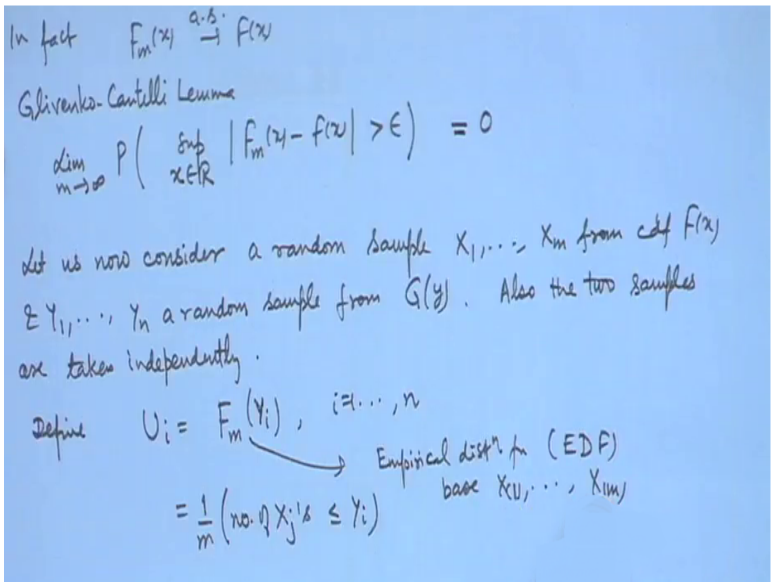
Statistical Methods for Scientists and Engineers
Prof. Somesh Kumar
Department of Mathematics
Indian Institute of Technology - Kharagpur

Lecture - 33
Nonparametric Methods - VI

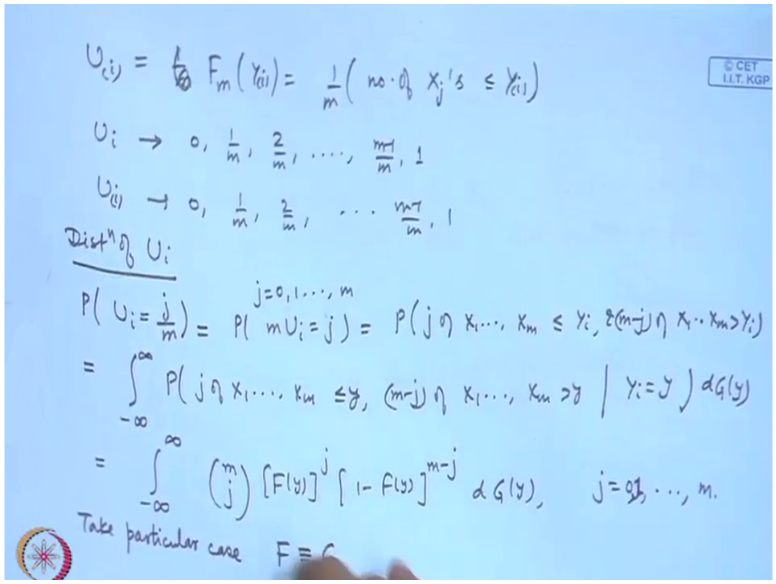
In the previous lecture, I had defined 2 types of a Statistics let me recall those things.

(Refer Slide Time: 00:27)



So we are considering that random sample X1, X2, Xn is taken from cdf F x and Y1, Y2, Yn is a random sample from cdf G y, and we also assumed that the 2 samples are taken independently, then based on the empirical distribution function of first sample we defined Ui=Fm of Yi.

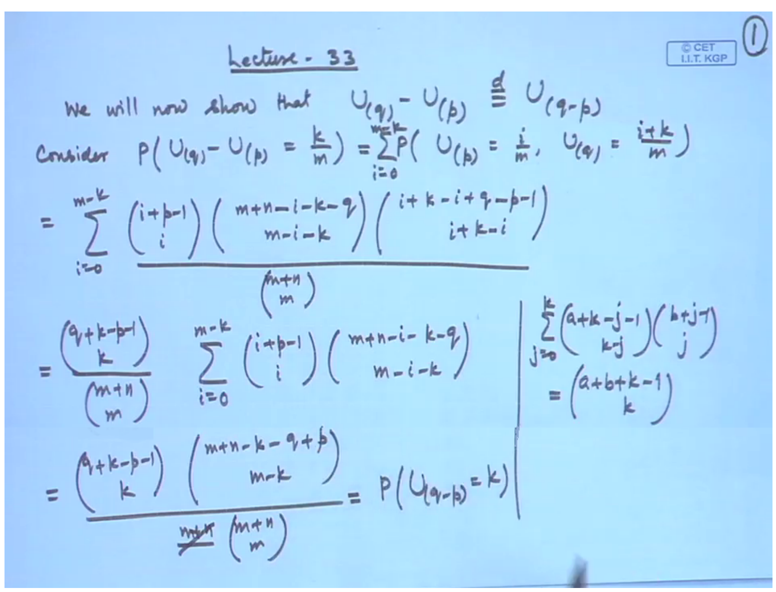
(Refer Slide Time: 00:56)



And similarly, we defined U bracket i=Fm of Y of bracketed I, here again Fm is the empirical distribution function of the first sample. So this is the number of Xj is <=Yi and of course divided by m, and here it is the number of Xj's <=Yi/m, and both Ui and U bracket i they take values 0, 1/m, 2/m, up to m-1/m, 1. And we consider the forms of the distribution of Ui, and the distribution of U bracketed i. We also considered the special case when F=G.

Now let us, and we also consider the joint distributions of Ui, Uj, U bracket i, Uj, we also looked at the moment structure of these quantities, let me now proceed from here.

(Refer Slide Time: 02:05)



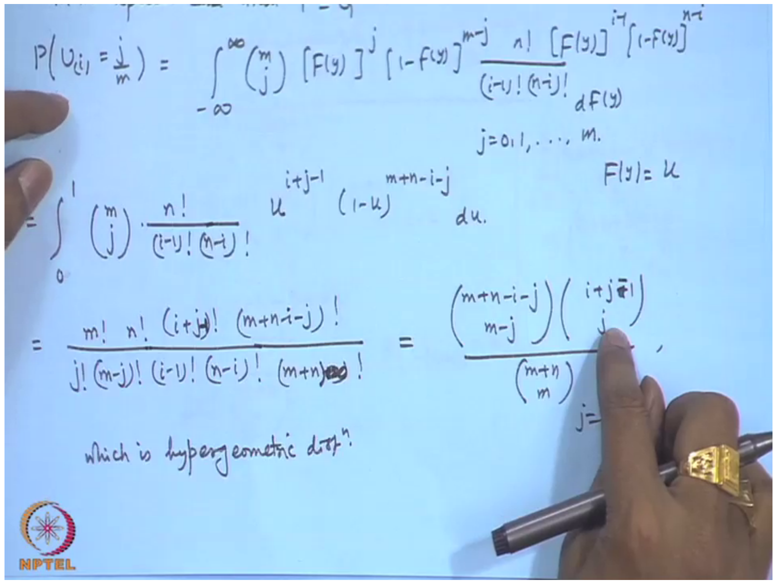
So first thing is that we consider the joint distribution of Up and Uq, now I look at the distribution of the difference so we have the following result. We will now show that the distribution of Uq-Up is the same as the distribution of Uq-p, so for that let us consider the distribution of Uq-Up, so Uq-Up= say something like k/m since the values taken by these are of the form p/m.

So we can consider like this the difference is that k/m type only that means one of them is i/m and other one is of the form i+k/m where i is varying from 0 to m-k that is= now here I write down the joint distribution of this which we derived in the previous lecture let me recollect that thing. The joint distribution of Up and Uq, so the values taken are j/m and l/m, then it is given by this expression j+p-1 c j and all these things.

So here for j I will substitute i, and for l I will substitute i+k, so when we do that we will get the expression as i+p-1 c i m+n-i-k-q c m-i-k i+k-i+q-p-1 c i+k-i, i=0 to m-k/m+n c m, so this term you can see becomes free from i so we can take it out, so that is become a q+k-p-1 c k/m+n c m, so the second one this is i+p-1 c i then this is m+n-i-k-q c m-i-k on this we apply this formula. Let me write that formula here we have a+k-j-1 c k-j, j=0 to k*b+j-1 c j that is=a+b+k-1 c k.

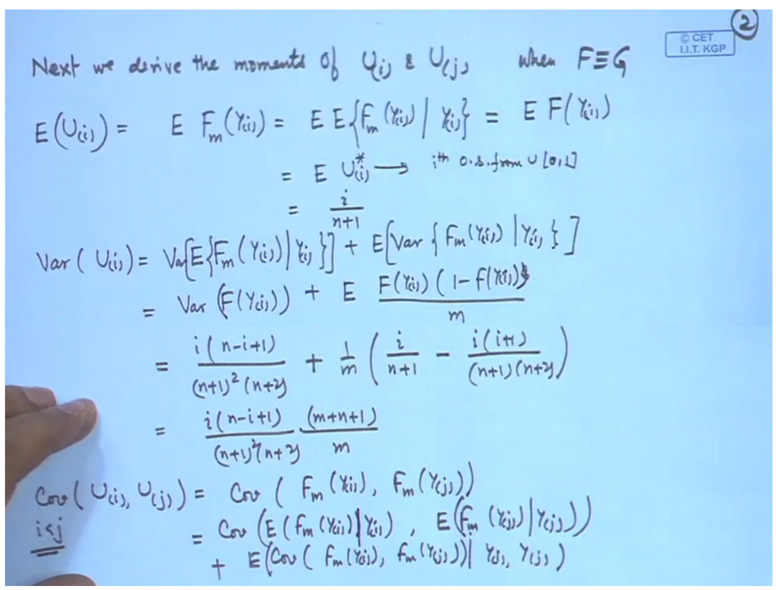
So this becomes simply= then q+k-p-1 c k m+n-k-q+p c m-k/m+n c m, but if you see this is also the probability of Uq-p=k.

(Refer Slide Time: 06:26)



We can look at the distribution that I derived for Ui here that is=j/m=m+n-i-j c m-j i+j-1 c j, so here if I put i=q-p and j=k then I get exactly this quantity, so this proves that the distributions of Uq-Up and Uq-p is same that means it is dependent upon the difference.

(Refer Slide Time: 07:14)



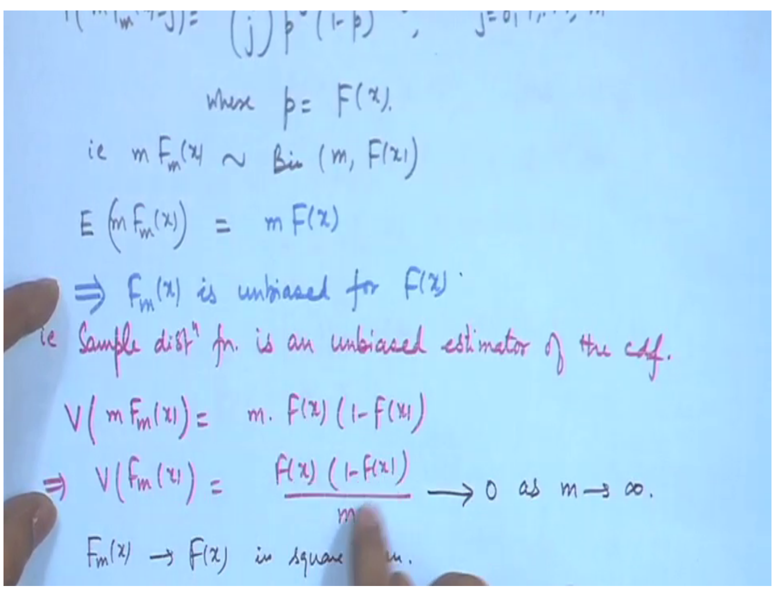
Now let us consider the moments of the Ui, Uj, we derive the moments of Ui and Uj of course we consider the case when F=G, when F !=G then only expressions can be written but here we can derive the exact values. So expectation of Ui that is expectation Fm of Yi, we can consider it as expectation of expectation Fm Yi given Yi okay. So what we are doing here that here the sample, the first sample and the second sample both are involved here.

So expectation means expectation with respect to both the distributions of F and G, so we do it iteratively firstly condition on Yi, so when we condition on Yi then it will become the expectation with respect to the X sample and after that we will do the second one. Now if Yi is fixed then this is the empirical distribution function and that we know it unbiased for the population cdf.

Now this quantity is nothing but expectation of Ui that is the Ui of the not this Ui, this Ui is actually the one which we derived as the ith order statistics from a uniform distribution, so this quantity okay let me call it Ui star that is the ith order statistics from the uniform 0,1. This we have already seen is=i/n+1, because for the order statistics from the uniform distribution we have seen that the it is a beta distribution with parameters i-1 with parameters i and n-i+1.

So the mean will become i/i+n-i+1 that is n+1, so it is = i/n+1. Similarly, if I consider variance of Ui we can consider it as a variance of expectation of Fm Yi given Yi+ expectation of variance Fm Yi given Yi. Once again this inner term will become F of Yi, and this term as we can see the distribution of Fm x that we have seen it is binomial distribution, we have derived it in the previous class.

(Refer Slide Time: 10:59)

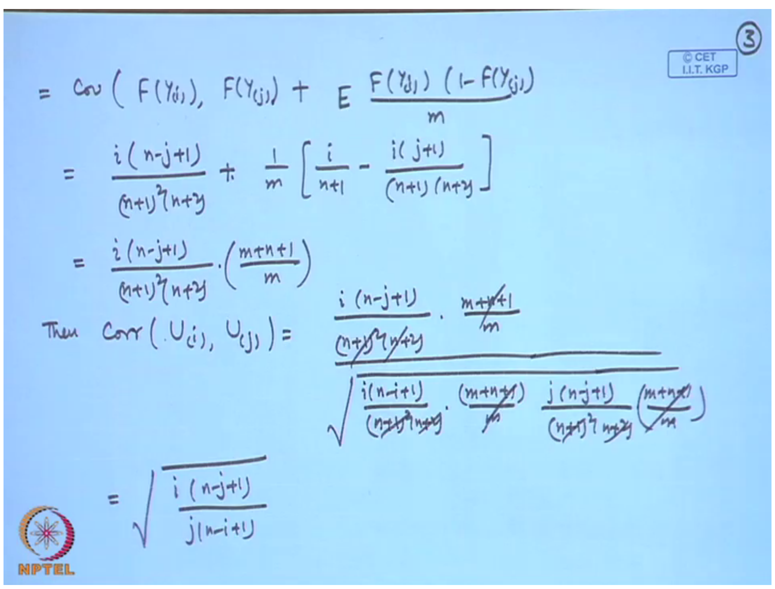


Let me just show the result once again that we had obtained the distribution of m Fm x is a binomial distribution, and from here the variance of Fm x was F x*1-F x/m, so if you use this then this will become =F of Yi*1-F of Yi/m, now this term is turning out to be the variance of the ith order statistics from the uniform distribution, now variance in beta distribution the formula is known, so it is becoming i*n-i+1/n+1 square*n+2.

In the second case, so 1/m we can keep outside, and this is becoming expectation of F of Yi that is i/n+1 and this will become the second moment that is i*i+1/n+1*n+2, so anyway this terms can be simplified and we get it as= i*n-i+1/n+1 square*n+2 m+n+1/m, so we are able to derive the mean and the variance of Ui, of course under the conditions when the 2 populations are the same. In a similar way we can try for the covariance term also.

Let us look at covariance between Ui and Uj, of course here I am taking i0=j so without loss of generality let us take say i<j, so that is covariance of Fm of Yi, Fm of Yj, so that is= once again we can write it as covariance of expectation+ expectation of covariance, so this one will become given Yi and this one will become expectation of Fm of Yj given Yj+ expectation of covariance between Fm Yi, Fm Yj this is given Yi, Yj.

(Refer Slide Time: 14:23)

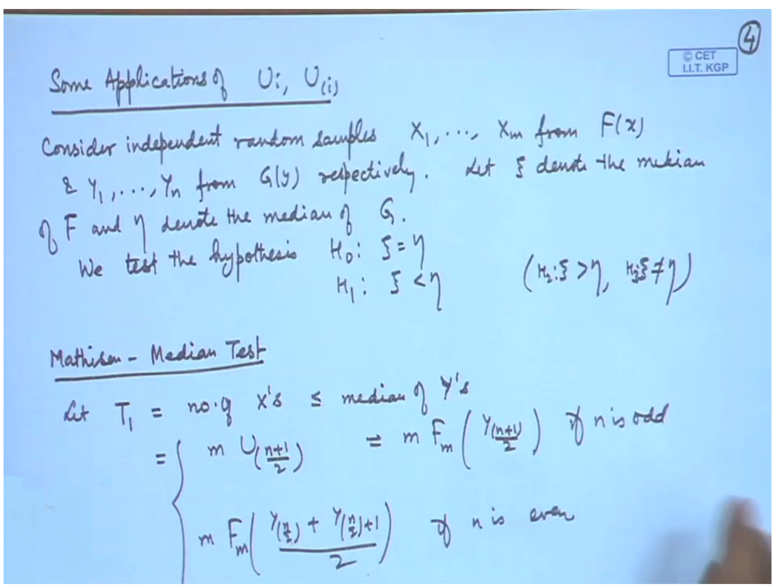


So this is= covariance between so you look at the first term here expectation of Fm Yi given Yi then this will become F of Yi, second term will become F of Yj, and the next one is the now this type of term also we have seen because if I look at 2 of them then this will become a F of Yi*1-F of Yj, so it will become expectation of this/m okay. So now these are the order statistics ith and jth from the uniform distribution.

So the formula for the formula for the covariance we have derived earlier i*n-j+1/n+1 square*n+2+ 1/m i/n+1-i*j+1/n+1*n+2, so after simplification this turns out to be i*n-j+1/n+1 square*n+2 m+n+1/m. Then let us also look at the coefficient of correlation between Ui and Uj and that is=covariance divided by the square root of the variances, so it is becoming i*n-j+1/n+1 square*n+2 m+n+1/m/.

So you can easily see that these terms will get cancelled out, and we left with here simply root of i*n-j+1/j* n-i+1, se we are able to completely determine the moment structure of the distributions of Ui and also joint distributions of Ui, Uj, and we have seen without the bracket and with the bracket also. Now we will see the applications of the certain 2 sample testing problems.

(Refer Slide Time: 18:10)



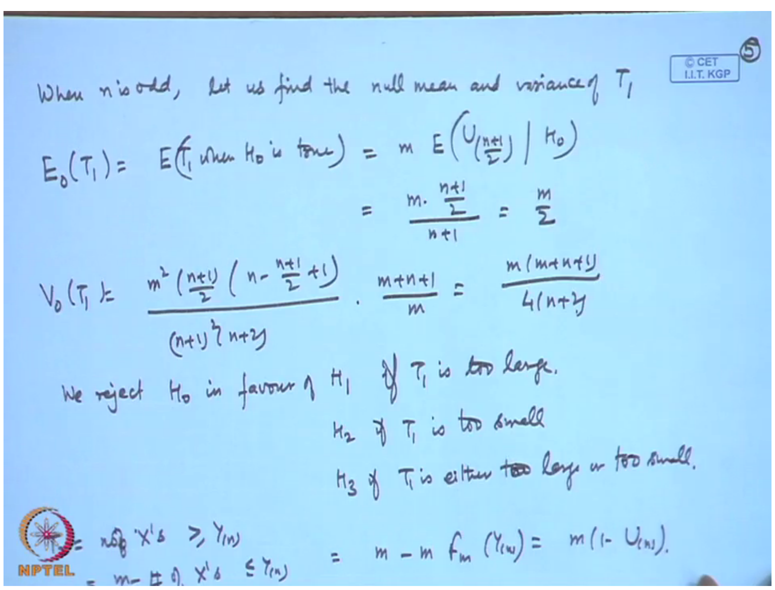
Some applications of these terms, so let us go to our original assumption that we are having independent random samples X1, X2, Xm say from F x, and Y1, Y2, Yn from say G y respectively. Let us assume that, let psi denote the median of F and say eta denote the median of G. So for the 2 distributions I am considering the medians, like in the classical parametric inference problems we assume the means to be mu 1, mu2 and variance is to be sigma 1 square, sigma 2 square.

Then our problem of interest is to test whether mu 1=mu 2 or sigma 1 square= sigma.2 square etc. So similarly, when we are considering the nonparametric situation we would be interested in testing whether psi=eta or psi<eta or psi>eta etc. so we can consider this hypothesis problem, psi=eta against say psi<eta, so it could be also psi>eta, psi !=eta all these type of testing problems can be considered, let me call it H2 and this as H3.

One of the first test is actually called Mathisan-Median test, in this test what we do? We define T1 to be the number of X's <= median of Y's, so actually depending upon what is the number of observations in Y second sample, so n could be odd, n could be even, so n is odd then you will have a unique medium, so how many x's are <that, that is exactly given by this Ui term, so we can consider it as m times that is U n+1/2 that is same as Fm Y n+1/2 if n is odd.

And it is=Y n/2+Y n/2+1/2, if n is even, now in the case of odd actually the distribution of this has already been worked out, we already know its mean and variance under the null distribution let me write that here.

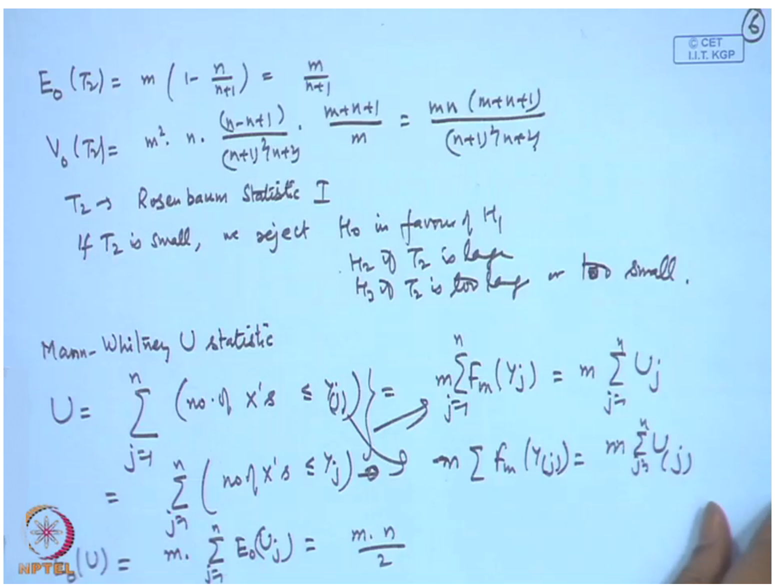
(Refer Slide Time: 22:28)



When n is odd let us find the null mean and variance of T1, so we consider expectations of T1 under H0 that is expectations of T1 when H0 is true that means when F=G, when F=G that means we can consider this= given H0 that=m times n+1/2/n+1 that=m/2, and variance of T1 of course it is m square n+1/2 n-n+1/2+1/n+1 square n+2 m+n+1/m so that=m*m+n+1/4*n+2, so this is the mean and this is the variance under the null hypothesis certainly we can consider the normalization actually.

And of course I know the distribution of T1 also here, so we can actually check whether it is too large or too small, we can also apply this Neyman-Pearson type thing that means we can consider the probability of type 1 error= alpha, and then we find the value of the critical point. So we can consider say, we reject H0 in favor of H1 if T1 is too large, and we can also define reverse of this that is in favor of H2 if T1 is too small, and in favor of H3 if T1 is either too large or too small.

(Refer Slide Time: 25:45)



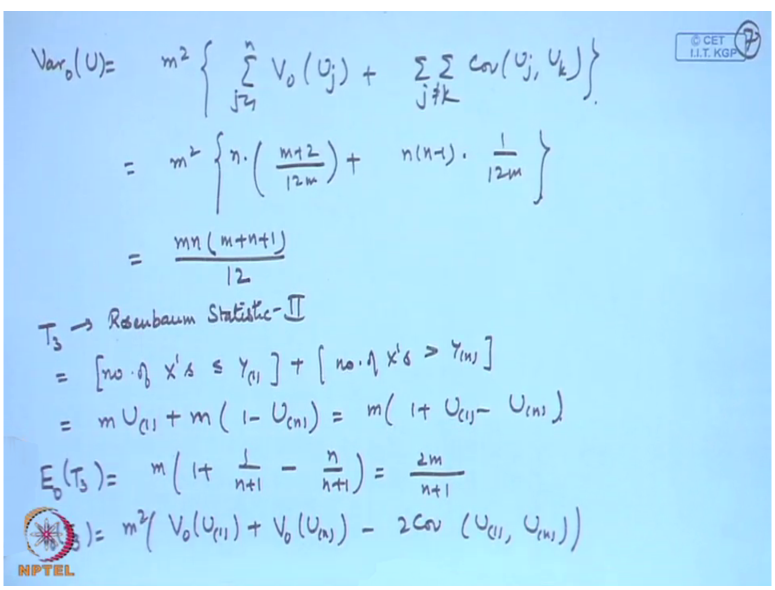
See we can actually consider T2 as the number of X's>=Yn that is m-number of X's which are <=Yn that=m-m Fm of Yn that=m times 1-Un that=m/n+1, variance of T2=m square n n-n+1/n+1 square*n+2 m+n+1/m that=mn*m+n+1/n+1 square*n+2, so one can use this also, this T2 is called Rosen Baum statistic 1. So we can actually do the testing based on this also, for example here it is greater, so if it is in the reverse way, we can consider that.

If T2 is small, then we reject H0 in favor of H1, so you can see it is the reverse of this, T1 is too large and here it is T2 is a small and reverse will happen against H2 if T2 is large, and in favor of H3 if T2 is too large or too small. The drawback with this 2 statistic that I defined that is Mathisan-median test is that this is based on median only, and in the Rosen Baum this is based on only the largest one.

Now one can think of using all of them, then that is called Mann-Whitney U statistic that is based on the summation of all such terms that is number of X's<=Yj and you sum from j=1 to n that that= summation j=1 to n, actually here if I put this or if I put this they are same, number of X's<=Yj, because when I am summing over all j's so when whether it is ordered or unordered both are same, so both of them I write like this.

Now the advantage of this term is that it is simply F times sigma Uj, and if I consider this form that is m times sigma Fm of Yj that=m times sigma of Uj, so this is known as the Mann Whitney U statistic we can consider the mean and variance under the null hypothesis, then it is simply=m times j=1 to n, we know that this was=1/2 so it is m n/2.

(Refer Slide Time: 29:39)



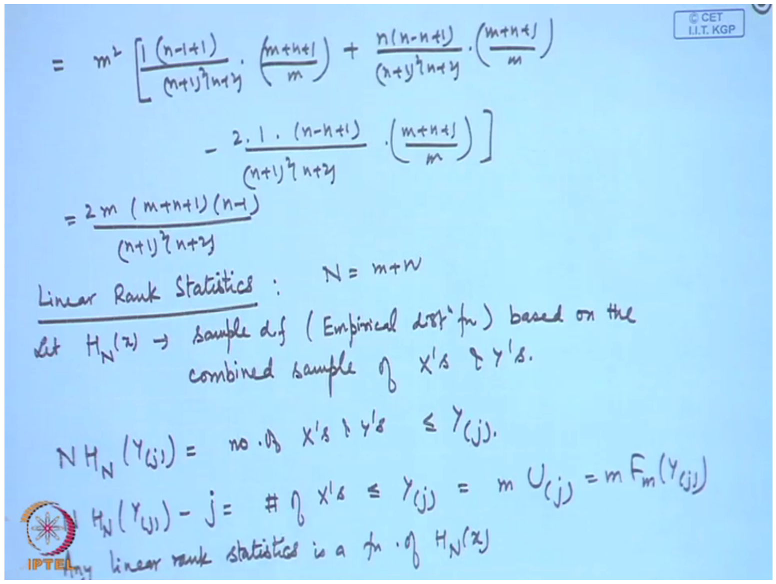
And if you look at the variance term, variance under the null hypothesis then it=m square summation variance of Uj+ double summation covariance of Uj, Uk j !=k, so that= m square and this is n and we actually derived this expressions in the previous lecture, let me just recall those expressions here. So you can see here expression for this was 1/2 and the expectations of this was derived, variance was derived as 2+m that is m+2/12m, so these expressions were derived.

And also the covariance term was derived here, the covariance terms of this was=1/12m, so we substitute all these terms here m+2/12m+n*n-1 1/12, so we can simplify this easily this=mn m+n+1/12. Once again if there are more number of X's which are<=Yj's then the median of X's will be >the median of Y's, so for large value of U will be rejecting H0 in favor of H1, and similarly for the other hypothesis.

So in a similar way we have another one which is called again Rosen Baum statistic 2, I had defined 1 that was T2 here, now I am defining T3 this I am defining as the number of X's <=Y1+ the number of X's>Yn, this one you can see it is more useful for the range that means if we are checking the variability of the scale parameter, then this will be more useful for example there are more X's which are outside of Y1 and Yn then certainly it means that the variability of X will be more than the variability of Y's okay.

So this can be written as m times U1+m times 1-Un=m*1+U1-Un, of course under this=m 1+1/n+1-n/n+1 that=2m/n+1, the variability of this can be calculated that will become m square*variance of this +variance of this and covariance term here. So this is becoming m square variance of U1+variance of Un- twice covariance of U1, Un.

(Refer Slide Time: 34:00)



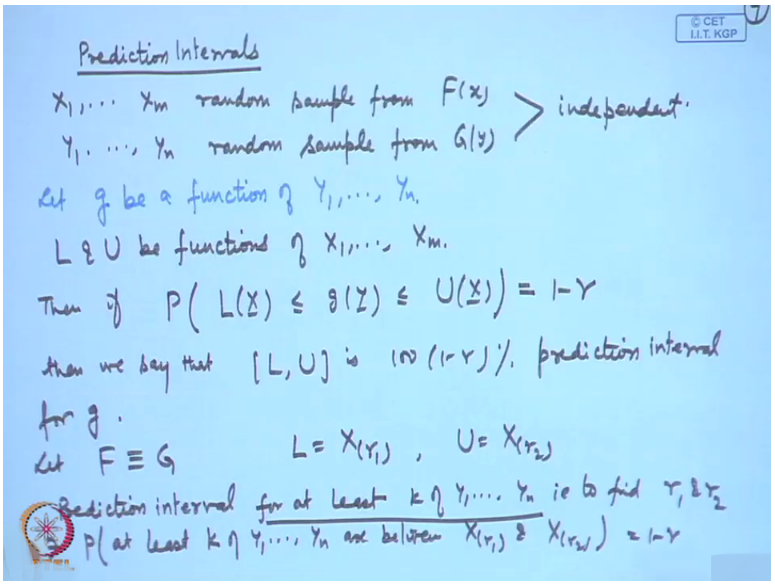
So we can substitute the expressions for this here and we will get m square n-1+1 since it is a first one, so 1*n-1+1/n+1 square*n+2 m+n+1/m+ n*n-n+1/n+1 square*n+2 m+n+1/m- 2*1*n-n+1/n+1 square*n+2 m+n+1/m, so this term is common and then you have n+1 square*n+2 that is also common, so you get m*m+n+1 and after simplification this term becomes 2m-2, so 2 times n-1/n+1 square*n+2, so once again we are able to obtain the null mean and variance of this thing.

And for testing purpose this is more as I mentioned it is more useful for the scale, so if we get a large value of T3 it means that the range of first sample is more than the range of the second sample, if T3 is too small then it means that the range of the second distribution is more than the range of the first distribution, so this can also be used for the testing for the range. In general, we can define Linear Rank statistics, 2 samples of sizes m and n are there let us consider the composite samples sizes=N.

And let us define HN x is the sample distribution function or empirical distribution function based on the combined sample of X's and Y's that means I consider all the observations together and then I looked at the order statistics of them not separately, it is not that I write X1, X2, Xm first and then Y1, Y2, Yn, I merge the 2 and then I considered the full ordering, and from that I define the empirical distribution that means if I consider N times HN of Yj then it is the number of X's and Y's which are<= say Y.

So if I consider N times HN of Yj-j then it= number of X's<=Yj that= m times Uj which is the one I defined earlier, that was m times Fm of Yj, so I establish a relationship between the empirical distribution of the first sample with the empirical distribution function of the combined sample in terms of the value of Yj. So in general we say that any linear rank statistics is a function of HN x. In other words, it is also a function of Fm Yj.

(Refer Slide Time: 38:22)



Next, we define Prediction Intervals, as before we have the 2 samples that is X1, X2, Xm is a random sample from F x, Y1, Y2, Yn is a random sample from G y and we assume that these are independently taken. Let us consider say g be a function of Y1, Y2, Yn, and L and U be functions of X1, X2, Xm. Then if we have the statement like that means probability of g y lying between 2 functions of X, if this=1-gamma.

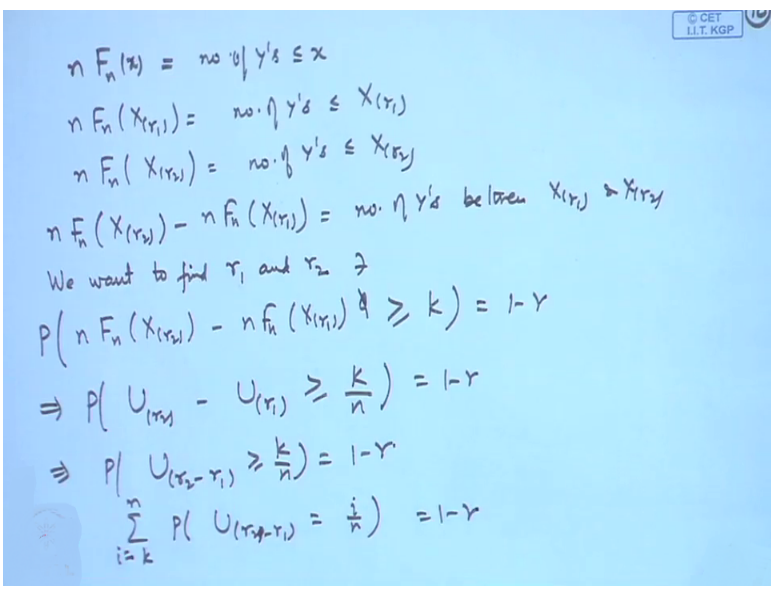
Then we say that L, U is 100 1-gamma% prediction interval for g, now let us consider the interpretation of this. See, we already seen the confidence intervals, in the confidence intervals the parametric term is to be fixed, so we find out the probability of 2 statistics including that parametric value =1-alpha, so that is called the 100 1-alpha% confidence interval. Now this there is a difference in the terminology.

So in general it could be that we may be using see there may be some relationship between the 2 distributions, so we may be using the relationship to predict the value of a function of second sample based on the values of the first sample, so this is basically based on the relationship that is available between the 2. Let us take the special case, let F be=G then how do we look at it? So let us consider say L to be the r1th order statistics and U to be say r2th order statistics.

So we consider say prediction interval for at least k of Y1, Y2, Yn that is to find r1 and r2 such that probability of at least k of Y1, Y2, Yn are between X r1 and X r2 that=1-gamma, at least k of Y1, Y2, Yn are between X r1 and X r2, and I want this probability to be 1-gamma, so we want prediction interval for this. So this is something like a location problem, but this is more general kind of location problem.

See if we consider parametric inference then we consider say if I am having 2 samples then we can consider theta 1-theta 2, so a confidence interval for that, we can consider confidence interval for linear combination of theta 1 and theta 2, we can also consider linear like sigma 1/sigma 2 some sort of parametric function, when we are not having that then we are considering order statistics here, and on the basis of that we are talking about some sort of location problem.

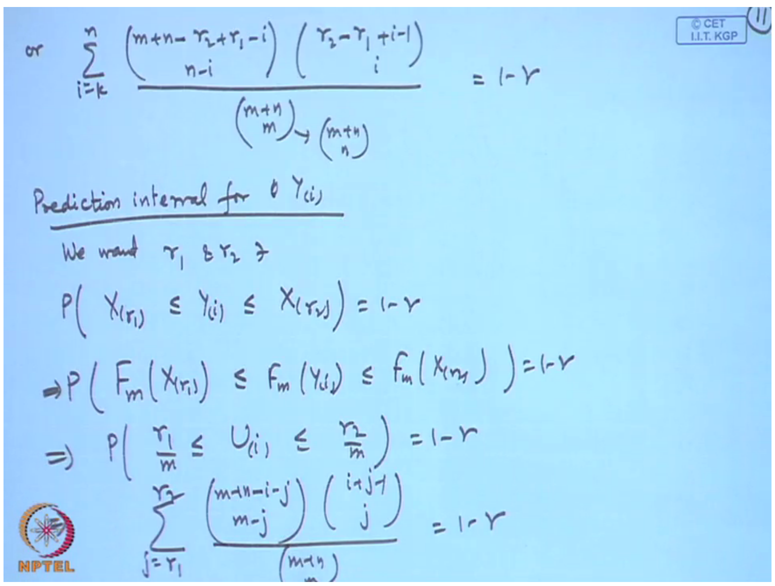
(Refer Slide Time: 43:40)



So let us consider then n times Fn x that= number of Y's<=x, then we can consider n times Fn of X r1 and that= number of Y's<=X r1, then we can also consider n times Fn of X r2 that= number of Y's<=X r2, so here Fn denoting the empirical distribution function based on the second sample rather than the first sample. So n times Fn X r2-n times Fn X r1 that is the number of Y's between X r1 and X r2.

So we want to find r1 and r2 such that probability of n times Fn X r2-n times Fn X r1 that is>=k that=1-gamma. Now the expressions for this is it is simply U r2-U r1>=k n times so I can divide here by n that=1-gamma, we have seen that the distribution of the difference is same as the U of Uq-Up as same distribution as Uq-p so that=1-gamma here, so this is nothing but sigma probability of U r2-r1= some i/n, where i=k to n that=1-is gamma.

(Refer Slide Time: 46:22)

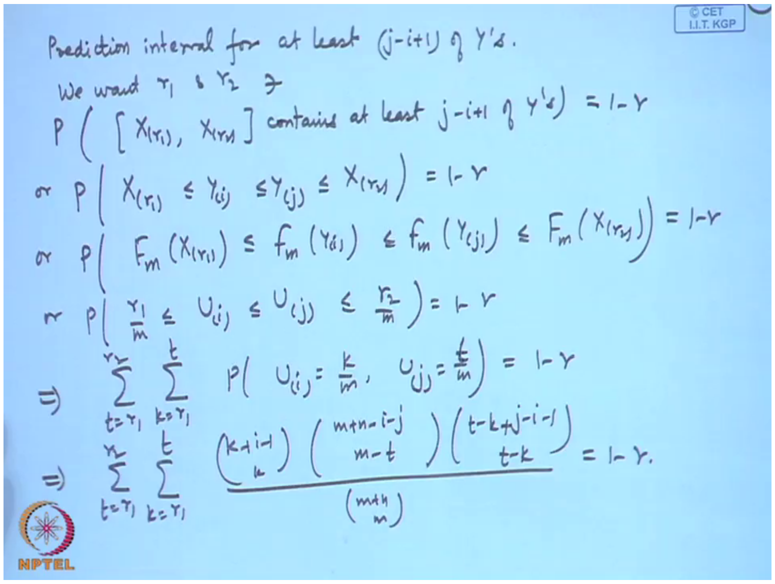


The distribution of this is known to us so we substitute this value it is turning out to be simply=i=k to n m+n-r2+r2-i c n-i r2-r1+i-1 c i/m+n c m that=1-gamma, so from the tables of the factorials of hypergeometric distribution, then this can be calculated here, so basically this is hypergeometric here, this is m+n c m that is also same as m+n c n so both are same therefore, this is the proper hypergeometric term here.

In a similar way I can consider prediction interval ith order statistics from the second sample say that means we want now r1 and r2 such that probability of X r1<=Yi<=X r2=1-gamma, let us take here empirical distribution function, then this is nothing but r1/m<=Ui<=r2/m that=1-

gamma, the distribution of this is known so it is simply reducing to simply m=n-i-j m-j i+j-1 c j/m+n c m from r1 to r2.

(Refer Slide Time: 48:56)



Similarly, we can consider prediction interval for at least j-i+1 of Y's, so we want r1 and r2 such that probability of the interval X r1 to X r2 contains at least j-i+1 of Y's that=1-gamma that is probability that X r1<=Yi<=Yj<= X r2 it=1-gamma, so Fm of X r1<=Fm of Yi<=Fm of Yj<=Fm of X r2, where Fm is denoting the empirical distribution function of X sample, so this is nothing but r1/m<=Ui<=Uj<=r2/m that=1-the gamma.

So we can consider it as a double summation probability of Ui=some k/m, Uj= some t/m, where t=say r1 to r2 and k=r1 to t that=1-gamma, the joint distribution of Ui and Uj has been obtained, so this is nothing k+i-1 c k m+n-t-j c m-t t-k+j-1-1 c t-k/m+n c m t=say r1 to r2 and k=r1 to t that=1-gamma. So this is the bivariate hypergeometric type of term here, and once again from the tables of factorials or the tables of bivariate hypergeometric this can be evaluated.

I have given here some elementary applications of this statistics which are based on the empirical distribution function. So we have the, let me summarize I have given here several applications here, one is to find out the test for the comparison of the medians, I also mentioned that one of them can be used for the scales also that means for the range. We have also defined what is known as the prediction interval.

So when we have the 2 samples, then what is the use of prediction interval? That means based on 2 values of the first sample or order statistics from the first sample, I can predict something about the value of the random variable, so basically we are talking about the probability that this much probability we can say that is value will lie in this interval. So in this one I discussed several of them, for example what is the prediction interval for Y's, i's, what is the prediction interval for at least j-i+1 of Y's, what is the prediction interval for at least k of Y1, Y2, Yn etc.

So these different type of formulae, they are useful for various type of nonparametric location or scalar matrix problems. We will talk more about this in the next lecture. Now there is another problem that is very commonly used by all the people working in different areas of Science and Engineering, that is given the data how to decide that which particular distributional model will be useful. One of the most popular applications or test for this is known as the Chi square test goodness of fit, which is originally given by Karl Pearson, so I will talk about that.

Later on another powerful test was given by Kolmogorov and Smirnov, so in the next lecture I will be discussing both these tests for fitting of a distribution, so they are called goodness of fit test, and as you can see that the structure of these test is extremely simple, they are not dependent upon what is the original distributional model only what assumption we are making it is dependent upon that. So in the next lecture will be discussing these.