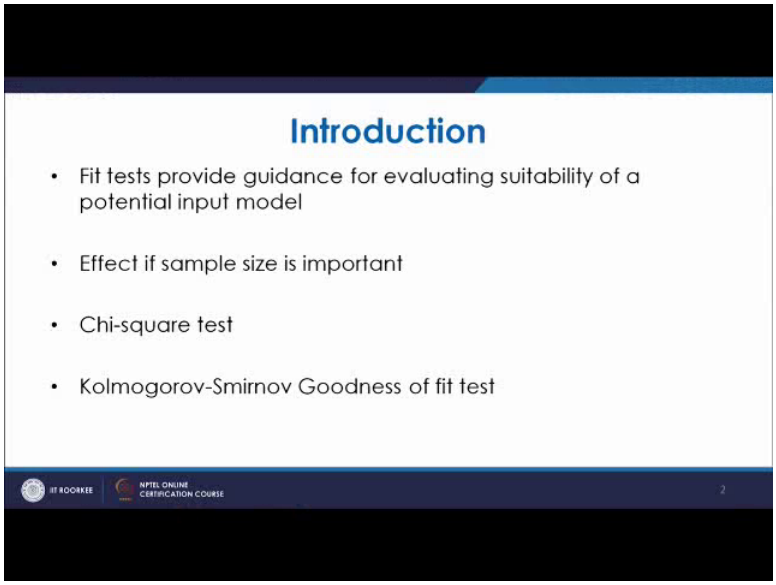
Modeling & Simulation of Discrete Event Systems
Dr. Pradeep K Jha
Department of Mechanical and Industrial Engineering
Indian Institute of Technology, Roorkee

Lecture - 23
Intput Modeling: Goodness-of-Fit Tests & Assessing Sample Dependence

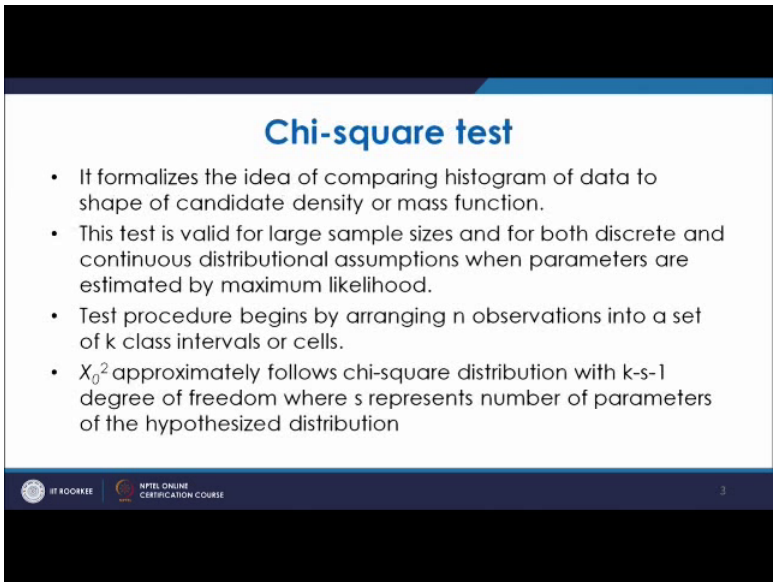
Welcome to the lecture on goodness of Fit Tests and Assessing Sample Dependence. So, we have been introduced about the fit tests and as we know that fit tests are providing the guidance for evaluating suitability of a potential input model.

(Refer Slide Time: 00:37)



So, in this goodness of fit tests normally we test for the uniformity. And we have seen that the sample size is important. Based on the sample size we have different type of test which are used to see that the distribution or the data which is there where it can be said to be forming a uniformity having uniformity or not. So, you have some hypothesis test well hypothesis is there, whether it is rejected or it is accepted that depends upon the calculation of the parameters. So, we have 2 types of tests chi square test and you have that is chi square test and then you have Kolmogorov Simonov test. So, these are the 2 test which we have earlier discussed in brief.

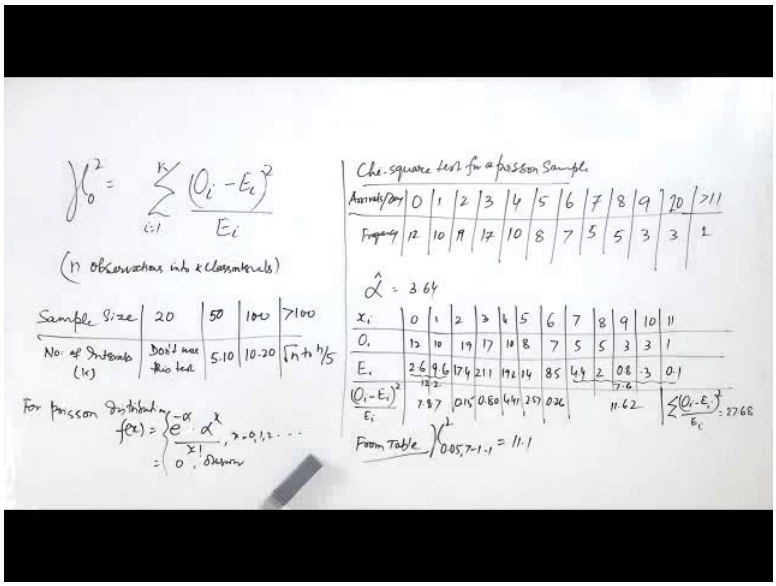
(Refer Slide Time: 01:38)



So, what we have seen that the chi square test is basically it is the it is formalising the idea of comparing histogram of data to shape of the candidate density or mass function. So, from the histogram you can have the you can see that what is the shape of the density or mass function, how it looks like you will have the data based on that as we have seen you draw the histograms and from there you can have a feel about the shape of that.

Now, this test is valid only for large sample sizes; so for both discrete as well as the continuous distribution assumptions, and when the parameters are estimated by maximum likelihood. So, normally we go for when we whenever we have the large number of sizes sample size and also we need to have the calculation of parameters, in those cases we use this chi square test. So, because that affects the degree of freedom calculation which basically we have to know for getting that table value. So, what we do is that basically you begin by arranging the n observations into set of k class intervals and then the chi 0 square. So, that basically is approximately following the chi square distribution with k minus s minus 1 degree of freedom; so this s where it is the representing the number of parameters of the hypothesized distribution. So, we get this value first the parameter values and this parameter value as we know what we see.

(Refer Slide Time: 03:42)



We have already seen that this will be basically summation O i minus E i of square by E i. So, it will be from i equal to 1 to k. So, we have k class intervals. So, you have n observations into k class intervals.

So once, what we do here is you have the class interval, you arrange the data in those intervals this is the observed frequency O i and this is the expected frequency if you going to fit for the uniform distribution. So, in that case this will be i p i basically. So, that way you have the i p i value, and then ultimately you are doing this you are dividing it with E i and then you take the summation of that. So, you get this chi 0 square and then that value you are gets basically comparing from the table with this k minus s minus 1 that much degree of freedom. So, this with that degree of freedom you calculate the value from the table. Now this value should be less than the value which we calculate from the table then we say that this hypothesis is not rejected I mean the null hypothesis. So, we should not reject that we can say that this data is forming a uniformity or it is the uniformity uniformly distributed.

Now, in this case you have also to see that your sample size normally you there is certain you know guidelines and sample size and number of intervals. So, how many number of intervals should be there; so for that depending upon the sample size. So, if the sample size is 20 you do not use this test.

Normally your sample size should be larger only when the sample size is large you sue this test. If the sample size is 50 the interval number of intervals can be 5 to 10. Similarly if it is 100 you go from 10 to 20 and if it is more than 100 then you go from under root n to n upon 5. So, this is what normally is accepted and we see that the value of chi 0 square, it has to be less than the critical value then the hypothesis assumes that random variable confirms to distribution assumption with parameters given by the parameter estimate.

Now, the thing is that size of E i t is also said that it should be normally ranging from 3 to 5. And when this value is less in that case you can combine it from the adjacent cells either above or below you should combine them. So, so that you get ultimately you have less number of data intervals. Now if the distribution is discrete then in that case every data should be the class. So, every data will be a class interval and in that case as we have seen that if the frequency is less in some cases in that case you can combine the adjacent classes. So, that you get adequate number of frequencies.

So, this is normally practiced in the case this chi square test now let us see you have you are doing for a Poisson sample. So, if we do the chi square test for a Poisson sample we have the. So, we are dealing with the same problem which we have discussed the number of arrivals and mean number of arrivals on every day. So, you had the data arrivals per day in a frequency and frequency. So, you have arrival per 0, 1, 2, 3, 4, 5, 6, 7, 8,9,10 and then more than 11. So, these are the arrivals per day data and this data was seen it was the frequency was calculated as 12, 10, 9, 17 then 10, 8, 7, 5 and further 5, 3, 3 and 1.

So, as we see that this type of data is given to you and you have to use the test to see the uniformity. Now in this case we have already computed alpha estimator as 3.64. So, we saw that you alpha estimator is nothing but x bar and for that you are computing the product and edited and added all together, this is 364 divided by 100. So, you are getting 3 0.64 now in that case now you are having this data. So, you will be having x i and in that case you will O i E i and further summation of them. So, you will have x i and O i E i and ultimately you have compute O i minus E i square by E i.

So, you have the values as we see you have the values as 0 1 2 3 4 5 6 7 8 9 10 and 11. So, once you have these values O i as we see you have again 12, 10, 19, 17, 10, 8, 7, 5, 5,

3, 3 1 and expected frequency it will be i p i. So, in this case you are expected frequency now how to get these expected frequencies. Now since it is a Poisson distribution. So, in that case what you do is you compute the probability value. Now for this Poisson distribution as we know that probability distribution function tells f x will be alpha raise to the power x and then e raise to the power minus alpha divided by x factorial.

So, what we see is for exponential distribution. So, for Poisson distribution, we get f x as probability mass function will be e raise to the power minus alpha into alpha raise to the power x upon x factorial. For x equal to 0 1 2 and all and it is 0 otherwise. So, that we know we know this that this is the probability mass function for the Poisson distribution and we know this alpha estimator 3.64. So, when the x value is 0 in that case you get here e raise to the power minus alpha.

So, this way you are getting the values. So, in this case if we get the probability value this value will come out to be 2.6, now what happens you are getting this is nothing but P 0. So, that will be from here and that value is coming as 0.026. So, 0.026 will be multiplied with hundred and that you get as 2.6 similarly for one. So, for one again you are putting it here as one. So, for one what is the probability that is the expected probability.

So, this will be coming as 9.6. So, for in this way you are going to calculate this probability value for all these values by placing the value of x here and in this way and multiplying it again with 100. So, you are getting. So, once you do that calculation you are getting the values for 2 is 17.4, for 3 it is 221.1, for 4 it is 19.2, for 5 it is 14 and then it is 8.5 for 6, 4.4 you have 2 0.8, 0.3 and 0.1.

This is the percent of probability this is the 2.6 this is 1.6. So, out of 100 if you have this has a frequency which is observed and this is the expected one because the process is Poisson. So, you get this, now based on that you will calculate this parameter. So, this will be 12 minus 2.6 a square divided by 2.6. So, this way your value now the thing is that in this case as we see that we have to see that how we can see you have to get may be tried not to see so many intervals. So we will try to combine some of them. So, if we combine here, it will be something like 12 12.2. So, it will be 12.2.

Similarly on this side it is getting low. So, you can combine all them; so 6.6 7.4 7.7 and 7.8. So, it will be coming as 7 point. So, we are going further to combine all of them. So,

it will be coming as 4.4 plus 2. So, this is 6.4 7.2 7.5 and 7.6. So, this is coming as 7.6. So, as we have seen that we can combine certain intervals, we have to combine we have not to take very less or very more and then further you are finding the values.

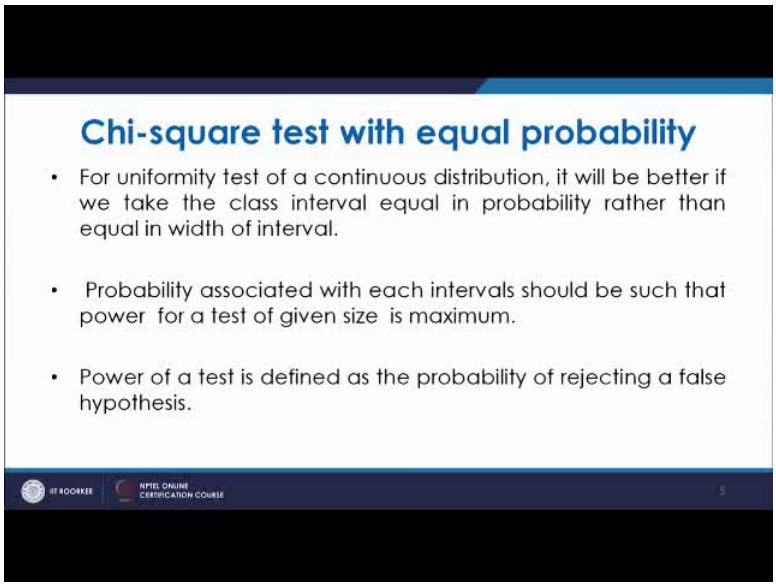
Now, in this case it will be 22 and this is 12.2; so 22 minus 12.2 square. So, that will be something like 9 point. So, the 2.9 plus 9.6 it is coming out 2.6 and. So, it is 12.2. So, further you can do the calculation and that will be closed to 7.87. So, it will be 7.87 further it will be 0.15, then you have 0.80, you have 4.41, you have 2.57, you have 0.26 and then for all of them together you are getting the value as 11.62. So, what we do is here we take this as 22 and this is 24.2.

Similarly here you have 10 plus 13 16 and 17 and this is 11 point and this is 7.6. So, we take this. So, once we get that the summation of O i minus E i square upon E i and this comes out to be 27.68. Now this chi 0 square this value is this is the value which is coming out to be 27.68, now we have to see it from the table. Now from the table you have to get. So, from table that table was shown in the earlier lecture.

Now, from table you have to see the value and if we take that 0.05 level of significance as 0.05 and degree of freedom. Now degree of freedom is k minus s minus 1. So, you have intervals of k n you can see here 1 2 3 4 5 6 and 7. So, you have 7 intervals k minus s. So, you have the parameter which is estimated that is 1, you have calculated that alpha estimator. So, it is o1and minus 1. So, that is 5. So, from the table you have to calculate this chi square value for 0.05 to and degree of freedom as 5 and that value is coming out to be 11.1. What we see is the value which we calculate is chi 0 square that is 27.68, now this is basically larger than this value whereas, for passing the uniformity test it must be lower than this theatrical value or from the table value.

So, since it is not less it is not passing that test. So, we say that it is not uniform ally or it is not distributed does not pass that uniformity test in the case of this data. Now the thing is next case is that many a times we have to see that when we take the intervals, the it should be for that interval which should have equal probability. We should do the test for those intervals where there is equal probability.

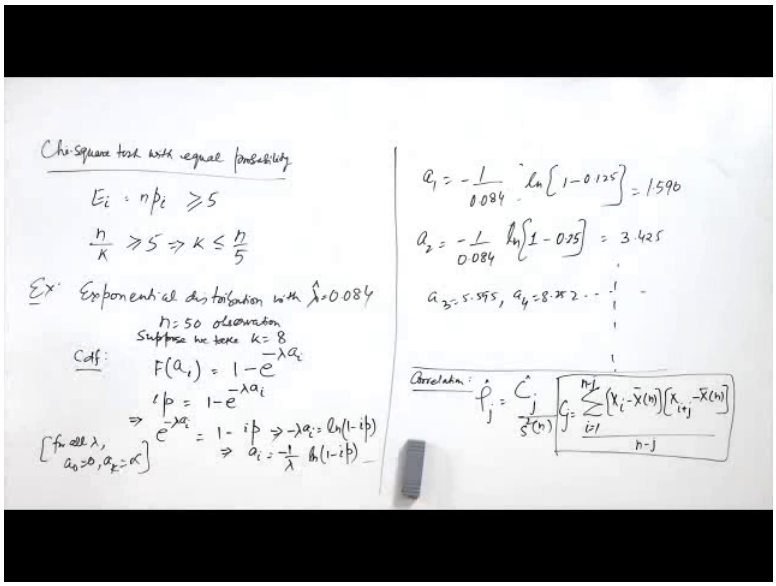
(Refer Slide Time: 19:51)



Now, the thing is that, that is why the test of a continuous distribution in that case it will be better if we take the class interval equal in probability rather than equal in width. So, we take equal in width whereas, we should take equal in probability. So, we should take that those intervals which is expected to have the equal probability. Now in that case you must know the point where to take the. So, interval for that interval you must have those n points known to us. So, probability associated with each interval should be such that the power of the test of given size is maximum means the power of test means the probability of rejecting a false hypothesis. So, so this way we I mean accepting a good thesis good hypotheses is maximum in that case. So, that is power. So, you should take the interval in such a manner that you must have the maximum probability of accepting that particular hypothesis or so.

Now, how can we proceed for such cases where you have the equal probability?

(Refer Slide Time: 21:07)



So, we will see that. So, we will see the chi square test. So, in that case as we have seen that in such cases of the test with equal probability, now in those cases as we have seen that the expected frequency E i it is normally equal to i into p i, and that should be if we take the maximum number of class intervals in that case it should be more than equal to 5. So, if we take that case. So, if we take.

So, E i is basically E i which is given as i p i and if we take the maximum number of intervals in that case it should be more than equal to 5 now the. So, it will be written as n by k it should be written as more than equal to 5. So, n by k how it come now actually E i has to be n p i. So, E i equal to n p i. So, and p i is equal to 1 by k. So, that is why n into 1 by k. So, n by k is more than equal to 5 it means k has to be less than equal 5 by n, n by 5.

So, k should be what should be the class interval size. So, k class interval. So, k should be less than equal to n by 5. So, if you have if you know n it should be. So, n by 5 and less than that the k should be number of class interval should be less than that. Now suppose you have the 50 number of data that is n is 50, in that case k should be less than equal to 10 or so. Now, suppose we take example of an exponential distribution with lambda as 0.084 and you have n as 50 observations. So, how to find the interval now in this case would not know we have the data from that data we got the x bar.

So, you get the sample mean, from there you get lambda estimator that will be 1 by x bar. So, for exponential we know that we get that lambda estimator as 1 by x bar. Now in that case what should be the interval length what should be the upper and lower limit of this interval every interval. So, that the probability is normally equal. So, as we know that you have in this case what we do is we first of all find the c d f. So, we find the c d f as f of. So, if we the a i is representing the n point of i-th interval. So, suppose a one will be the n point of the first interval, a 2 will be the n point of the second interval. So, it will be starting from a 0 to a 1 a 1 to a 2 a 2 to a 3 like that.

Now, if we take if we take this. So, in this case in the case of 50 observations k has to be less than equal to 50 by 5 that is 10 suppose we take 8 intervals suppose we take k as 8. So, in that case the E i value will be going as 100 by 8 means everything will be you know 0.1 to 5 probability will be there 12.5 percent. Now again here f a i we will. So, if we write the cumulative distribution function F a i will be written as 1 minus e raise to the power minus lambda a i. So, we know that the cumulative distribution function is f x equal to 1 minus e raise to the power minus lambda x. So, that is how F a i will be one minus e a to the power minus lambda a i and this a i should be basically we have to get the expression for it and this a i it will be basically i into p. So, this will be i p equal to 1 minus e raise to the power minus lambda a i.

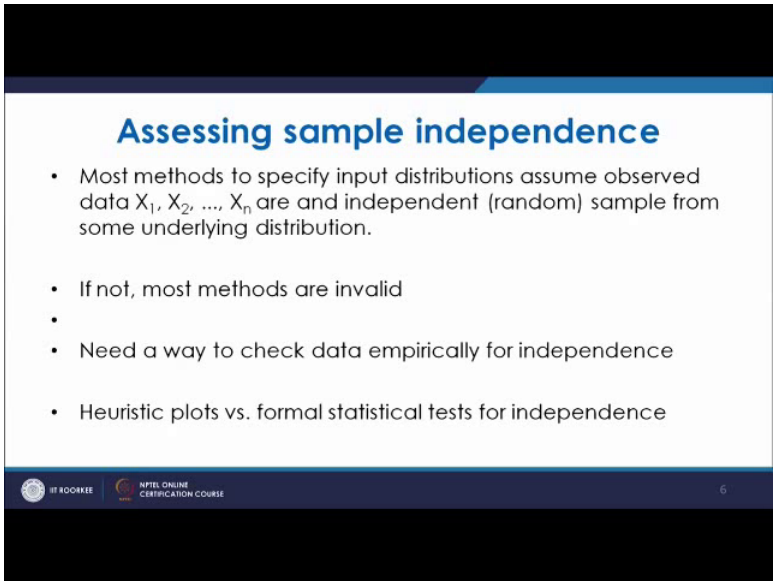
Now, every time we go you will have these values. So, in every class. So, i p will be equal to 1 minus e raise to the power minus lambda I, now we have to solve this and once we solve it we get the values as a equal to minus of. So, we get we can have the expression more simplified. So, we will have e raise to the power minus lambda a i will b equal to 1 minus i p. So, in that case we get lambda a i as l n of 1 minus i p and a i will be minus of 1 by lambda into l n of 1 minus i p. Now as we move from interval to interval this a i has to be calculated. So, that the probability is equal in every interval. So, in that case as we see a 0 will be always 0 and a 8 if you have 8 intervals a 8 will always be the infinity.

So, if you take the first interval I will be varying from 1 to 8. So, when we go from for the first interval the lower limit is always 0 and the upper limit of the last interval will always be infinity. So, what we see is that for all lambda a 0 is 0 and a k is infinity. So, that data is the condition is there with us now what we see that since we know the lambda and every time we go in the first interval. Now this is because we have 8

intervals. So, what the first interval this this value will be 0.125, so in that case 1 by 8. So, p will be 1 by 8 and i is 1; so for the first one.

So, a 1 will be minus of o1 by lambda, lambda is given to you and lambda is given as 0.084. So, it is 0.084 and then you d it l n of 1 minus 0.125. So, this value is basically computed and this you get as equal to 1.590, it means you are going to have the first data first interval as from 0 to 1.590, the second interval for that will a 2 and a 2 will be minus 1 by 0.084 and in that case it will be l n of 1 minus i is 2 now; so 1 minus 2 into 1 by 8. p is 1 by 8 you have 8 intervals. So, it will be 2 by 8 that is 0.25. So, l n of 0.75 and then you do that. So, this way you get the value as 3.425. So, this way you get the different kind of you know the ranges a 1 a 2 a 3 if you compute further you get a 3 as 5.595, a 4 as 8.252 and so on. So, for such typical problem you can calculate the a 1 a 2 a 3 like that.

(Refer Slide Time: 30:09)

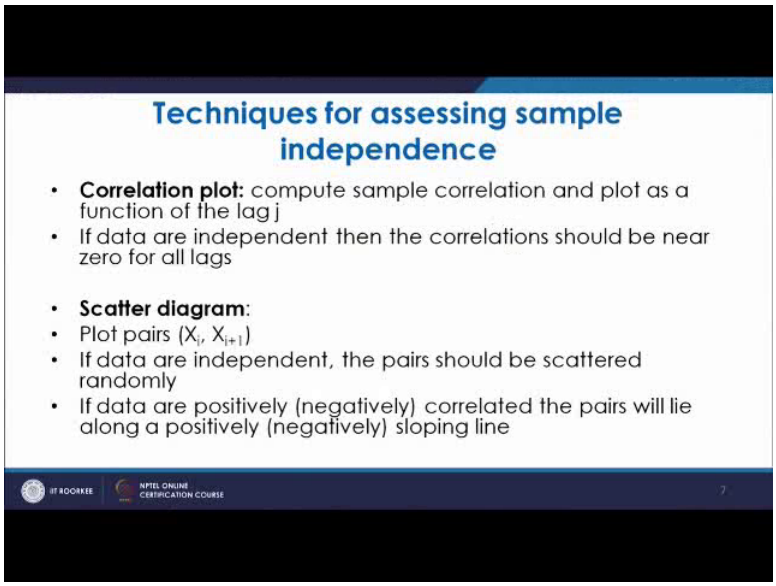


Next is the assessing sample independence, we have discussed about the uniformity check in the case of sample how to check the sample independence. So, most of the methods statistical methods assume that the data which we have been taken they are independent, and if they are not then these methods which we used to calculate the estimators or so, they are invalid.

So, we are assuming that the data are is independent now you need to say you need to check the dependence or independence of the data. So, for that you have the plots and

these plots basically tell us and heuristic plots are there which tell us that whether the data can be said to be independent or they are dependent they are correlated, and you have even formal statistical test like run test or so. So we will talk about certain some number of plots what are the plots.

(Refer Slide Time: 31:08)

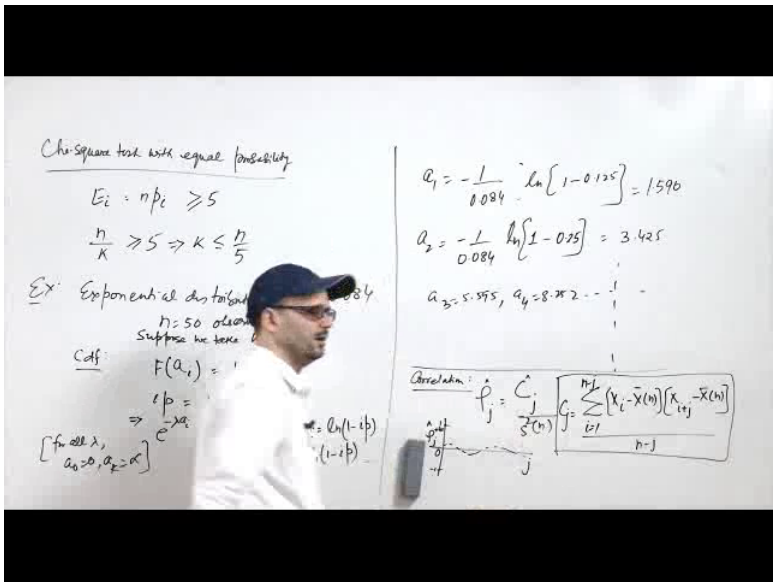


So, one of them is correlation plot. So, by in the correlation plot what we do is we find the correlation. Correlation is nothing but we have if we have the random number x i and x i plus j you have the time varying the data is there, in that case you define the correlation rho i j and that will be basically depending upon the co variance between the 2 and then it is further divided by the variance. So, that is. So, that variable is rho i j. So, this rho j or this correlation of I mean from i to j. So, that that is computed and once you compute that that will is plotted. So, once you plot that it should be basically near to 0. If this value is near to 0 they are said to be independent and if they are plus one or minus one closed to that then they are said that they are not independent.

So, what we do is we find in that case when we go for the correlation, we find this correlation as. So, we find rho j as c j estimator divided by s square n. So, this is the covariance we will discuss later, how this covariance is computed c j square is compute this is c j square c j is computed as summation of i equal to 1 to and minus j and then it is x i minus x bar n and similarly x i plus j minus x bar n and divided by n minus j. So, if we start from i and go up to i plus j, in that case what is the correlation between them for

that you get the co variance and then divided by s square n. So, this is basically c j this this is actually the value of c j. So, c j is computed using this formula and once you get the c j then divided by s square n. So, you get the correlation and once you get the correlation then you plot it. So, you are plotting and plotting this correlation value.

(Refer Slide Time: 33:49)



So, you are plotting with j. So, you will be plotting like you will have j on this x axis and you have rho j on the y axis. So, if this is 0 this is a positive one, and this is the negative one. So, in this if the data is like this its fine if they can be said to be you know that they are not related they are independent otherwise they will be said to be correlated. So, that it should be independent.

Similarly, another diagram is scatter diagram in that scatter diagram what we do is we plot x i and x i plus 1. So, we have the x i and we have x i plus 1 both are plotted x i will be here on the x axis and x i plus 1 will be on the y axis, and if it is completely distributed uniformly then we say that that it is scattered randomly then it is we say that it is having no correlation. And if they are having some slope positive or negative in that case we say that they are basically positively or negatively correlated.

So, this way this scatter test scatter diagram is also used for assessing the sample independence.

Thank you very much.