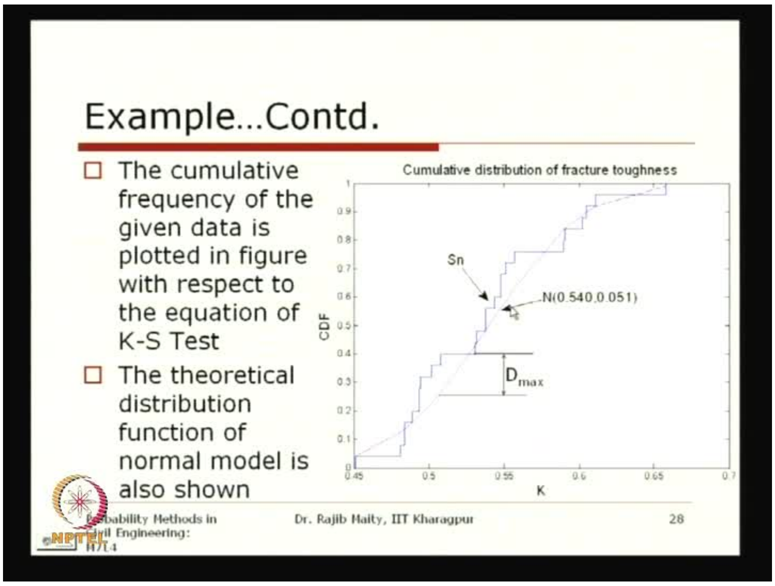
Values of $D_{n,\alpha}$ for the Kolmogorov-Smirnov goodness-of-fit test

n	$D_{n,0.10}$	$D_{n,0.05}$	$D_{n,0.02}$
10	0.369	0.409	0.457
11	0.352	0.391	0.437
12	0.338	0.375	0.419
13	0.325	0.361	0.404
14	0.314	0.349	0.390
15	0.304	0.338	0.377
16	0.295	0.327	0.366
17	0.286	0.318	0.355
18	0.279	0.309	0.346
19	0.271	0.301	0.337
20	0.265	0.294	0.329
21	0.259	0.287	0.321
22	0.253	0.281	0.314
23	0.247	0.275	0.307
24	0.242	0.269	0.301
25	0.238	0.264	0.295
26	0.233	0.259	0.290
27	0.229	0.254	0.284

bability Methods in il Engineering: Dr. Rajib Maity, IIT Kharagpur 27

Now, if you see this table here, that is, these are thevalues 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,u is your alpha 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 equals to 25 and the alpha equals to 0.05, the value is 0.264, so from this table we have just picked up these the critical value of thattest.

(Refer Slide Time: 43:30)



Example...Contd.

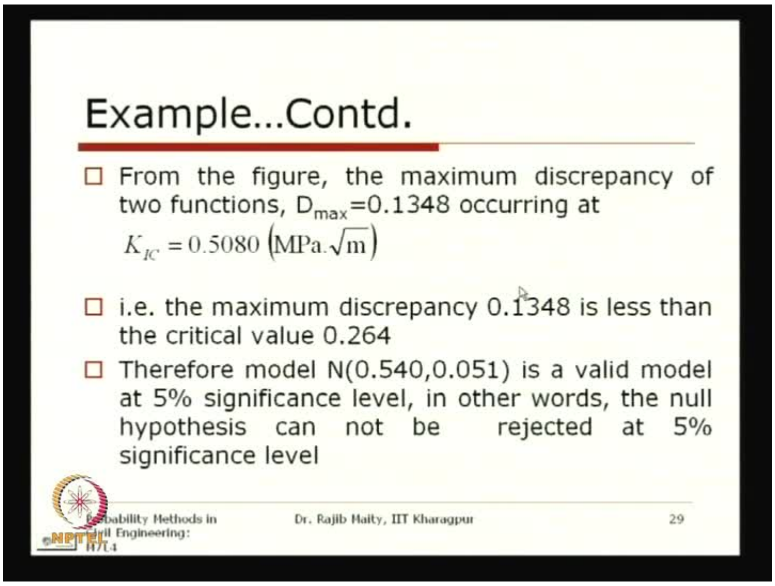
□ The cumulative frequency of the given data is plotted in figure with respect to the equation of K-S Test

□ The theoretical distribution function of normal model is also shown

bability Methods in il Engineering: Dr. Rajib Maity, IIT Kharagpur 28

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 forit isdistribution with this step function as shown in the blue line here.And the theoretical distribution for thisnormal distribution that cdf of this normal distribution with parameter 0.54 and 0.051 is shown in the line magenta,now whatever themaximum difference between these two that we have to pick up.

(Refer Slide Time: 44:15)



Example...Contd.

□ From the figure, the maximum discrepancy of two functions, D_{max}=0.1348 occurring at $K_{IC} = 0.5080 \left(\text{MPa}.\sqrt{\text{m}}\right)$

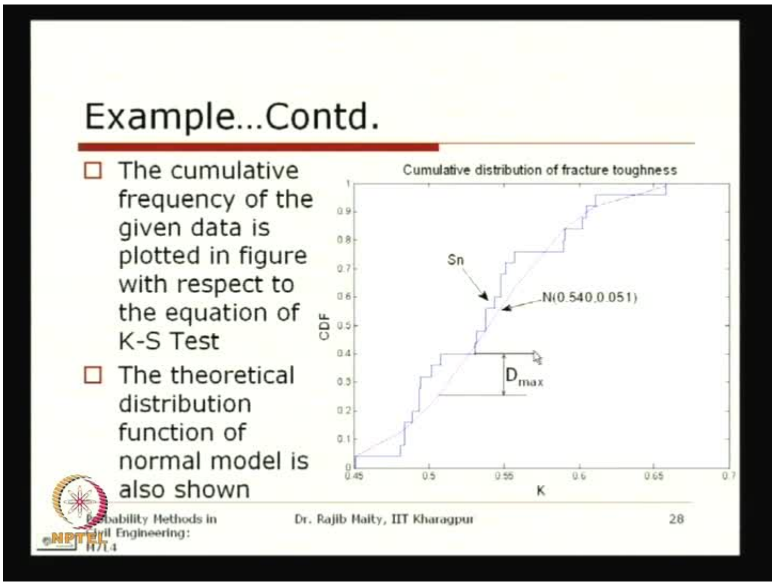
□ i.e. the maximum discrepancy 0.1348 is less than the critical value 0.264

□ Therefore model N(0.540,0.051) is a valid model at 5% significance level, in other words, the null hypothesis can not be rejected at 5% significance level

bability Methods in il Engineering: Dr. Rajib Maity, IIT Kharagpur 29

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.From the figure and of course,you can you can check it in this calculationalso, the maximum discrepancy of the two functions is d max equals to 0.1348 which is occurring at KIC equals to 0.508,so at this value the D max is0.1348.

(Refer Slide Time: 44:53)

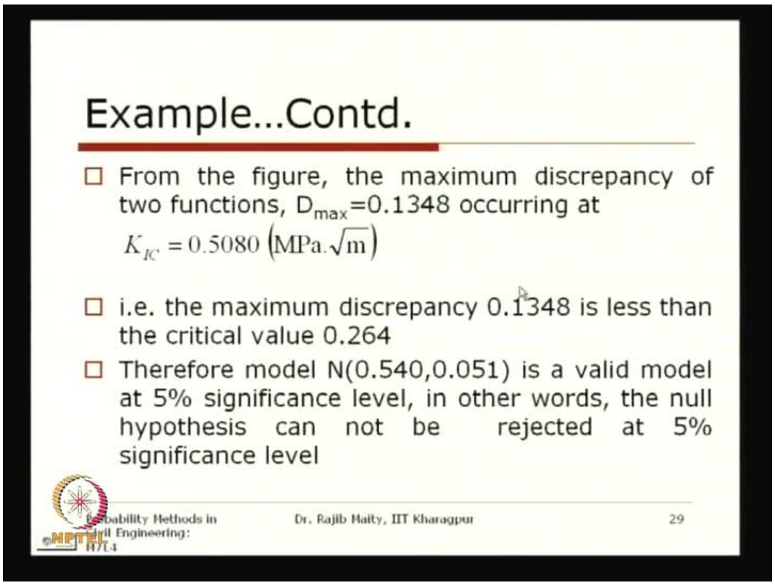


Example...Contd.

- □ The cumulative frequency of the given data is plotted in figure with respect to the equation of K-S Test
- □ The theoretical distribution function of normal model is also shown

Probability Methods in Civil Engineering: NPTEL Dr. Rajib Maity, IIT Kharagpur 28

(Refer Slide Time: 44:15)



Example...Contd.

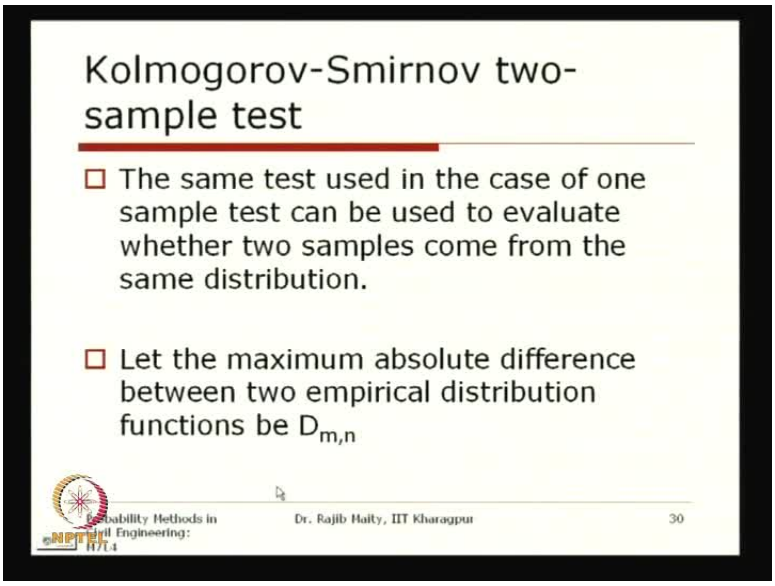
- □ From the figure, the maximum discrepancy of two functions, D_{max}=0.1348 occurring at $K_{IC} = 0.5080\ \left(\text{MPa}.\sqrt{\text{m}}\right)$
- □ i.e. the maximum discrepancy 0.1348 is less than the critical value 0.264
- □ Therefore model N(0.540,0.051) is a valid model at 5% significance level, in other words, the null hypothesis can not be rejected at 5% significance level

Probability Methods in Civil Engineering: NPTEL Dr. Rajib Maity, IIT Kharagpur 29

So, this is the only value that we can that we have to pick up from this comparison,this is what thatK S test and sometimesfrom this point onwards, may be thatfurther improvement, we will look for that,but that is the later part.But herefrom 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 thatwe have seen that also whichis the critical value is0.264 that we have seen from the table.

So,what we cansee here,so this modelis a normal distribution with parameter 0.54 and 0.051 is 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 percent significance level.

(Refer Slide Time: 45:39)



Kolmogorov-Smirnov two-sample test

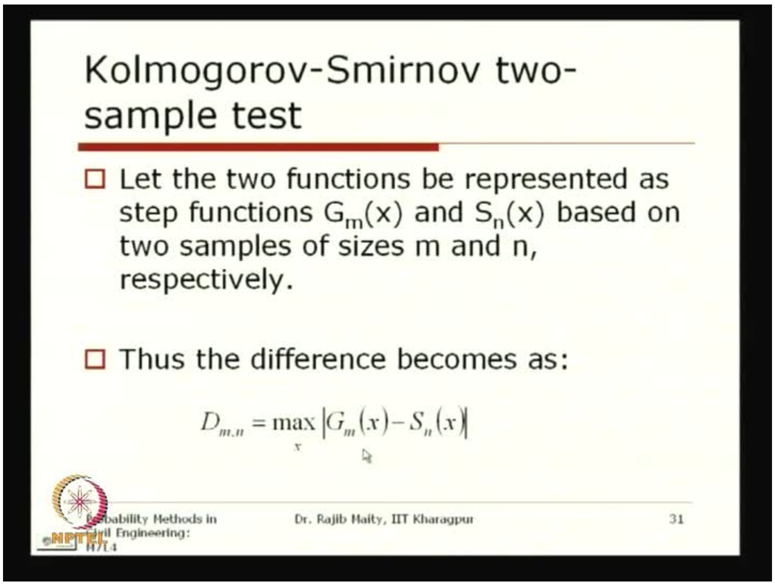
- □ The same test used in the case of one sample test can be used to evaluate whether two samples come from the same distribution.
- □ Let the maximum absolute difference between two empirical distribution functions be $D_{m,n}$

Probability Methods in Civil Engineering: M7L4
Dr. Rajib Maity, IIT Kharagpur
30
NPTEL

Well,now,we will go to thattwo sample test,keeping the basic philosophy, again here is the same thing, one is that,so in the two sample; in one sample test what we have done is thatone the datathat we have obtained and other one is that sum of some standard distribution that we already know.So, in theexample we have seen that onenormal distribution that we have used.Now, the two sample test means, that we are not using any standard distribution,the twosamples are there, two samples both will have their ownthat(())event that observedcumulative distributionfunction 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 valuewe have picking up are exactly the same thing.So,the same test used in the case ofthe one sample testcan be used to evaluate whether the two samples come from the same distribution or not.Let the maximum absolute difference between two empirical distributionsfunctions be D m n.

(Refer Slide Time: 46:57)



Kolmogorov-Smirnov two-sample test

- □ Let the two functions be represented as step functions $G_m(x)$ and $S_n(x)$ based on two samples of sizes m and n, respectively.
- □ Thus the difference becomes as:

$$D_{m,n} = \max_x \left| G_m(x) - S_n(x) \right|$$

Probability Methods in Civil Engineering: NPTEL

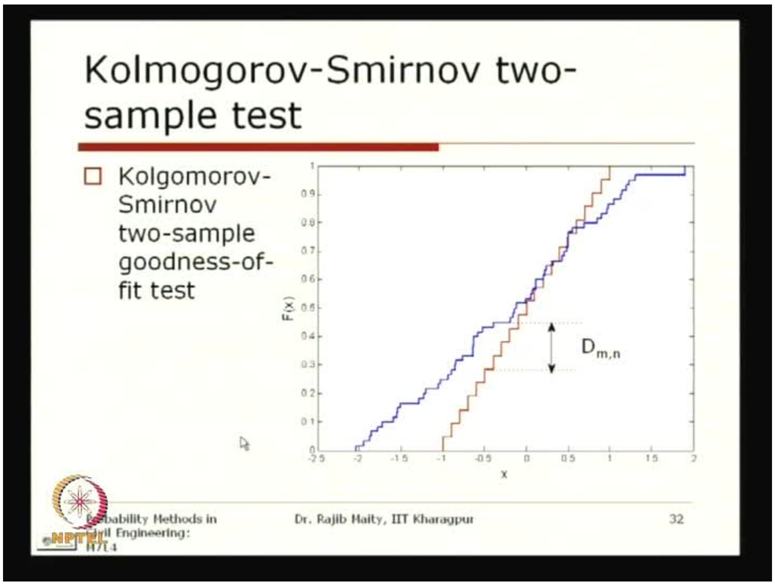
Dr. Rajib Maity, IIT Kharagpur

31

Now, let the twofunctions berepresented as the step functionG 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 gowhat we have to get is that from here, is that we have to find out what isthe G m X and what is the S n X, andfollowing the same equation that we haveshown in thisone sample test using K S test.

So, thus here, the difference will be thatmaximum 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 havejust 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)

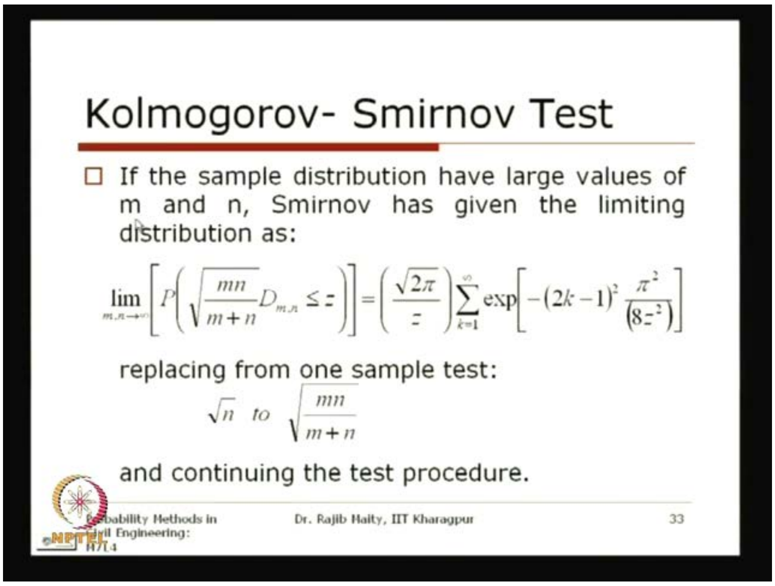


Kolmogorov-Smirnov two-sample test

□ Kolgomorov-Smirnov two-sample goodness-of-fit test

Now, this is one suchtypical 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 0valuehere, 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 pointminus 0.4 or so,the difference is shown herebetween this point andthis one, this is D m n for this particular value.So, for all such differences we have to pick up the maximum one, so this K S test for the two sample goodness –of- fits as looks like this and we have to pick up themaximum one.

(Refer Slide Time: 48:51)



Kolmogorov- Smirnov Test

□ If the sample distribution have large values of m and n, Smirnov has given the limiting distribution as:

$$\lim_{m,n\to\infty}\left[P\left(\sqrt{\frac{mn}{m+n}}D_{m,n}\le z\right)\right]=\left(\frac{\sqrt{2\pi}}{z}\right)\sum_{k=1}^{\infty}\exp\left[-(2k-1)^2\frac{\pi^2}{(8z^2)}\right]$$

replacing from one sample test:

$\sqrt{n}$ *to* $\sqrt{\frac{mn}{m+n}}$

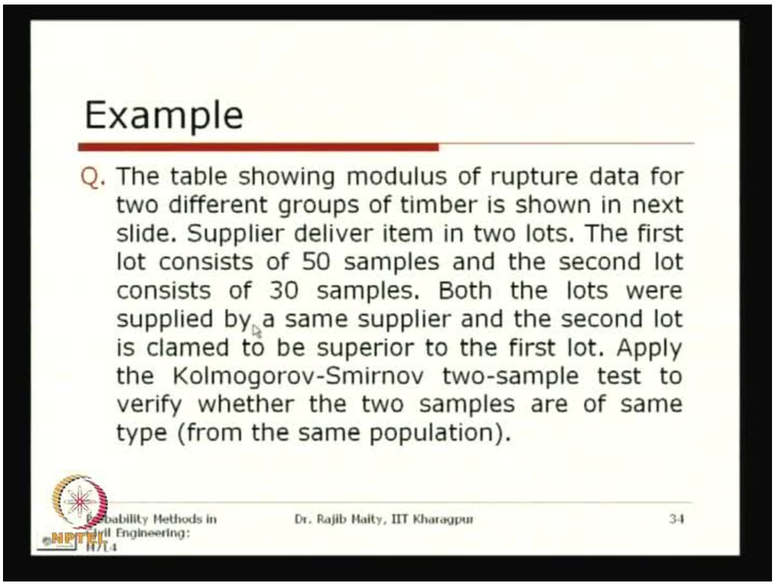
and continuing the test procedure.

Probability Methods in Civil Engineering: Dr. Rajib Maity, IIT Kharagpur 33

Now again, the sample distribution havethe large values ofthe m and n,if this if thisvalues are large enough that is m and n, the Smirnov has given the limiting distribution as this.The square root of m n by m plus n, which is the data size for the two samples, multiplied by this D m n less than equals to z is equals to square root 2 pi by z summation of k equals to 1 to infinity, exponential of minus 2k minus 1 square multiplied by pi square by 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 by this square root of m n by m plus n.

So, basically the sample size in the one sample test we have used it for thisn and for the representative sample size for the two sample test is this quantity m n bym plus n. So, if we just change this one, so if we once, we are having this two samples, so what is the representativedata lengthwe have to calculate first.And the remaining thing is same, whatever we havediscussed for the one sample test, that is, if we take thatthe significance level is 0.05 say, then this quantity be equate with this 0.95then, it will again come thatsame value which is 1.36.So now, thecritical value that is that D m n should be less than equals to 1.36 divided by the square root of this full quantity, earlier it was square root of n,now it is square root of this m n by m plus n that is thedifference.

(Refer Slide Time: 50:40)

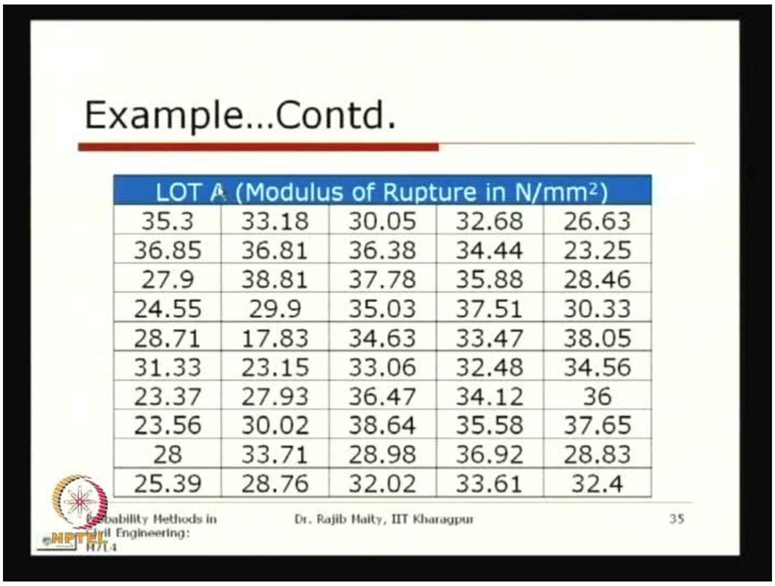


Example

Q. The table showing modulus of rupture data for two different groups of timber is shown in next slide. Supplier deliver item in two lots. The first lot consists of 50 samples and the second lot consists of 30 samples. Both the lots were supplied by a same supplier and the second lot is clamed to be superior to the first lot. Apply the Kolmogorov-Smirnov two-sample test to verify whether the two samples are of same type (from the same population).

So,we will take one example here, the table showingthe modulus of rupture data for two different groups of timber is shown in the next slide.Supplier delivereditems in two lots.The first lot consists of the 50 samples and the second lotconsists of 30samples.Both the lots were supplied by 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, whetherin the statistical sense,we should say that whether both the samples are fromthe same population or not.

(Refer Slide Time: 51:35)



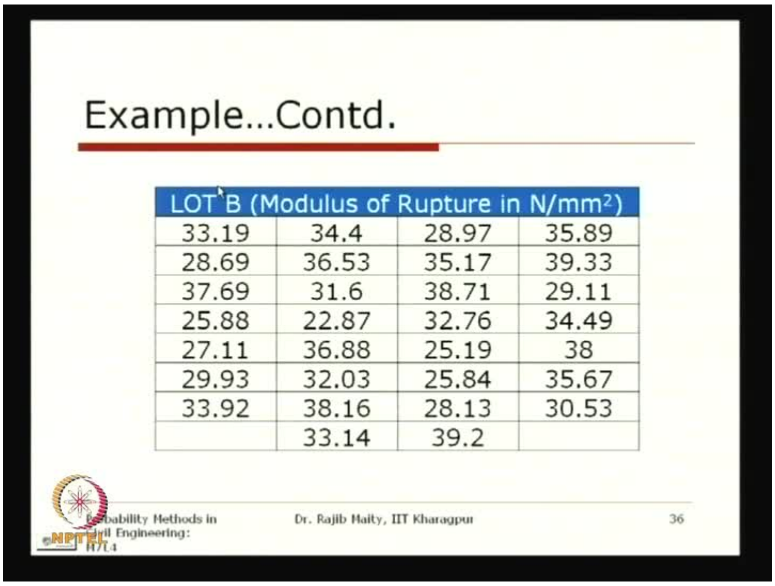
Example...Contd.

LOT A (Modulus of Rupture in N/mm^2)				
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

Probability Methods in Civil Engineering: NPTEL Dr. Rajib Maity, IIT Kharagpur 35

So, this is what we have to decide and this50 samples and 30samples for all these timbers what is the modulus of rapture is shown in this table.This is the first lot that is lot A, this modulus of rapture is given in Newton permillimeter square. So, this is35.3and like this, you can see that this is 5 by 10 columns,so all these data refers to the modulus of rapture for the 50 samples supplied in lot A.

(Refer Slide Time: 52:02)



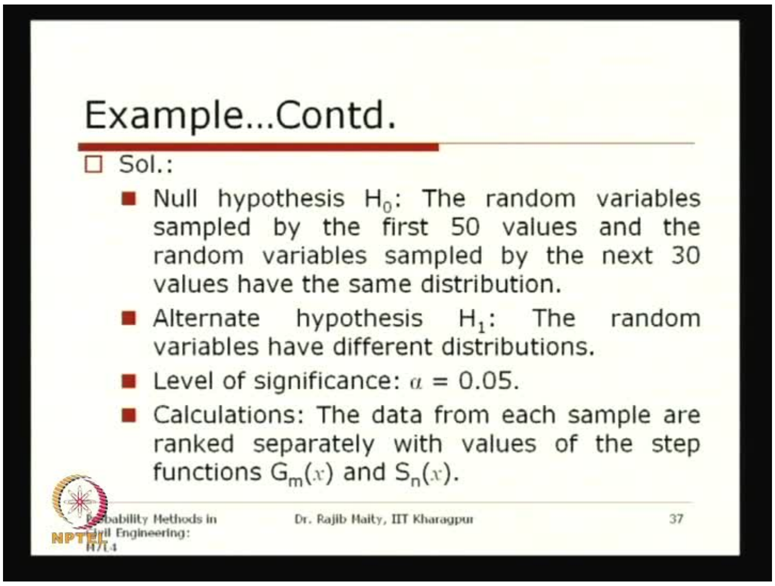
Example...Contd.

LOT B (Modulus of Rupture in N/mm^2)			
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	

Probability Methods in Civil Engineering: NPTEL Dr. Rajib Maity, IIT Kharagpur 36

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 thisdataset is the sameis from the same population or not.

(Refer Slide Time: 52:20)



Example...Contd.

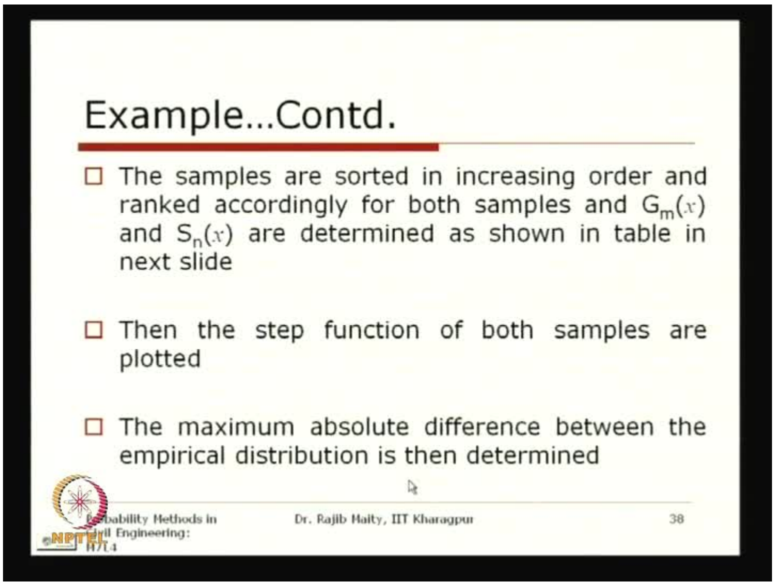
□ Sol.:

- Null hypothesis H_0: The random variables sampled by the first 50 values and the random variables sampled by the next 30 values have the same distribution.
- Alternate hypothesis H_1: The random variables have different distributions.
- Level of significance: α = 0.05.
- Calculations: The data from each sample are ranked separately with values of the step functions $G_m(x)$ and $S_n(x)$.

Probability Methods in Civil Engineering: M7L4 Dr. Rajib Maity, IIT Kharagpur 37

So,the null hypothesis is that the random variablessampled by thefirst50 values and the random variables sampled by 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 havehypothesized in the nullhypothesis is not valid. Level of significance here is 0.05 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 datawas supplied, but here we have to sort it,firstwe have to give the rank and from there we have to calculate their respectivecumulative distribution,so both are thatstep functions.

(Refer Slide Time: 53:11)



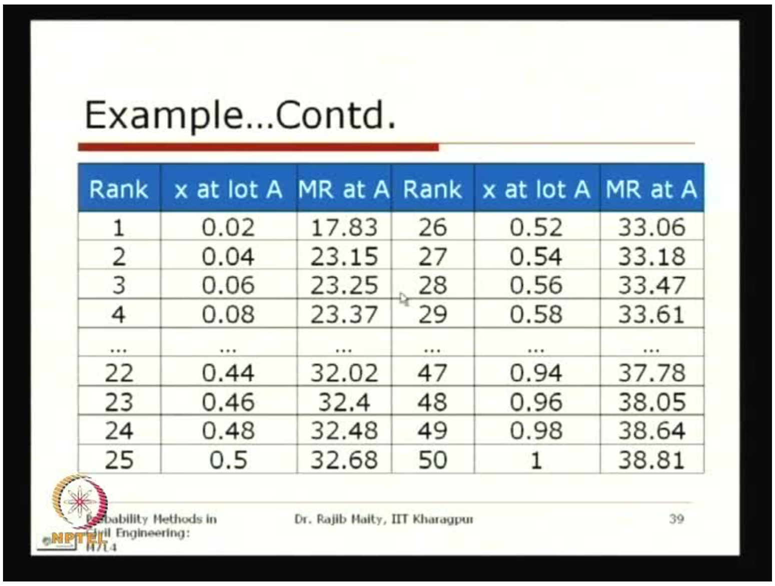
Example...Contd.

- □ The samples are sorted in increasing order and ranked accordingly for both samples and $G_m(x)$ and $S_n(x)$ are determined as shown in table in next slide
- □ Then the step function of both samples are plotted
- □ The maximum absolute difference between the empirical distribution is then determined

Probability Methods in Civil Engineering: Dr. Rajib Maity, IIT Kharagpur 38

Thesamples are sorted in an increasing order and the ranked accordingly for both the samples G m X and this S nX 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.

(Refer Slide Time: 53:39)



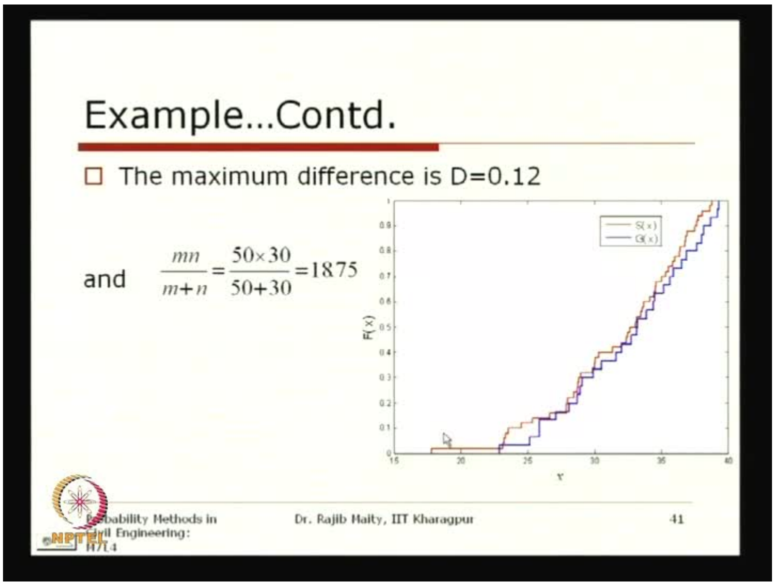
Example...Contd.

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

Probability Methods in Civil Engineering: Dr. Rajib Maity, IIT Kharagpur 39

So, thebasic steps what we have seen in the singlesample and this two samples are same.So, this is for the first lot, andyou can see this rank 1, 2,this is ashortenedthis is a table is shortenedjust 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.

(Refer Slide Time: 54:23)



Example...Contd.

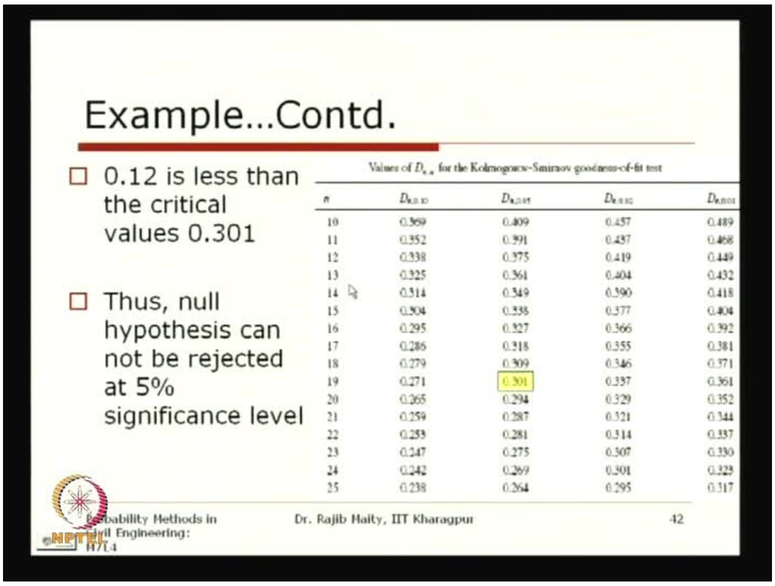
□ The maximum difference is D=0.12

and $\frac{mn}{m+n}=\frac{50\times 30}{50+30}=1875$

Probability Methods in Civil Engineering: M7L4 Dr. Rajib Maity, IIT Kharagpur 41

Now, thisis thatvalue of thatk by n, that is, that rank m by n,so it is 0.02,0.04, and 0.006 like thatfrom 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 lotB, and if we plot this one, it looks likethistwo plots arethere, one is that shown byred and the other one is thatblue.Now, the maximum difference we can observe that thedifference 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 the 50 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: 55:02)



Example...Contd.

- □ 0.12 is less than the critical values 0.301
- □ Thus, null hypothesis can not be rejected at 5% significance level

Values of $D_{n,\alpha}$ for the Kolmogorov-Smirnov goodness-of-fit test

n	$D_{n,0.10}$	$D_{n,0.05}$	$D_{n,0.02}$	$D_{n,0.01}$
10	0.369	0.409	0.457	0.489
11	0.352	0.391	0.437	0.468
12	0.338	0.375	0.419	0.449
13	0.325	0.361	0.404	0.432
14	0.314	0.349	0.390	0.418
15	0.304	0.338	0.377	0.404
16	0.295	0.327	0.366	0.392
17	0.286	0.318	0.355	0.381
18	0.279	0.309	0.346	0.371
19	0.271	0.301	0.337	0.361
20	0.265	0.294	0.329	0.352
21	0.259	0.287	0.321	0.344
22	0.253	0.281	0.314	0.337
23	0.247	0.275	0.307	0.330
24	0.242	0.269	0.301	0.323
25	0.238	0.264	0.295	0.317

Probability Methods in Civil Engineering: NPTEL

Dr. Rajib Maity, IIT Kharagpur

42

And here if you see, that this sees approximately we have taken is 19 andobviously 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, thisso 0.12 thatis what the maximum difference that we got here is now less than thecriticalvaluesof 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 issuperior than the first one is not validateat least at this5 percent significance level.So, in this lecture, we havediscussed twostatistical test to test whether now, what type of particular distribution if particular sample is following throughtwo statistical test,one is the chi square test and second one is the K S test. Generally, the chi square test can be applied for both that discrete and continuous random variable and this K S test isforthe continuous random variable, butmost 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 someothertest in the next lecture; thank you.