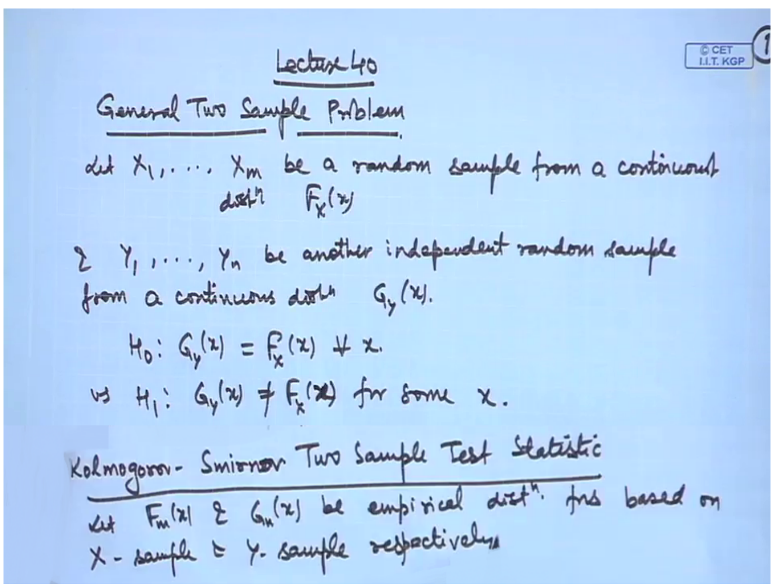
Statistical Methods for Scientists and Engineers
Prof. Somesh Kumar
Department of Mathematics
Indian Institute of Technology - Kharagpur

Lecture - 40
Nonparametric Methods - XIII

As I mentioned that, in place of considering location alternative or the scale alternative, one may consider general alternative that means we simply say that the 2 distributions are not the same. So now an analogous situation you can think of in the case of parametric or nonparametric situation is that we may think of not specifying the distribution and then talking about the location parameter or the scale parameter rather we considered simply the specification of the distribution itself.

Now in the one sample problem one solution was given by the Chi-square test for goodness of fit, and another solution is given by the Kolmogorov-Smirnov test statistics. Now let us define the general 2 sample problems and discuss the analogous test here.

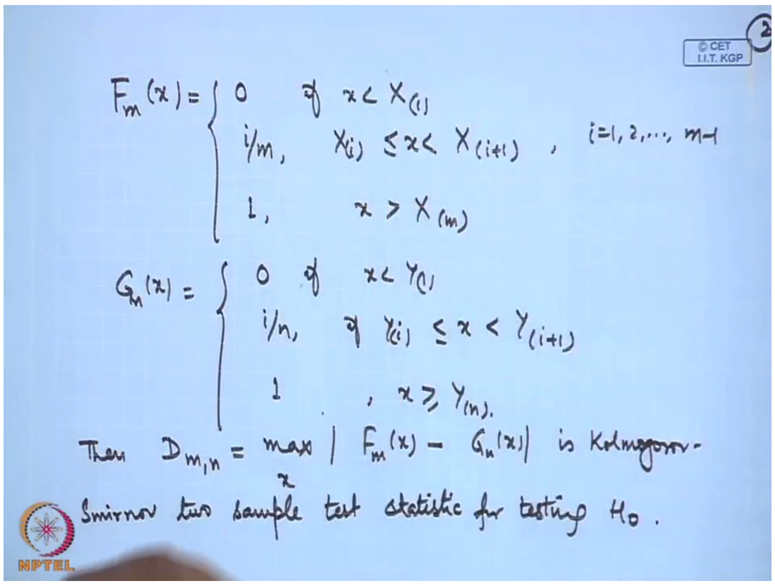
(Refer Slide Time: 01:18)



So the general two sample problems, so let X1 , X2 , Xm be a random sample from a continuous distribution Fx, and say Y1 , Y2 , Yn be another independent random sample from a continuous distribution Gy okay. So we want to test the hypothesis whether Gy x=Fx x for all x against the alternative Gy x !=Fx x for some x, for this problem one of the you can say most possible solutions is the Kolmogorov-Smirnov two sample test statistic..

So let us define firstly the empirical distribution functions based on X1, X2, Xm and Y1, Y2, Yn, so let Fm x and Gm x be empirical distribution functions based on X sample and Y sample respectively.

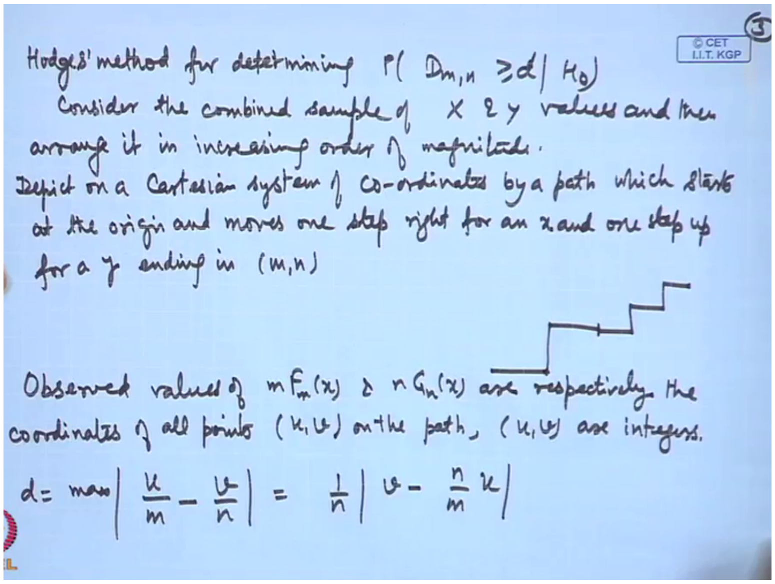
(Refer Slide Time: 03:57)



So we are defining like Fm x=0 if x<X1, it is=i/m if Xi<=x<Xi+1 for i=1, 2, m-1, it is = 1 if x>Xm. Similarly, Gm x=0 if x<Y1, it is=i/n if Yi<=x<Yi+1 and it is=1 if x>=Yn. Then Dm, n that is= maximum of Fm x-Gn x, if you remember in the case of one sample problem it was defined as Fm x-F x, here now I am considering the difference between the distribution, the 2 distributions basically, so this is actually the Kolmogorov-Smirnov two sample test statistic for testing H0.

Now naturally if the 2 distributions are close, then naturally their sample values will be closer to each other and therefore, there will be region where Fm and Gm will be similar. So therefore, the differences will be smaller otherwise the difference will be large, so we can simply say that large values of Dm, n indicate that H0 is false, so Dm, n is actually distribution free. We will actually discuss the method how to determine the probability, the probability of type 1 error, type 2 error based on this thing.

(Refer Slide Time: 06:59)

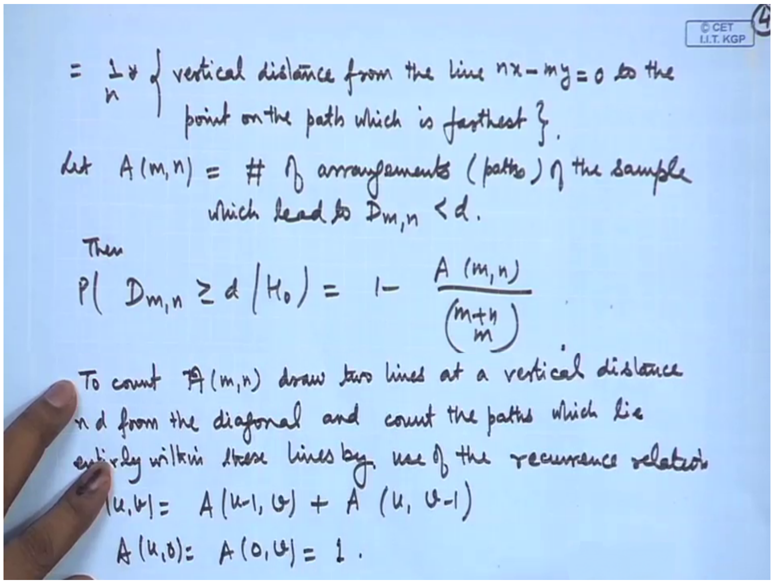


So this is distribution free let us discuss these methods now, so one method that has discussed it is actually by Hodges' method, Hodges' method for determining probability of Dm, n>=d under H0, so first thing is that we consider the combined sample of X and Y values, and then arrange it in increasing order that means we rank them of magnitude. So we consider a Cartesian system in the following fashion.

Depict on a Cartesian system of coordinates by a path which starts at the origin, and moves one step right for an x, and one step up for a y, ending in m, n, so it is like this if x is next then we come like this otherwise it go like this so it could be like this and so on. So observed values of m Fm x and n Gn x they are actually the all the coordinates of the points of u, v on this path, where u and v are integers.

Observed values of m Fm x and n Gn x they are respectively the coordinates of all points u, v on the path, where u and v are integers, so d= maximum of basically u/m-v/n which we can write us i/n v-n/m u.

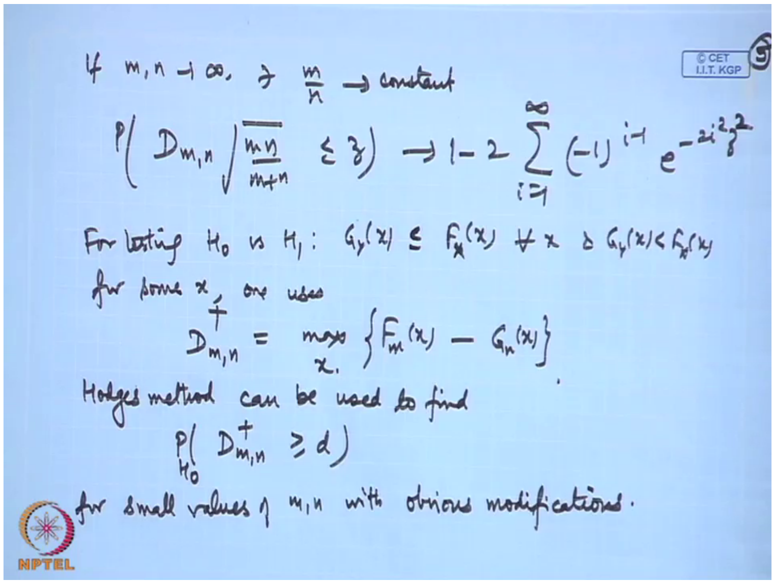
(Refer Slide Time: 10:18)



We can express it as i/n*the vertical distance from the line nx-my=0 to the point on the path which is farthest, so m, n is a number of arrangements that is paths of the sample which lead to Dm, n<d. Then you will have probability of Dm, n>=d under H0 that is=1-A m,n/m+n c m because m+n c m will be all the paths. We can have the extreme cases when all the Xi's are below all the Yj's are one could be in which the all of them are like 1 X1 Y like that all, it could be like that also.

So Am, n is the arrangements of the paths of the sample which leads to Dm,n<d, then it is probability that Dm,n>=d=1-A m,n/m+n c m, so to count A m,n we have to draw 2 lines at a vertical distance n d from the diagonal, and count the paths which lie entirely within these lines by use of the recurrence relation that is A u,v is actually=A times u-1,v+ A of u, v-1, and of course A u,0 that will be=A 0,v that is=1.

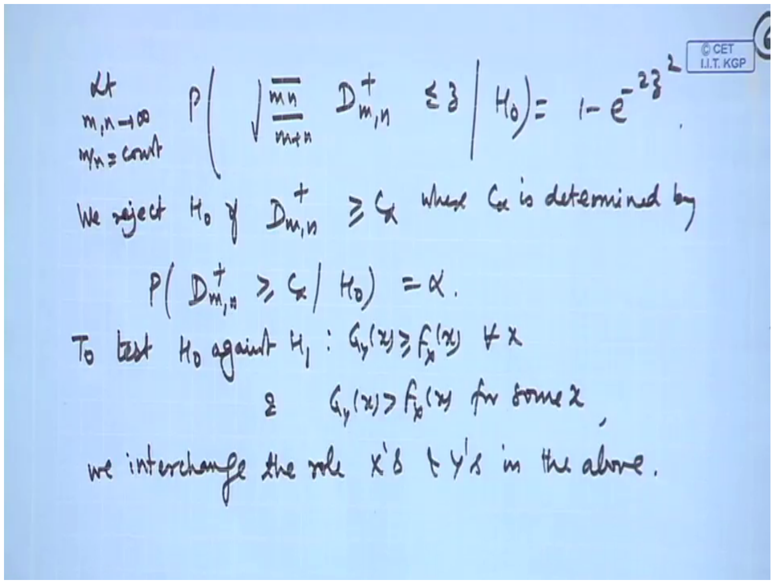
(Refer Slide Time: 13:31)



Let us also consider this, if m,n tend to infinity such that m/n goes to a constant, then probability that Dm,n root mn/m+n<=z it convergence to 1-twice-1 to the power i-1 e to the power -2i square z square i= 1 to infinity. So asymptotic distribution is obtained here for testing H0 versus H1, where H1 could be like stochastic ordering that is say Gy x<=Fx basically it means Y<=X<=probability X<=x.

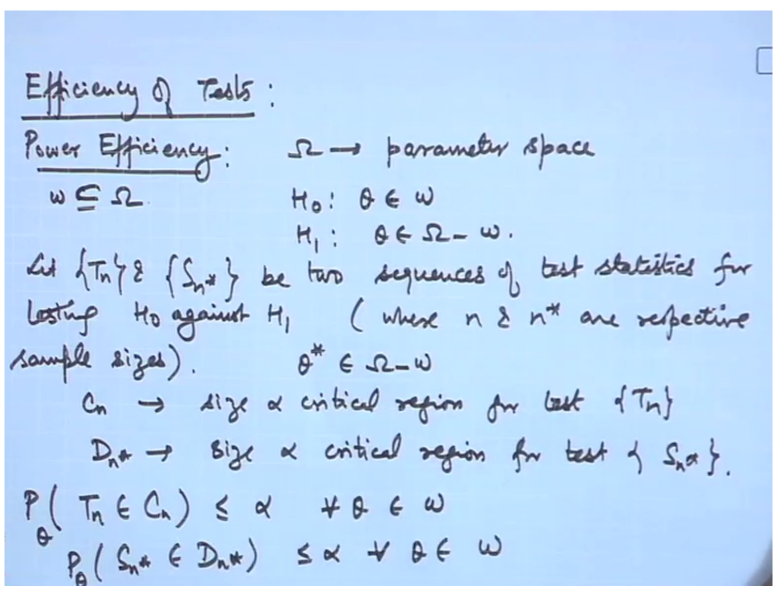
That means basically what we are saying is that X is stochastically smaller than the X is stochastically smaller than y, and for all x Gy x<=Fx for some x, so one uses a version of this let us call it Dm, n+ that is maximum of that means we consider only one sided maximum of Fm x-Gn x. Hodges Method can be used to find probability of Dm, n+ >=d under H0 for small values of m, n with obvious modifications

(Refer Slide Time: 15:57)



One can talk about the full-fledged asymptotic distributions of Dm, n+ also, let me give that here, limit of m, n tending to infinity such that m/n is a constant of the probability root mn/m+n Dm, n+ <=z under H0 that is 1-e to the power -2z square. We reject if Dm, n+ >= C alpha, where C alpha is determined by probability of Dm, n+ >=C alpha under H0=alpha. To test H0 against H1 Gy x>=Fx for some x, we interchange the role of X's and Y's in the above. So basically you can consider here Gn x-Fm x, and then the similar things will be holding here.

(Refer Slide Time: 17:37)



Another concept is that of efficiency of the tests here, like the consistency of the test, like consistency of estimators we have an analog of the efficiency of the test. So first is the power efficiency, let us consider that power efficiency. This concept, I am introducing because in the usual classical theory of testing of hypothesis in the parametric situations, we have the concepts of the most powerful test, uniformly most powerful test, unbiasedness and other things.

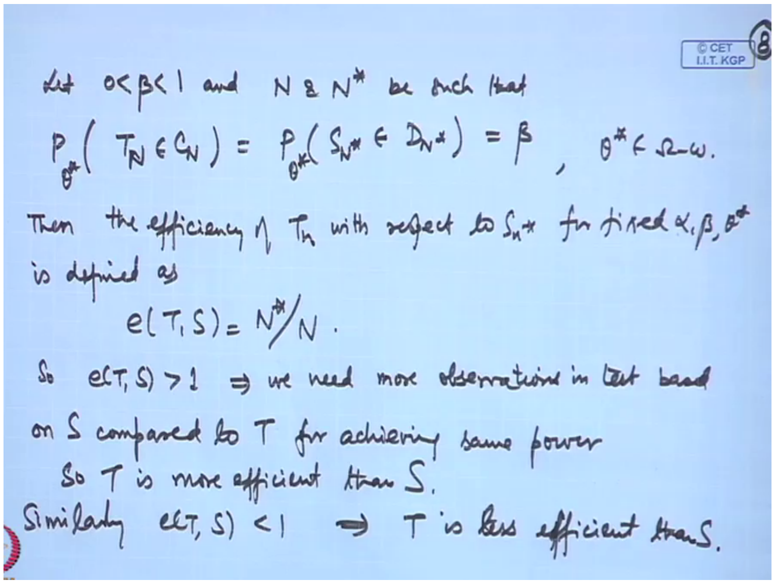
Now when we are dealing with the distribution free situations or the non-parametric situations, then certainly those type of most powerful and other concepts are not valid, and therefore, we need to look at the asymptotic properties, so that is why consistency and now we are talking about efficiency, so first thing is that about the power efficiency. So let us consider say omega is the parameter space here.

So suppose I have a point a subset of omega, so we are considering the hypothesis testing problem H0 theta belonging to small omega against H1 theta belonging to capital omega-small omega. So let us consider say Tn and Sn so let me put a Sn* here be 2 sequences of test statistics, we considered sequence because they are based on a sample size, and so sample size may vary, so that is why we are calling it for testing H0 against H1.

So here n and n* are the respective sizes, so let us take n arbitrary point theta* belonging to alternative hypothesis that, so let us consider Cn is the size alpha critical region for test Tn and let us call say Dn* is the size alpha critical region for test Sn*. So now let us consider probability of Tn belonging to Cn, then this will be certainly<=alpha for all theta belonging to omega.

And probability of Sn* belonging to Dn* that will be <=alpha for all theta belonging to omega. So these are 2 size alpha test and critical regions are identified by Cn and Dn* respectively based on Tn and Sn*, let us consider now the power here.

(Refer Slide Time: 21:39)

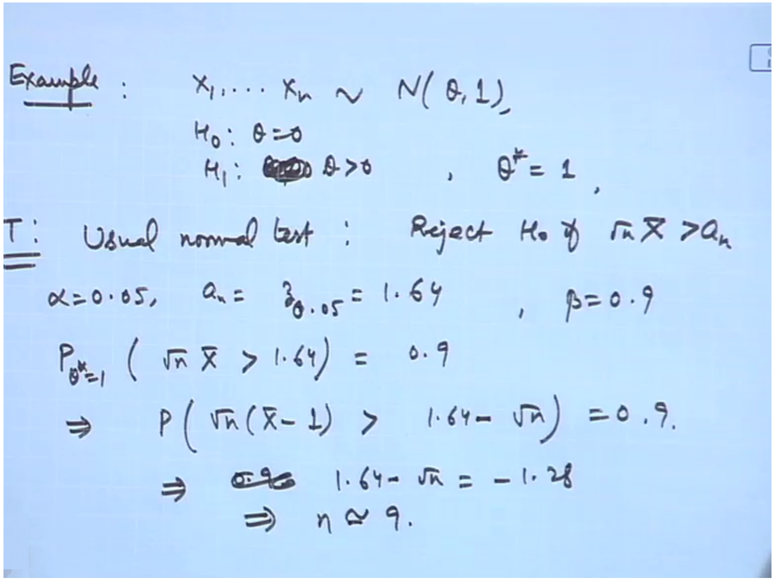


Let us fix powers say beta and suppose we require N and N* as the sample sizes that probability of TN belonging to CN* that is= beta under some value say theta*, then and here this theta* belongs to actually omega-omega. So then the efficiency of Tn with respect to Sn* this is for fixed alpha, beta and theta*, this is defined as e T, S=N*/N, so naturally you can have the interpretation here.

So for example if e T, S >1, if it is >1 that means for achieving the same level of power I need more observations in as compared to T that means T is more efficient than S, if this is <1 e T, S <1 then it will mean that I need more observations in T compared to S for achieving the same level, so T is less efficient than P. So we can write that, so e T, S >1 implies that we need more observations in test based on S compared to T for achieving same power, so T is more efficient than S.

Similarly, e T, S<1 implies T is less efficient than S, so for this efficiency we have some sort of interpretation here, but of course this definition is having unlimited usage here because we have fixed here alpha, beta and theta*, if we change the theta* and the underline distribution changes here, then I mean for different underline distributions it may quite complicated to adopt here, but anyway this is the definition let me consider some example here.

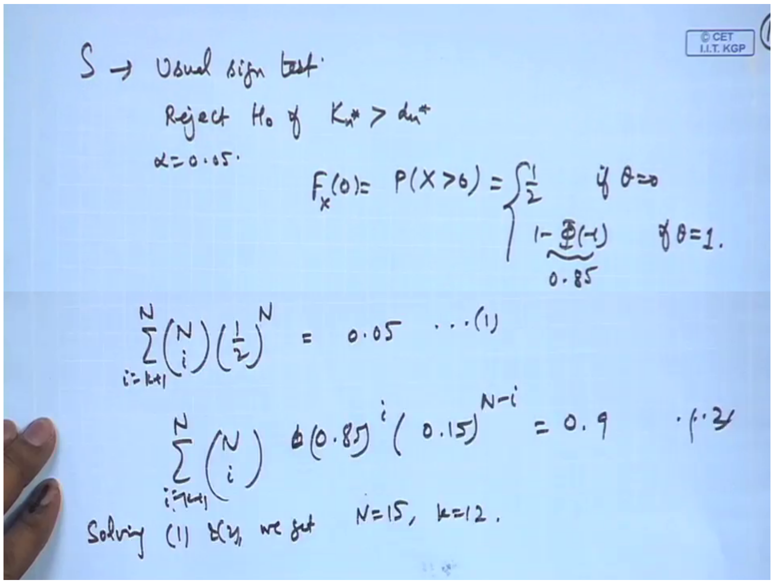
(Refer Slide Time: 25:08)



Let us consider the usual problem based on say normal distribution say normal theta 1, and I consider H0 theta=0, H1 theta !=0 or say theta >0 for example let us take, so let me take theta* in the alternative hypothesis set as 1. So the usual normal test is okay, the usual normal test that is reject H0 if root n X bar >a n basically that is say alpha, so if I am taking say alpha=0.05 then a n=z0.05 that is= 1.64.

So if I considered say probability theta*=1 root n X bar>1.64=say 0.9, so I am taking say beta= 0.9 say for example, so this will mean probability root n X bar-1> so this will become 1.64-root n=0.9, so that means 0.9= sorry 1.64-root n=-1.28, so this you can simplify so this is n is approximately 9 okay.

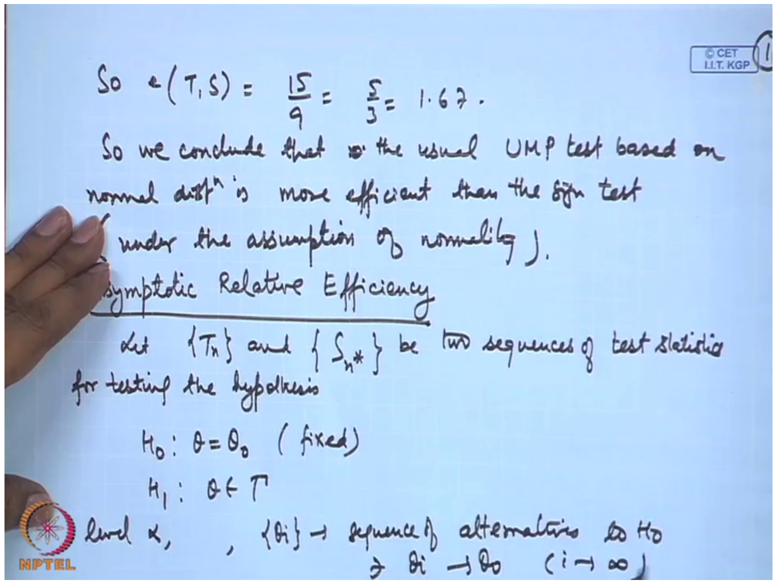
(Refer Slide Time: 27:13)



Let us take second test say based on the sign test function here S say usual sign test, for usual sign test you are saying reject H0 if Kn>dn so let me put here *, so if I am taking alpha=0.05 and so okay, let us see Fx 0 that is= probability X>0 that is 1/2 if theta=0 it is=1-phi of -1 that is approximately 0.85 if theta=1 for example okay. So now we are having 2 equations sigma N c i 1/2 to the power N that is=0.05, this is for i=k+1 to N, this is one equation.

And the second equation is sigma N c i of 0.85 to the power i and 0.15 to the power N-i for i=k+1 to N that should be=0.9, so you need actually to solve these 2 equations to get the value of N, and actually K is also required here because at which point that cutoff will come. So solving 1 and 2 we get N=15 and K=12.

(Refer Slide Time: 29:01)

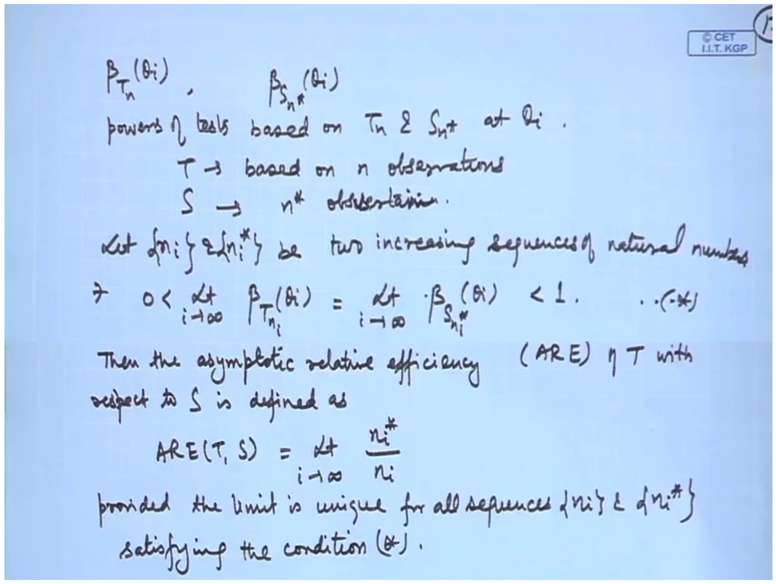


so what we conclude then, so efficiency of normal test corresponding to this one will become=15/9 that is=5/3 that is=1.67 etc. so basically we conclude that the usual UMP test based on normal distribution is more efficient than the sign test, now this is under the information that the under the assumption of the distribution, if the normal distribution is not then certainly you cannot use this, under the assumption of normality.

So if the distribution is known certainly we will use that information, now this is for the power 0.9, if we use some other power say 0.95 or 0.92 then certainly this efficiency level will change, but certainly it means that the test based on the normal test is certainly more efficient than the sign test here. Since this is completely dependent upon the values of alpha, beta and the point that we have chosen, there is another concept that is called the asymptotic relative efficiency of the test, let us consider this concept now.

Asymptotic Relative Efficiency, let Tn and Sn* be 2 sequences of test statistics for testing the hypothesis H0 theta=theta 0 this is fixed against the alternative theta belonging to some gamma, and we are considering level alpha and theta i's this is a sequence of alternatives to H0 such that theta i converges to theta 0 as i tend to infinity, so we are considering alternatives but they approach the null hypothesis value.

(Refer Slide Time: 32:14)

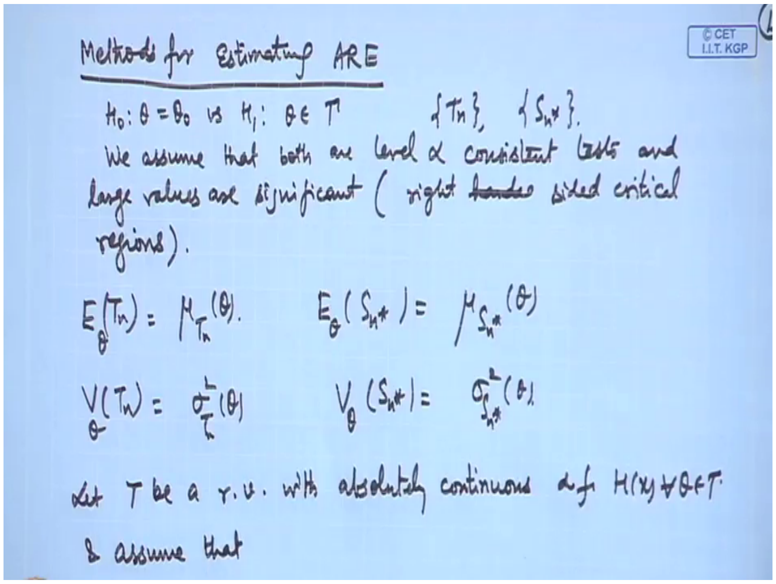


So let us now consider say beta Tn theta i, and beta Sn* theta i, so these are the power of tests based on Tn and Sn* at theta i, so T is based on n observation, and S is based on n* observations, so we consider this ni and ni* sequences and I consider the conditions such that ni and ni* increasing sequences of natural numbers such that 0< limit as i tends to infinity beta Tni theta i= limit as I tends to infinity beta Sni* theta i<1.

Then the asymptotic relative efficiency that is ARE of T with respect to S is defined as ARE T, S= limit as i tends to infinity ni*/ni provided the limit is unique for all sequences ni and ni* which satisfying the condition* okay. So what does it roughly mean? It roughly mean that suppose I say this values= say 1.2, then it would mean that in test based on S I need on the average 120 observation if there are 100 observations needed by the test based on T to achieve the same level of power.

So now since this definition is a limiting definition, this is you can say better definition than the definition of efficiency which I gave just before this, because that was a fixed sample size definition here, and here it is an asymptotic kind of definition.

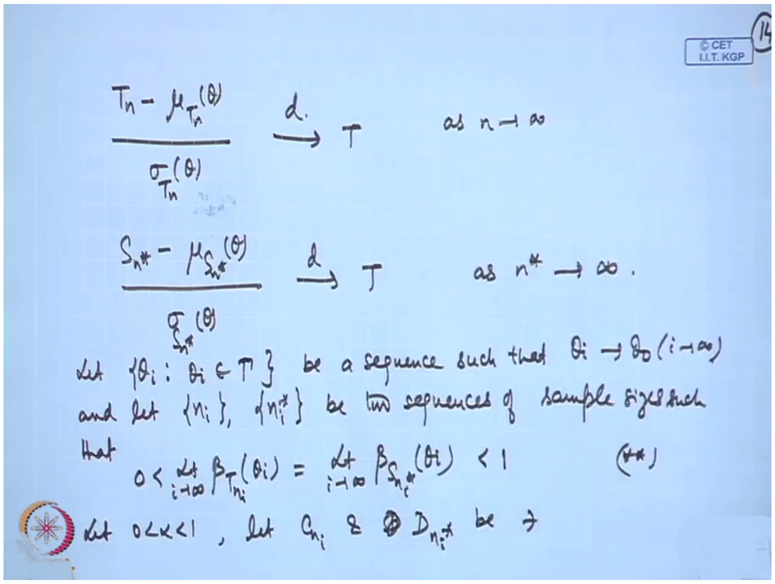
(Refer Slide Time: 36:25)



Now there are several ways of calculating the asymptotic relative efficiency, let me briefly mention about that. Methods for estimating the asymptotic relative efficiency, so we are having 2 test sequences, so we have the hypothesis testing problem theta=theta 0 against H1 belongs theta to some gamma, and we have Tn as a test sequence and Sn* is having another test sequence. And we assume that both are level alpha consistent tests.

That means they are one sided let me say and large values are significant, basically it means that we are having right handed critical regions, right sided or right handed critical regions this is as for convenience otherwise to compare 2 test it will become quite complicated. So let us consider that say the expectation of Tn is say mu Tn theta, expectation of say Sn* that is=mu Sn* theta, variance of Tn say that is=sigma square Tn theta, and variance of Sn* say = sigma square Sn* theta.

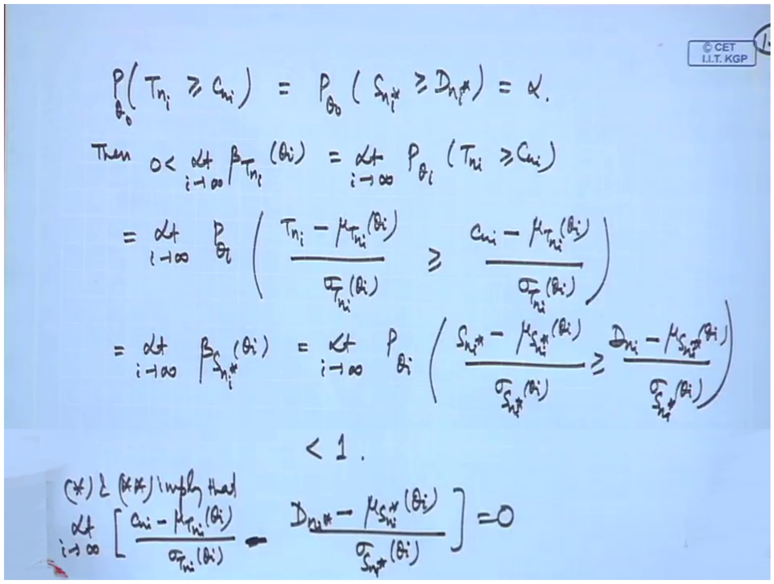
(Refer Slide Time: 39:01)



So let us assume that let T be a random variable with absolutely continuous distribution function H x for all theta belonging to gamma, and we assume that the asymptotic distributions of standardize variables of Tn and Sn they are based on the T, they are basically T we are saying that they converges to distribution T as n tends to infinity, we also assume that Sn*-mu Sn* theta/sigma Sn* theta converges in T.

Let theta i, theta i belonging to gamma be a sequence such that theta i goes to theta 0 as i tends to infinity, and let ni, ni* be 2 sequences of sample sizes such that 0<limit beta Tni theta i that is=limit beta Sni* theta i, i tending to infinity, i tending to infinity is <1. So let us take say alpha between to be 0 and 1, and let the critical region points be such that.

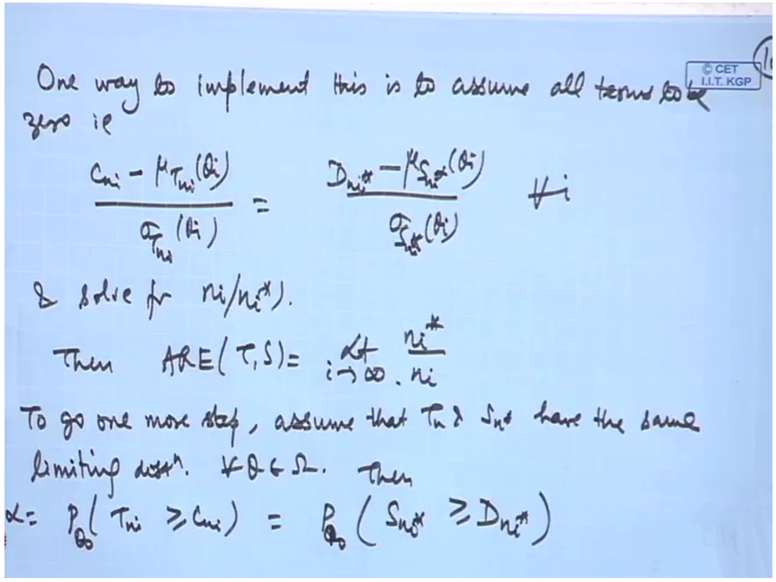
(Refer Slide Time: 40:59)



Probability of Tni>=Cni that= Sni* >= Dni*= alpha, so then you can consider here 0< limit of beta Tni theta i, as i tends to infinity that is= limit of probability theta i Tni>= Cni as i tends to infinity. Now based on the asymptotic distribution of Tni, we can express it as limit as i tends to infinity probability of theta i Tni-mu Tni theta i/sigma Tni theta i>= Cni-mu Tni theta i/sigma Tni theta i,

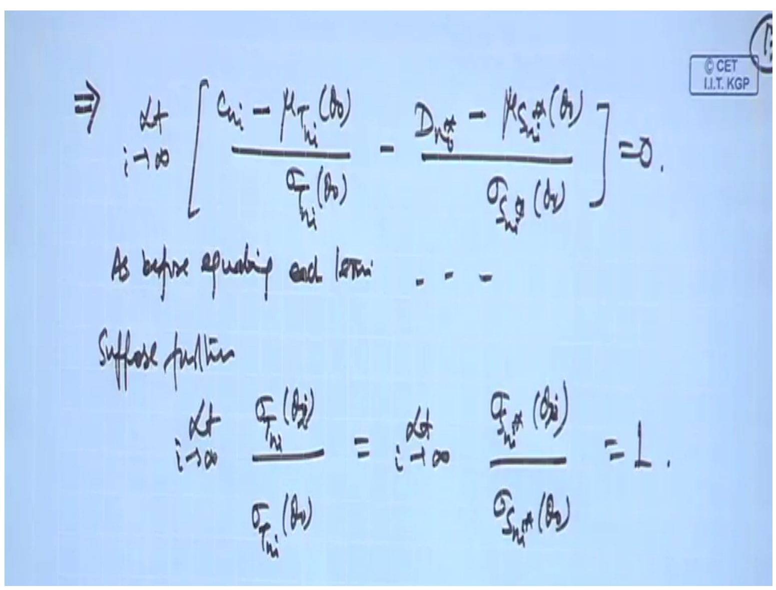
That is=limit as i tends to infinity beta Sni* theta i that is= limit as i tends to infinity Sni*-mu Sni* theta i/sigma Sni* theta i>=Dni*-mu Sni* theta i/sigma Sni* theta i that is<1. So if we compare the let me call this as *, so from these 2 relations we get this imply that limit as i tends to infinity of Cni-mu Tni theta i/sigma Tni theta i=-Dni*-mu Sni* theta i/sigma Sni* theta i=0 okay.

(Refer Slide Time: 43:47)



So, one way to implement this is to assume all terms to be 0 that is Cni-mu Tni theta i/sigma Tni theta i= Dni*-mu Sni* theta i/sigma Sni* theta i for all i, and solve for ni/ni*. Then ARE of T, S= limit as ni*/ni for I tending to infinity that is one way. Now to go one step further, you can assume that Tn and Sn* have the same limiting distribution for all theta, if that is so then you will have alpha= probability theta 0 Tni>= Cni=P theta 0 Sni*>=Dni*.

(Refer Slide Time: 45:48)



So this will imply that limit as i tending to infinity Cni-mu Tni theta 0/sigma Tni theta 0-Dni*-mu Sni* theta 0/sigma Sni* theta 0=0, so if we can equate each term then again we can define, equating each term we can get the ARE. We can also assume, so let me just briefly mention this thing because all of them are leading to an additional condition here, we can also assume the standard deviation that is this one that is=1, then again we can get the condition here.

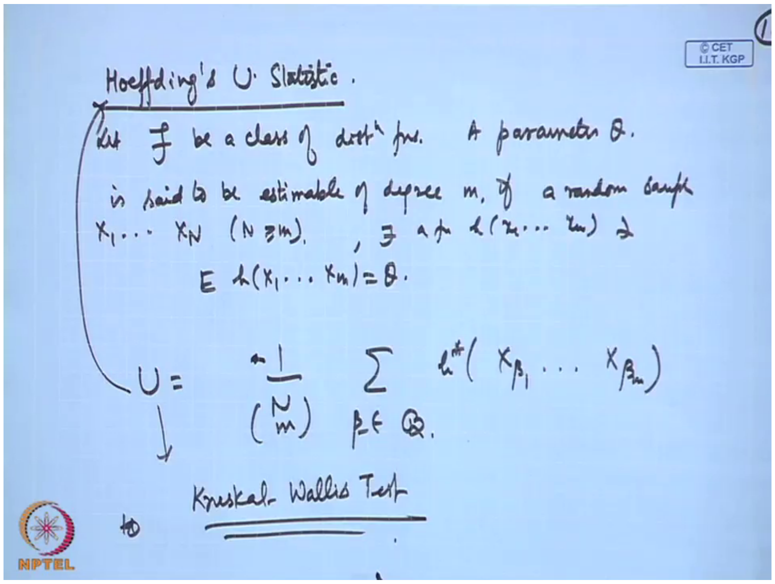
So these are the methods for estimating the asymptotic relative efficiency, there are certain results regarding this, but I am not getting into this thing. In nonparametric statistics I have discussed main methodologies that is based on the, firstly we discussed in detail the role of the order statistics, asymptotic distributions of the order statistics under various conditions. We also discuss the estimation of the distribution function based on the empirical distribution function.

We considered the ranks, the distribution of the ranks based on the ranks we considered several 1 sample and 2 sample testing problems basically based on the location and scale, we gave certain test. Towards the end I have introduced the concept of asymptotic relative efficiency, the concept of the efficiency of the test. The subject of a non-parametric test or you can say the distribution free procedures is quite vast, it has developed a lot in past 50 years.

And there are various other aspects of it, for example there is a huge class of problem which is known as the Hoeffding's U statistics, let me briefly mention about this one, we will very briefly talked about the Kruskal-Wallis and also the say Kendall's Tau and Spearman's thing, in the remaining 5 minutes or so or 10 minutes or so, let me just briefly mention about these topics. There are other things like density estimation based on the Kernel method.

And there are so many other topics that have developed over the past few years in the field of nonparametric, but in this particular course we can cover only this much, maybe in another course we will discuss in detail about those procedures also.

(Refer Slide Time: 49:38)



So let me talk now about the Hoeffding's U statistic, so what is a Hoeffding's U statistic? Hoeffding's U statistic is defined as that let F be a class of distribution functions, a parameter theta is said to be estimable of degree m, if a random sample X1 , X2 , XN N>=m, there exist a function h of x1, x2, xn such that expectation of h of X1, X2, Xm= theta, so this is based on this we define a Kernel, so h of X1, X2, Xm is called a Kernel.

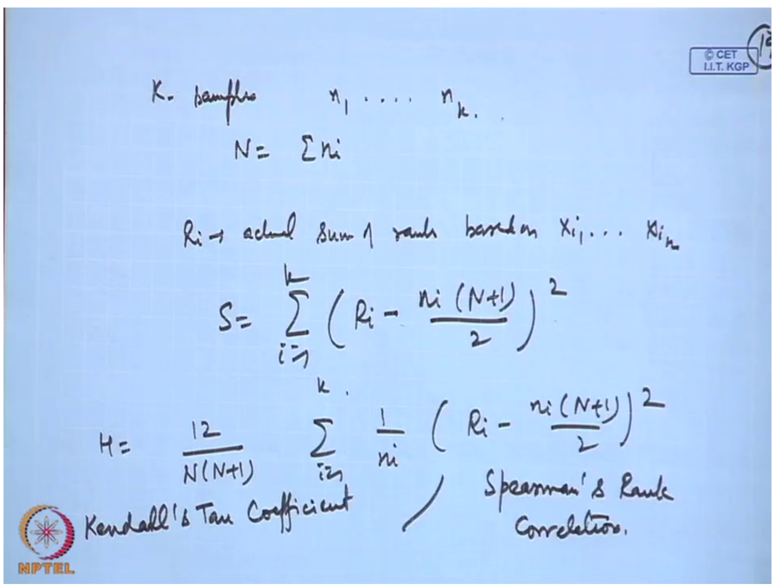
Now based on this we can define a Hoeffding's U statistic as sigma 1/N c m sigma h* X beta 1 up to X beta m by taking all the permutations of that means by taking all m permutations from N here that means it based on that, this is called Hoeffding's U statistic, so this is actually the name of this. So basically this is a symmetric function and its augments, and you can actually see that most of the standard estimates in the parametric situations.

And also in the nonparametric situation like Sign test, Wilcoxon signed-rank test, Mann Whitney test statistic they will be actually based on the Hoeffding's test. The good thing about the Hoeffding's U test statistic is that the asymptotic theory is established that means the asymptotic normality is there. And therefore, various procedures for this can be established for this one. As I mentioned in nonparametric methods are also applicable, when the normality assumption is not satisfied, there also situation like when we have large scale data.

So when we were unable to fit a distribution or it is impossible to say that which distribution will be applicable, because there may be too much variations when you are doing the fitting of the distributions, then certainly nonparametric distributions non-parametric methodologies are used. One of the important methodologies for example you can think of is the analysis of variance, in analysis of variance we assume that the errors are normally distributed.

But if the normal errors are not normally distributed, then we go for a distribution free procedure that is called Kruskal-Wallis test, so Kruskal-Wallis test procedure is based on the again ranks here actually, so if the ranks are here we are considering let me just very very briefly mention about this one.

(Refer Slide Time: 53:01)



So we have K samples and the sample sizes are n1, n2, nk and n=sigma ni, we consider all the observations mixed up, and we consider that Ri is the actual sum of ranks based on Xi1, Xi2, Xin in the combined sample, then you define the statistic Ri-ni*N+1/2 whole square i= 1 to k, so this is actually for the Kruskal-Wallis, and then we define H=12/n*N+1 1/ni Ri-ni*N+1/2 whole square i= 1 to k. So based on this the testing procedure has been developed, so this is called the Kruskal-Wallis formula.

The general linear rank statistics for a sample problems have been also define and then again versions of like Wonderwaden statistics theory, test statistics etc. are also defined. Another thing is regarding the coefficient of correlation or the coefficient of association etc. so we have the usual Karl Pearson coefficient and the distribution for theory for that has been developed which I mentioned we have discussing the multivariate statistics.

But again that is based on the normality assumption, if the normality assumption is not there then we define measure of association and also a correlation coefficient, which is based on the ranks. So we have the 2 following things, one is called Kendall’s Tau Coefficient, and another one is called Spearman's Rank Correlation, so I mentioned about this thing how they are used for example when we are discussing the same side concordance or discordance then Kendall’s Tau Coefficient is used.

For example we have seen which side the values are like Xi’s<Xj then Yi’s also <Yj, Xi’s>Xj then Yi’s>Yj, in such cases the Kendall’s Tau Coefficient is used here. And there are other situations like we are considering the coefficient of correlation then it is based on the observed values Xi, Yi that is X1 X2, X1 Y1, X2 Y2, Xn Yn are taken and we based on this we define the coefficient, but here the actual observations are used.

That means for example we may considering weights, we may considering heights and so on for 2 sets of this one. But suppose we are not interested in the actual values, rather we are interested in the ranks, because it could be a measurements maybe related to something which may not give actual picture for example if you consider the how much agreement is there between the ranks given by 2 judges for ranking certain athlete.

So for examples some gymnastic competition is there, and then there are several competitors and there are 3 judges or 2 judges are there, so those 2 judges will give ranks on the basis of each performance, now then whether there is an agreement that means whether they are on the same side that means the good performer is appreciated similarly, by the 2 judges or they are biased that means whether 1 judge judges one candidate as better compared to the other one, in such cases we use Spearman rank correlation coefficient and it is much more useful.

We end this course at this note, as I mentioned that there are so many other things in the statistical methods for scientists and engineers, there are so many new methodologies have been developed for example large literature is available in the regression methodologies, time series and other thing. But of course then certain other course will be available for that. There is a vast area of Bayesian methodologies and frequent test decision theory, so one may look at those topics for applications. I will end this course at this particular note.