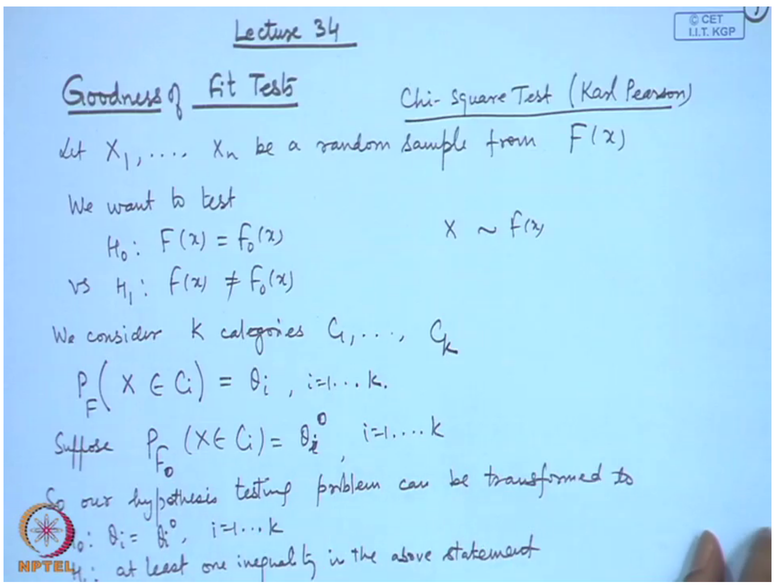
**Statistical Methods for Scientists and Engineers**
**Prof. Somesh Kumar**
**Department of Mathematics**
**Indian Institute of Technology - Kharagpur**

**Lecture - 34**
**Nonparametric Methods - VII**

Now, we move on to the next problem that is called the goodness of fit test, in this problem we want to test whether the distribution is a particular distribution or not. So basically this is the problem of modelling of distributions.

**(Refer Slide Time: 00:45)**

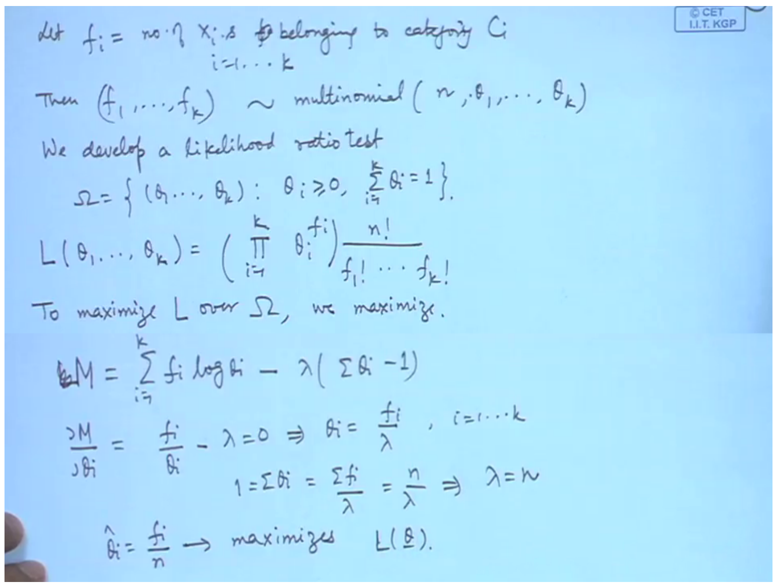



So roughly speaking we have a sample, so let us consider say X1, X2, XN this is a random sample from say if F x alright, and we want to test whether F x=F0 x or not. If you want to test F x=F0 x against F x=F1 x where F1 is different from F0, then we have the most powerful test using the Neyman–Pearson lemma. However, it is not that, here we want to null hypothesis we are specifying completely, but alternatively we are not able to specify.

Therefore, we cannot apply the Neyman–Pearson lemma here, so what we do? We consider say we classify the data into K categories say C1, C2, Ck, and we calculate probability of X belonging to Ci where X is of course F x, what is the probability of X belonging to Ci let us denoted by theta i okay, when the distribution is F for i= 1 to k. Now suppose this probability that X belonging to Ci that is say theta i0, when F is taken to be F0.

So actually we make use of this fact that means under the null hypothesis the probability of each category is specified, so we actually frame it as a multinomial testing problem, so our hypothesis testing problem can be transformed to H0 theta i=theta i0 i= 1 to k, against at least 1 inequality in the above statement. So you can see that in this Kolmogorov this Chi square test for goodness of fit, I am going to discuss it is one of the oldest nonparametric test it was developed by Karl Pearson.

**(Refer Slide Time: 04:32)**



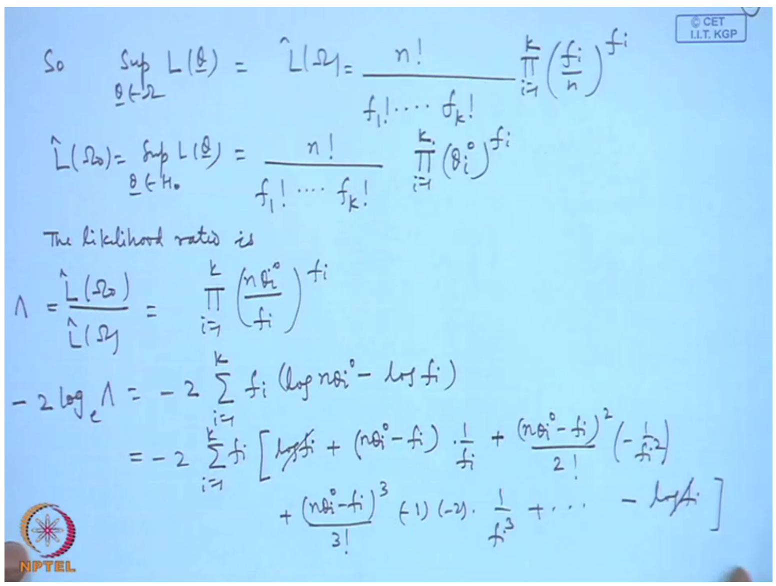

In this test actually cleverly the problem of full testing has been transformed to checking K categories, so we can actually consider let fi be= number of Xi's which are belonging to category Ci for i= 1 to k, then we can say that this f1, f2, fk this is having a multinomial distribution, total number of observations is n and the probabilities of categories C1, C2, Ck they are theta 1, theta 2, theta k respectively.

The simplest things that we can do is we consider a likelihood ratio test for this problem, so for likelihood test we know that we would like the likelihood function, we develop a likelihood ratio test, so here the full parametric space that is theta 1, theta 2, theta k, where theta i's are >=0 and sigma theta i that is =1, theta i to the power of fi n factorial/f1 factorial and so on fk factorial i=1 to k. So basically this part is constant, so maximization problem is reduced to this part only.

So to maximize L over the parametric space omega, we maximize fi so we take a log here, log of theta i sigma i= 1 to k, subject to the condition that summation theta i=1, so introduce Lagrange's multiplier, let us called this term as =Lagrange's multiplier let we call it LM here or we can give some other notations say M here. So that means we differentiate with respect to each theta i, then I will get fi/theta i-lambda that is = 0 that gives me theta i=fi/lambda for i=1 to k.

This gives me the value of n, because I can apply the condition here sigma theta i= sigma fi/lambda so this=1, this=n/lambda this means lambda=n, so theta I head=fi/n here. So this is the maximizing L theta.

**(Refer Slide Time: 08:14)**



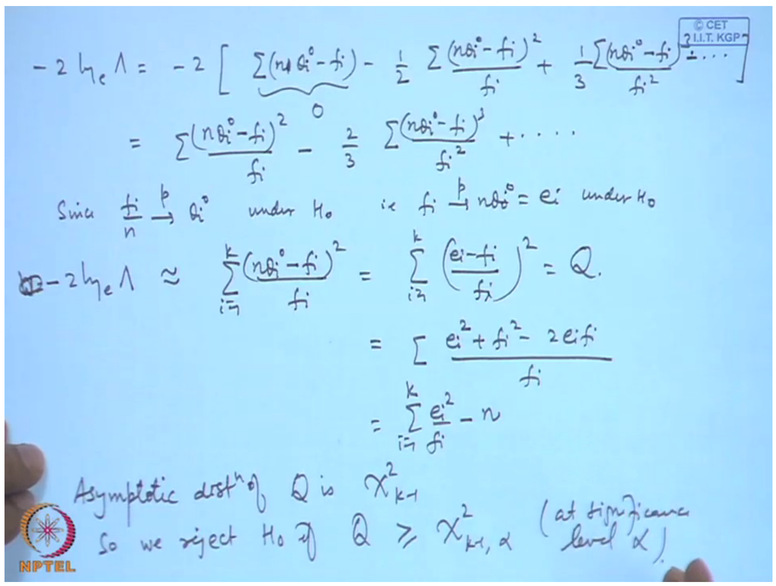

So what is the maximum value then, so the supremum value of L theta for theta belonging to omega=L head omega=n factorial/f1 factorial and so on fk factorial product of fi/n to the power fi i=1 to k, because I have substituted the values of theta I is a fi/n here, and L head omega 0 that means supremum of L theta for theta belonging to H0 that=n factorial/ f1 factorial, f2 factorial, fk factorial product of theta i0 to the power fi i=1 to k.

So if we consider the likelihood ratio that is L of omega 0/L head omega, so this term will get cancelled out you will left with product n theta i0/fi to the power fi, let me call it say lambda then -2 log of lambda that=, so this what we are doing to develop the distribution here, that is -2 fi log of n theta i0-log of fi for i=1to k. Now this term log of n theta i0 I expand around fi, so that=-

twice sigma i=1 to k fi log of fi+n theta i0-fi then the derivative of this that is becoming 1 by that at fi so it is simply becoming i/fi.

Then second term will give me n theta i0-fi square/2 factorial second derivative will give me -1/fi+ n theta i0-fi cube i/fi cube, this will be square here and so on - log of fi that is this term here So we can now adjust the terms here, this term will get cancelled out, this term will give me n theta i-fi this fi will get cancelled out, and in other terms here fi will come here.

**(Refer Slide Time: 12:00)**




So let us write this there is-2 log of lambda that=-2 sigma-n theta i0-fi-1/2 sigma n theta i0-fi square/fi+ 1/3 n theta i0-fi cube and of course summation will be there and so on, this term is simply so sigma theta i=1 so sigma theta i0 is also=1, this is n then sigma fi is an so this first term will become 0, second term is giving me sigma n theta i0-fi square/fi-2/3 sigma n theta i0-fi cube/fi square and so on.

Now fi/n converges to theta i0 in probability under H0 that is you can say fi converges to n theta i0 that we denote by ei under H0 in probability. So therefore, I can say that - 2 log of lambda is asymptotically= sigma n theta i0-fi square/fi for i= 1 to k, this is written as ei-fi/fi square i=1 to k that= Q here, so this is converging so therefore, the higher order terms we neglect here and we are writing only this, fi's are called observed frequencies, ei's are called expected frequency of the ith class.

And we have an alternative formula for this, this is also see if I expand it this will becoming ei square +fi square-2 ei fi/fi that= sigma of ei square/fi, +sigma fi-twice sigma ei fi both are n here so this will becoming -1 i=1 to k. So this is the test statistic which is coming from the likelihood ratio, in the likelihood ratio we know that we accept the null hypothesis, if L head omega 0/L head omega is so we reject the null hypothesis if the denominator is log.

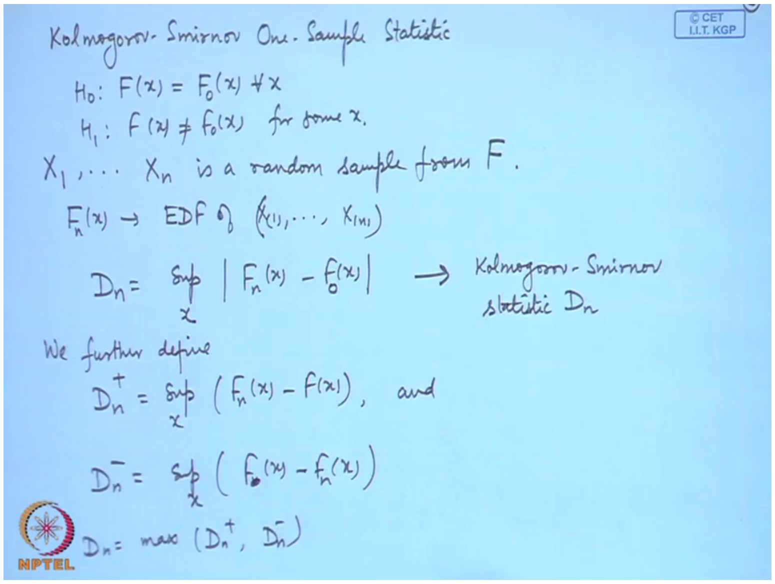
So basically we have taken -2, so lambda is should be small for closer to H0, and for rejection lambda should be small, so this -2 I have taken - 2 times log of this so outcome will be reverse that means for large values of -2 log lambda will be rejecting. Another interpretation you can make out from here this is actually the difference between the observed and expected frequencies square, so if the 2 distribution is not f0, then there will be large difference here that means these differences will propagate and this term will become large.

So basically this gives an indication the value of -2 log lambda whether the null hypothesis is true or not, now that gives a rough indication, but to get a real picture of this we will need the distribution of that, for that we see that this is multinomial. Therefore, asymptotic distributions of this simply will become the some of the chi-square, because we considered 2 then binomial convergence to normal, so here it is converging to k-1 dimensional thing.

And therefore when we are taking some of the square it will convert to chi square on k-1 degrees of freedom, so asymptotic distribution of this quantity let me call it W or Q we have called it that is chi square k-1, so we reject H0 if Q is >= 5 chi square k-1 alpha at significance level alpha. This test is widely used in all the applications for modelling of the statistical distributions and it is extremely useful.

However, since it is asymptotic certain assumptions are there, for example the expected frequency of each cell must be >5 for a good approximation, if that is not so then this test is not very good.

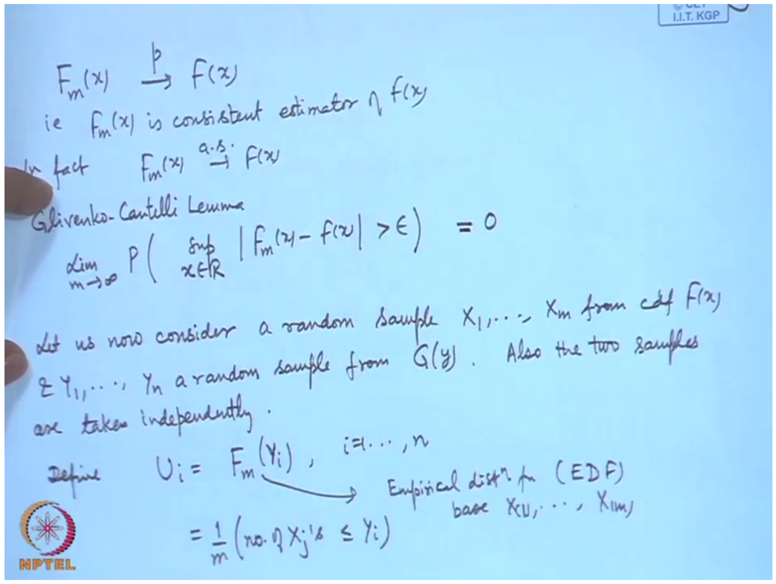
**(Refer Slide Time: 18:10)**




Now another test for the goodness of fit was developed by Kolmogorov and Smirnov, so this is called Kolmogorov-Smirnov one sample statistic, as before we are writing down our hypothesis testing problem as F x=F0 x for all x or F x !=F0 x. We have at our disposal a random sample from this population, we define the sample distribution function that is Fn x that is empirical distribution function of X1, X2, Xn that is the ordered statistics from this.

We define the maximum absolute difference between the empirical distribution function and the assumed the distribution function, so here actually you take F0. So what is the idea for this? The idea for this is the result about the empirical distribution function, which we gave earlier that was that it is strongly consistent not only strongly consistent.

**(Refer Slide Time: 19:58)**

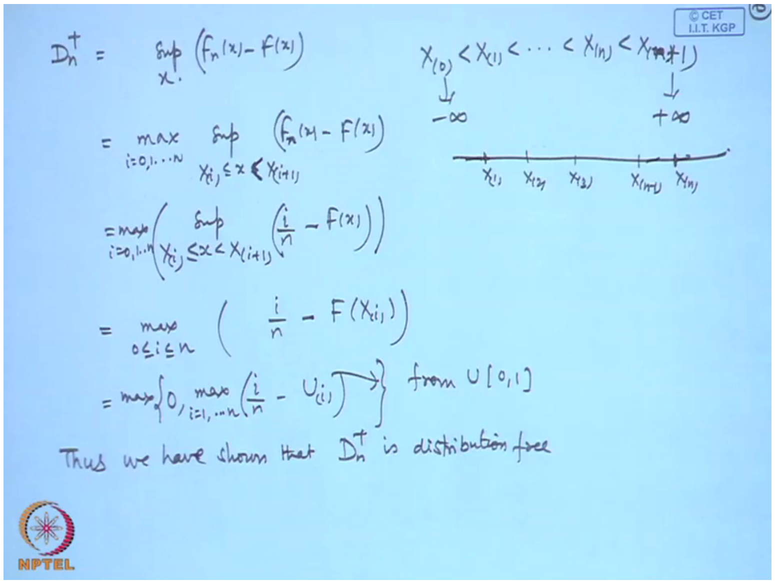



We had actually proved that the limit of the probability supremum Fm x-F x >epsilon actually goes to 0, so because of this, this is a very good indicator of the actual discrepancy between the assumed model and the sample that means based on the sample actually we calculating the empirical distribution function, so if there is too much discrepancy then this statement or this value will be large. So based on this idea this Kolmogorov-Smirnov they have defined the statistic called the Dn.

Now suddenly as in the previous chi-square goodness-of-fit, we need to discuss the distribution of Dn, if we are looking at Fn then certainly we knew the distribution, but since we are considering the maximum here so then this problem become slightly different. So we further define, so this one is actually the Kolmogorov-Smirnov statistic we call it Dn okay, so we further define 2 quantities called Dn+ that is= supremum of Fn x-F x.

And Dn- that is= supremum of F x-Fn x, so here I will put reverse that is Dn is actually = the maximum of Dn+ and Dn- that means actually I am taking the maximum positive difference and maximum negative difference. Now we will try to analyze this distribution of Dn+ and Dn- separately.

**(Refer Slide Time: 22:21)**




So Dn+ that is the supremum of Fn x-F x overall x, we have ordered X1<=X2<=Xn, and I take X0 and X infinity also or Xn +1, so this is I am taking to -infinity, this I am taking to +infinity. So we can then express it as maximum of supremum value between Fn x-F x, let me put this into one side, this is for i= 0, 1, 2, n that means I am considering that supremum value in the intervals, so this is X1 that is -infinity to X1, then between X1 to X2, then between X2 and X3, then between Xn-1 and Xn and then Xn to infinity.

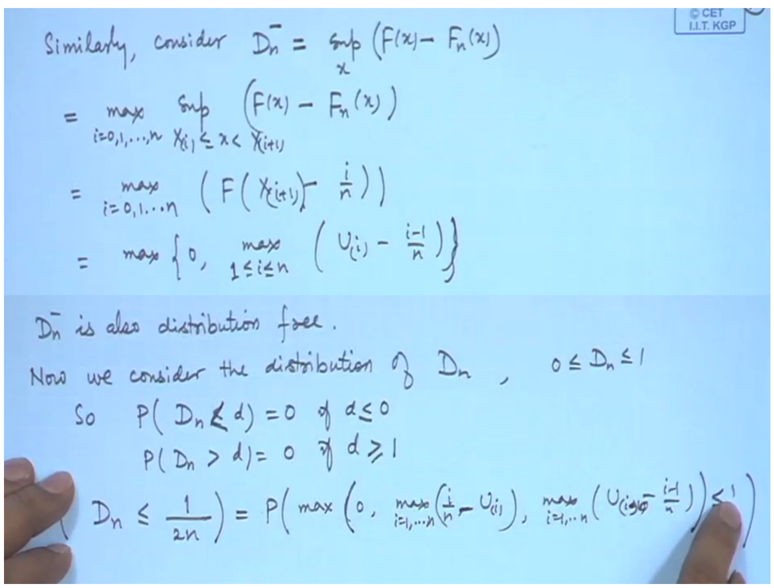
So I have divided this problem into looking at the difference into each of the intervals, but the advantage of this approach is that actually the value of the empirical distribution function in this interval is i/n, so basically I am looking at that what is the difference of F x from i/n, when X is in the interval Xi to Xi+1 and this we are doing over all i's. Now capital F is an increasing function because it is a cdf, so when I look at the supremum here.

Now it is supremum over X, so this will becomes a fixed quantity, so this becomes actually the minimum value of F the minimum value of F in this interval is attained at Xi, so this is becoming =maximum of i/n-F of Xi i=0, 1, 2, n, and now this F Xi's are actually Ui's that we have already seen, so this is i/n-Ui's, where Ui's are the order statistics from the uniform 0,1 okay this is from uniform 0, 1 here, and we are looking at the maximum for i=1 to n, and corresponding to 0 then this is actually 0.

So this is maximum of this and this, now this is very interesting here, I started with some sample here okay, now based on that sample I have considered the difference between Fn x-F x that is a empirical distribution function-F x, but this quantity if you look at this quantity has become free from the original distribution, because this is nothing but from the uniform 0, 1. Thus we have shown that Dn+ is distribution free.

As I mentioned earlier in the beginning of this particular section on nonparametric methods that here we develop those methods which are free from the distribution original distribution assumption, so that means that whatever the distribution originally it does not affect our distribution that means distribution assumption is not required except of course we consider continuity etc. here.

**(Refer Slide Time: 27:01)**




Now in a similar way I can consider Dn- here, so let us consider Dn- now what is Dn-, Dn-= supremum of F x-Fn x that means I am taking the negative value here, so this= maximum of supremum F x-Fn x for Xi<=x<Xi+1 i=0, 1, 2, n that is= maximum of i= 0, 1, 2, n that is = F of Xi+1-i/n, that is= maximum of 0, maximum of 1<=i<=n Ui- i-1, so I have shifted by 1 here because I am taking 1 to n, so I can in place of i+1 I can write i then this becomes i-1.

So once again as in Dn+ here also you see so Dn- is also distribution free, now first thing is that we are able to derive the form of Dn+ and Dn- in terms of the order statistics from uniform 0,1.

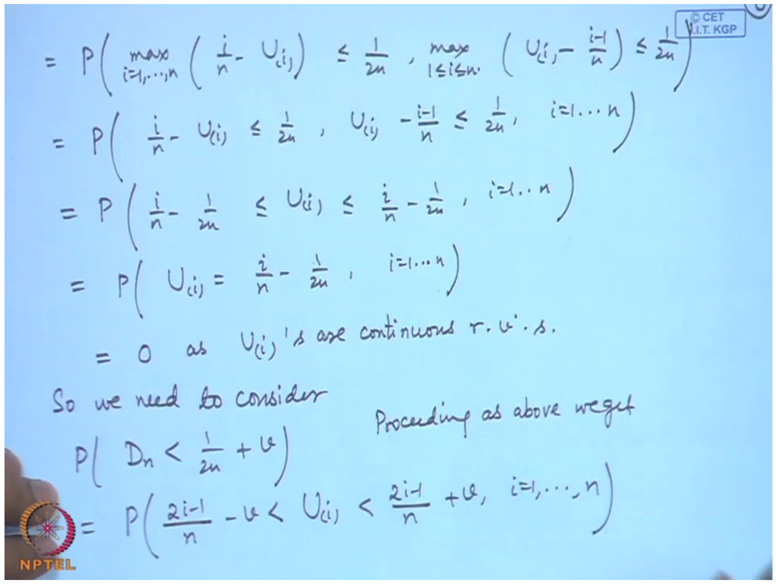
The distribution of U bracketed i is known that we have derived as the beta distribution, now here the form that is coming out is the maximum that means when we are considering several dependent distributions or dependently distributed random variables.

Then what is the distribution of the maximum of that, and then again maximum of the to2, let me express it here. So now we consider the distribution of Dn, now one thing that you note these values are between 0 to 1, and what are these values here i-1/n these values are also between 0 to 1, so these all values actually always lie between 0 to 1. If you look at this also this is i/n these values lie between 0 to 1, so this values will also lie between 0 to 1 only okay.

So the entire thing is that Dn lies between 0 to 1, so that means when we are considering distribution of this <= say some d then it is =0 if d is <0, and probability of Dn >say d that is =0 if d is>=1 okay. Now let us consider Dn<d between 0 to 1, so we put a particular form here, why that particular form? It will clear when we derive the expression here. Let us consider say probability of Dn<=1/2n.

So there is a reason that why I am considering 1/2n, the reason is that if you look at these values here they are of the form 1/n etc. in each interval if I look at this, so the differences will be of this nature and let me firstly derive this here. So Dn is nothing but probability of maximum of 0 maximum of i/n-Ui where i is from 1 to n, and maximum of Ui-i-1/n and again here i=1 to n, so what we are saying is that this is <= 1/2n, so maximum of this and this, and this <= 1/2n.

**(Refer Slide Time: 32:43)**



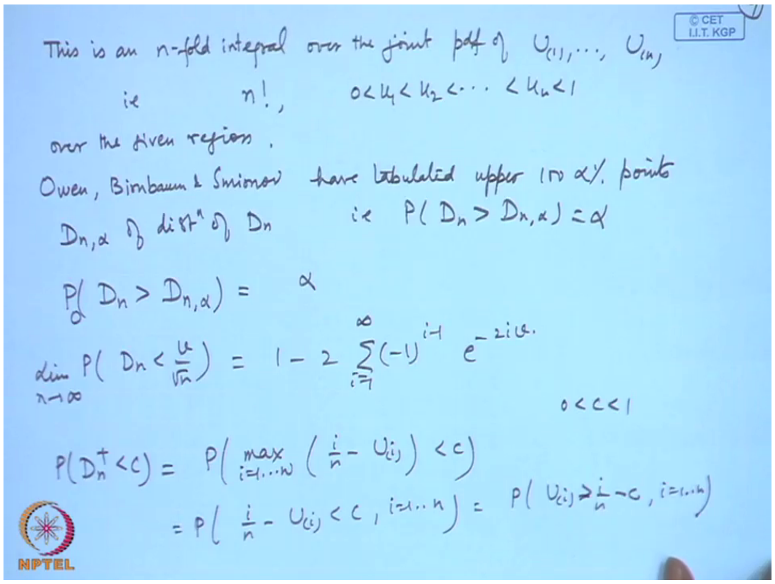

We splitted, so 0 is <=1/2n is always true, so this probability then can be expressed as probability of maximum of i/n-Ui <=1/2n for i=1 to n, and maximum of Ui-i-1/n<= 1/2n this is 1<=i<= n, this is = probability of now these are already ordered here, so this we can considered as i/n-Ui<=1/2n and Ui-i-1/n<=1/2n this is true for i=1 to n, these 2 statements I can combine. Then this is nothing but Ui i/n-1/2n and on this side I have I now if you take it to the other side then it is becoming i/n-1/2n i=1 to n.

So this is interesting both the sides are the same so this is actually becoming= probability of Ui=i/n-1/2n so that was the reason that I mentioned that why I am considering Dn <=1/2n, because for this particular part this is giving extremely simple expression that is the probability of Ui, so certainly Ui's are continuously random variable therefore, this probability will be=0. So what we are finding here that the probability of Dn <=1/2n.

That means Dn has to start from 1/2n from the original definition it is not clear what is the starting point, so we said Dn lies between 0 to 1, but now we see that even Dn <=1/2n it is giving you probability 0. So we consider then probability of Dn<1/2n+something okay ae going as before what will happen here? Here I will get 1/2n+v, here I will get 1/2n+v, so here 1/2n+v, here I will get 1/2n+v, then if I am having this term here I will get -v here and here I will get +v.

So proceeding as above we get this= probability of $2i-1/n-v < U_i < 2i-1/n+v$ $i=1$ to n, now this is the joint probability for the random variables U1, U2, Un which are the order statistics from uniform 0,1, the joint pdf of this is known, so it is nothing but the n-fold integral over this region like for U1 you will have from $1/n-v$ to $1/n+v$, for U2 it will be maximum of so actually then this region because you also have even $U_1 < U_2 < U_n$ and they are lie between 0 to n.

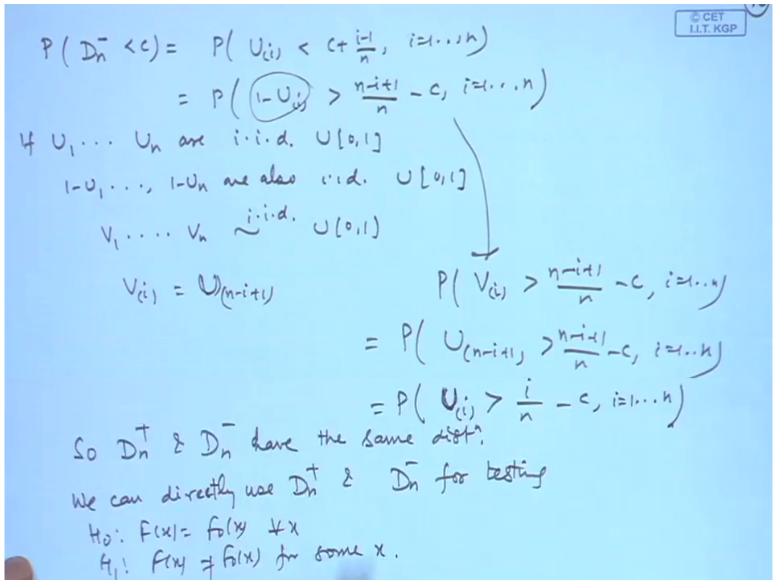
**(Refer Slide Time: 37:16)**




So this is nothing but, this is an n-fold integral over the joint pdf of U1, U2, Un that is n factorial for $0 < u_1 < u_2 < u_n < 1$ over the given region okay, so if I have to take say n=2 or n=3, then these things can be evaluated. Then Owen, Birnbaum and Smirnov, they have tabulated the values of the upper 100% alpha points of the distribution of the Dn, let us call it Dn, alpha of the distribution of Dn that is probability of Dn>Dn, alpha=alpha for various values of n and alpha. We can actually consider probability of Dn>Dn, alpha that= under H0= alpha.

So he also considered some asymptotic distribution also that is probability of $D_n < v/\sqrt{n}$ if we consider as limit n tends to infinity, it was shown that it is=1-twice i=1 to infinity-1 to the power i-1 e to the power i-2i v. And if I consider say a number c between 0 and 1, then probability of $D_n+1 < c$ that is= probability of maximum $i/n-U_i < c$ for i=1 to n that is= probability of $i/n-U_i < c$ i=1 to n that is= probability of $U_i > i/n-c$ for i=1 to n. Once again you see that this can be evaluated in terms of the joint distribution of the U1, U2, Un.

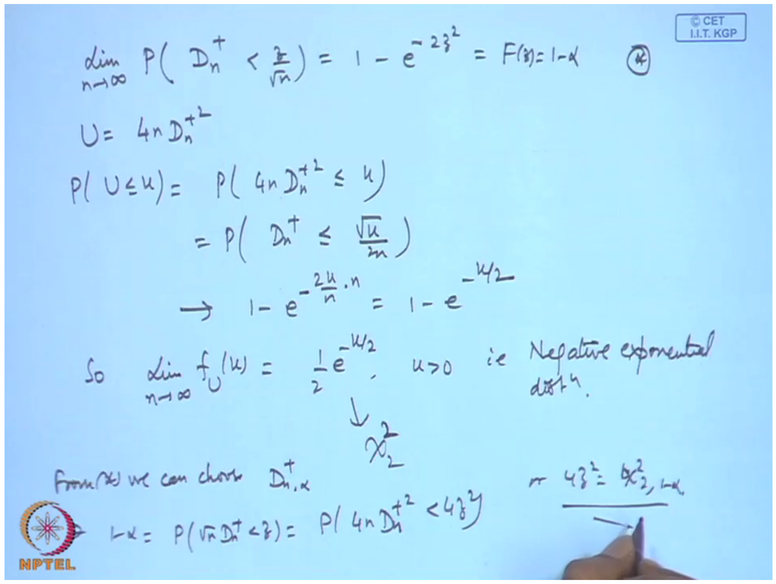
**(Refer Slide Time: 40:13)**




Similarly, if we consider Dn-1<c then this will be Ui<c+i-1/n i=1 to n that is= probability of 1-Ui>n-i+1/n-c, i=1 to n, if we consider say U1, U2, UN they are i i d uniform 0,1 then 1-U1, 1-U2, 1-Un are also i i d uniform 0,1 that means if I consider V1, V2, Vn which is 1-Ui then they are i i d uniform 0,1, so this is actually the same that means Vi=U of n-i+1, so this one then we can write as probability of Vn-1-Vi>n-i+1/n-c, i=1 to n.

Then this can be written as Un-i+1>n-i+1/n-c, i=1 to n, then this is nothing but Ui>i/n-c, i=1 to n, now you compare this with this expression here, probability of Dn+1<c is probability of Ui> i/n-c and it is the same thing here also. So what we are getting that Dn+ and Dn- have the same distribution, so one can directly use of the Dn+ and Dn- for the testing problem. We can directly use Dn+ and Dn- for testing H0 F x=F0 x, H1 F x!=F0 x for some x.

**(Refer Slide Time: 43:24)**

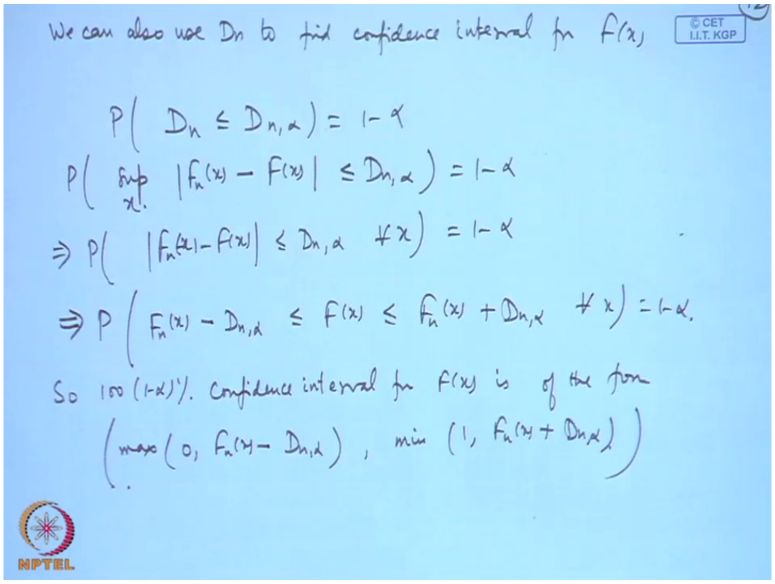



The asymptotic distribution of Dn+ etc. is also being worked out, if I look at Dn+1<z/root n as n tends to infinity that is= 1-e to the power-2 z square that is F z=1-alpha, if I define U=4n Dn+ square then probability of U<=u then that is= probability of 4n Dn+ square<=u that is probability of Dn+<= root u/2n. So if I apply this formula the limit will become=1-e to the power-2u/n*n that is 1-u/2.

So the limiting pdf of u that=1/2 e to the power-u/2 that is negative exponential distribution which also can be said as the Chi square distribution on 2 degree of freedom, of course since it is a negative exponential distribution the percentage points of this can be easily calculated, and we can express the test in terms of this also, so asymptotic test, asymptotic confidence interval can be obtained in terms of this, if we call this as say star.

Then from star we can choose say Dn+ alpha such that 1-alpha= probability root n Dn+1<z= probability 4n Dn+ square<4z square, that is 4z square=Chi square 2 1-alpha okay, so this can be easily calculated, one can easily find out the confidence interval for F x.

**(Refer Slide Time: 45:58)**



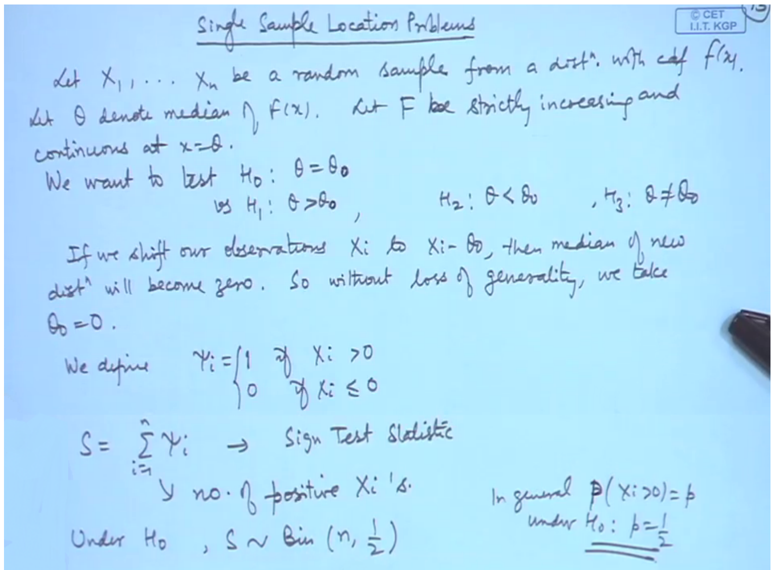

We can also use Dn to find confidence interval for f x that is probability Dn<=Dn, alpha that is= 1-alpha, so we write it as supremum of Fn x-F x this is equivalent to saying Fn x-F x this is <=Dn, alpha for all x that is=1-alpha, this is equivalent to saying probability Fn x-Dn, alpha<=F x<= Fn x+Dn, alpha that is=1-alpha. So 100 1-alpha% confidence interval for F x is of the form maximum of 0 Fn x-Dn, alpha to minimum of 1, Fn x+ Dn, alpha.

So we have seen that this test Kolmogorov-Smirnov test it is actually this is using more values compared to the Chi square test for goodness of fit was developed by Karl Pearson, in the Karl Pearson test essentially we reduced it to category problem that means we consider K classes out of the full distribution, and therefore, the test is more sensitive because what categories you are choosing, how many categories you are taking it will be dependent upon that.

Whereas this test is more robust, of course it is having sensitivity in the heavy tailed distribution but that is beside the point, there have been some modifications they have been proposed but essentially what we have seen is the distribution of the Dn is actually derivable here, so this is a much you can say improved thing compared to the chi-square test for goodness of fit, only thing is that the use of Kolmogorov-Smirnov is not that straight forward for the persons who have no idea about use of the statistics.

Because they need to understand the tabular version of the distribution of Dn that means how the percentage points are calculated, whereas for the Chi square test the percentage points are simply the percentage points of a chi-square distribution, so with a little knowledge of distribution theory on can actually apply the test. So it is a say compromise ease of applications is there in chi-square test, but the robustness is more in the Kolmogorov-Smirnov test.

**(Refer Slide Time: 49:40)**




Next we consider single sample location problems, actually in the beginning I have given you some applications of that 2 sample functions that F m, Yi that means which are based on F and G you have 2 samples and based on that some restrict is for the location etc. are given. Now I am getting into use of all this ranks here and to derive the few test for the location problem, so first we let us consider 1 sample problem.

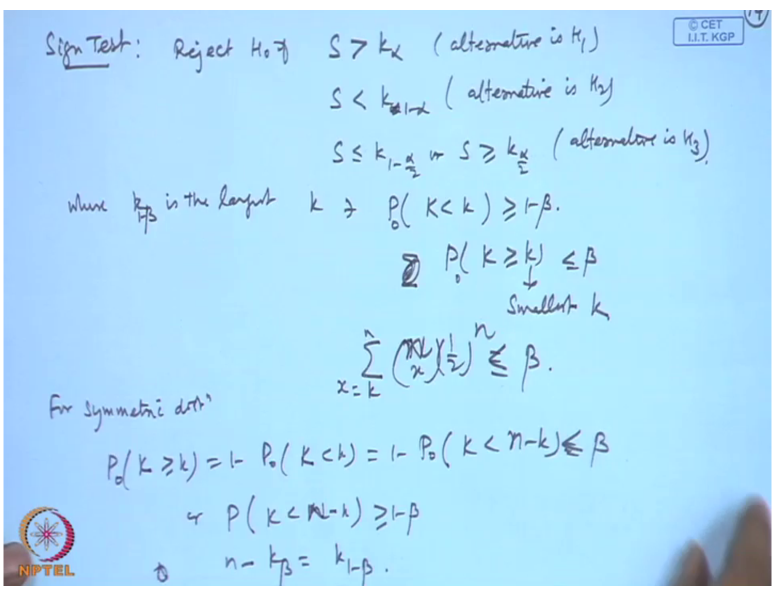
So let us consider X1, X2, Xn be a random sample, suppose you have the cdf f x, let theta denote median of F x, and we assume let F be strictly increasing and continuous at x=0 that means we are assuming median to be the unique. So we want to test theta= theta 0 against say theta is >theta 0, or theta <theta 0, or theta !=theta 0, so these 3 types of alternative hypothesis we will be considering, and you can compare it with the parametric testing problem.

In the parametric testing problem for one sample we were testing whether the mean value is = something < something > something != something, mu=mu 0 etc. we have done that testing problems under the assumptions of the normality. So here there is no distribution assumptions made except that it is a continuous distribution and strictly increasing and continuous at theta that means the medium is uniquely defined.

So now we want to test about some value theta 0, so whatever be the value theta 0 since this is known, if we shift our observations Xi to Xi-theta 0 then median of new distribution will become 0, so actually we do that thing. So without loss of generality, we take theta 0 to be 0, so that the problem becomes slightly simpler. We define say psi i= 1 if Xi is>0, it is 0 if Xi is<=0. And we define S= sigma psi i, i=1, this is called sign test statistic.

Because this is giving you the number of positive Xi's, how many Xi's are positive did exactly telling that thing. Actually under H0 if the null hypothesis is true then the 0 will be the median then under H0 S will follow binomial n, 1/2. See in general you will have probability of Xi>0 some p, but under H0 we will p=1/2, so under H0 the distribution of S is binomial n, 1/2 so we can actually device a simple heuristic test based on the sign test.

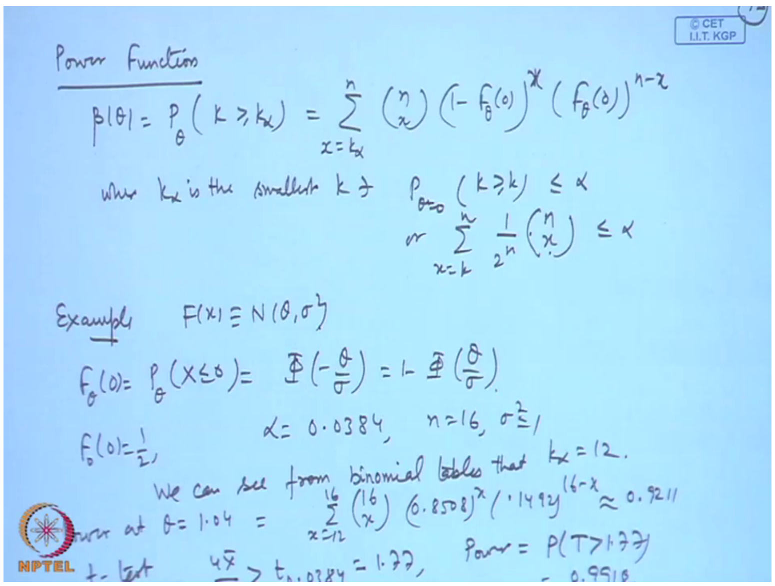
**(Refer Slide Time: 54:41)**




So sign test, reject H0 if S is> let us say some k alpha, this is if alternative is H1, and if S is <k alpha that is k 1-alpha if alternative is H2, and S<= k 1-alpha/2 or S>=k alpha/2 if alternative is H3. Where k beta is the largest k such that probability of k 1-beta is the largest k such that probability K<k is >= 1-beta, we have to take this largest etc. because the distribution is assumed to be discreet, so we may not actually achieved equality here that is probability of K>k alpha need not be alpha.

So that is why we choose the largest such cut off point, so basically we are saying that this condition is equivalent to that sigma, basically we are saying probability K>=k is<=beta, so then this small k is the smallest k that is sigma N C x 1/2 to the power n, x=k to n that is<=beta, so this can be easily calculated from the tables of the binomial distribution. If the distribution is symmetric, then we have this K>=k that is=1-probability K<k that is 1-P0 K<n-k that is<=beta. Then we can say that probability of K<n-k is>=1-beta, so you will have n-k beta=k 1-beta.

**(Refer Slide Time: 57:37)**




We can also calculate the power function here, power function of the sign test, probability of K>=k alpha when theta is the true value true medium then it is= x=k alpha to n, n c x 1-F theta 0 to the power x*F theta 0 to the power n-x, where k alpha is the smallest k such that probability K>=k under theta=0<=alpha or we can say 1/2 to the power n c x is<=alpha, for x=k to n. Let us take us an example here.

Suppose, I consider F x to be normal theta sigma square, so F theta 0 that is probability X<=0 that is=phi of –theta/sigma, that is 1-phi theta/sigma, so F 0 0=1/2, let us take say alpha= 0.0 384 n=16, sigma square=1. Then we can see from the binomial tables that k alpha=12, so if we consider the power function at say theta =1.04, then it is= sigma 16 c x 0.8508 to the power x, 0.1492 to the power 16-x, x=12 to 16, then that is approximately 0.9211.

If we consider the corresponding t test that is 4 x bar/S>t 0.0384 that is 1.77, then power of this test is= probability of T>1.77 that is= 0.9918, so certainly we can say that the power of the sign test is < the power of the usuality test that we already know, but this is under the assumption of the normality, if we actually have known all this about the normality then this test will be measurable, it will fail because this will simply give a wrong thing.

Another thing is that asymptotically also we can use the sign test, because we are saying k follows the S follows the binomial distribution, so this binomial distribution asymptotically becomes a normal distribution. So one can actually use this also, so in both the cases the results can be obtained and the cutoff point, the critical point and the power of the test can be easily calculated.

In the following lectures, I will be describing some other test statistics which are based on the order statistics, in the sign test actually the order statistics are not used only the sign of the term is important. So therefore, in that sense you can say extremely simplistic test for the median of the distribution. Next, we will define certain test which will be based on the actual values or the actual measurements, so that I will be starting like Wilcoxon signed-rank test statistic, Mann Whitney and so on.