Statistical Methods for Scientists and Engineers

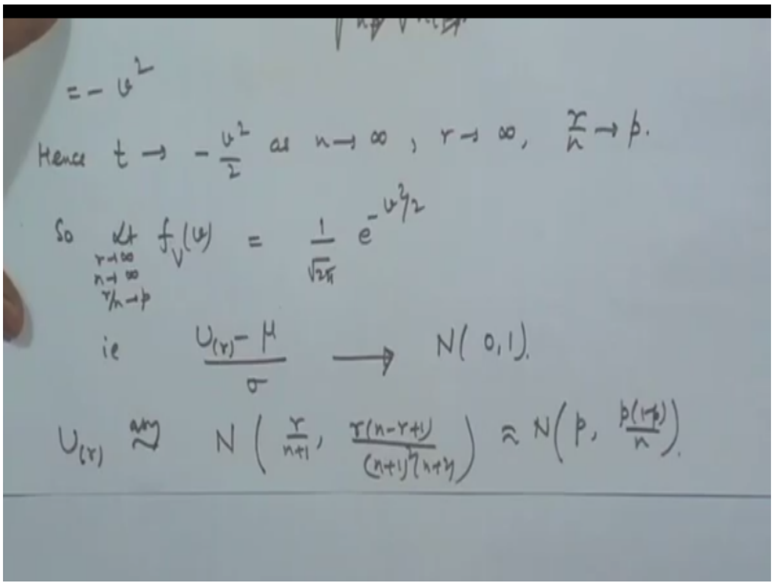
Prof. Somesh Kumar
Department of Mathematics
Indian Institute of Technology - Kharagpur

Lecture - 31
Non parametric Methods - IV

Yesterday, we have been discussing how to derive the asymptotic distributions of the order statistics. We found out the asymptotic distribution of the rth order statistics under 2 conditions, 1 was when r is kept fixed but the sample size n tends to infinity. Under this condition, the rth order statistics from the uniform distribution has a gamma distribution.

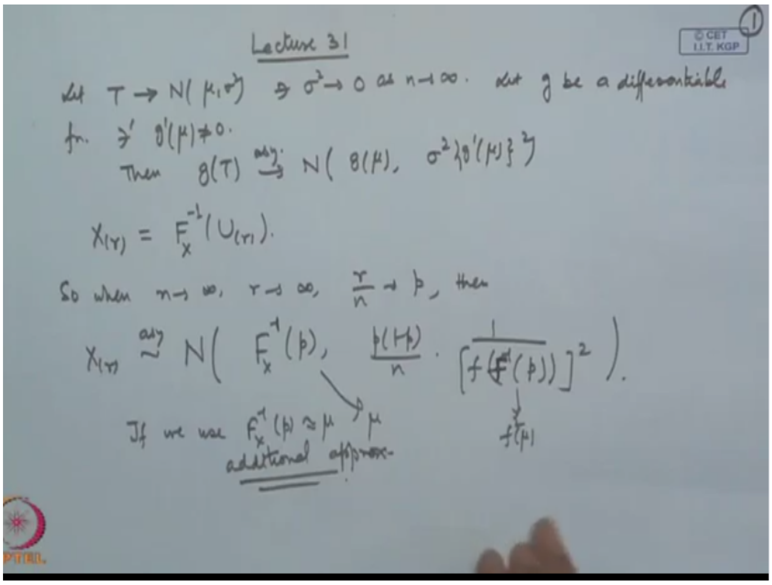
And therefore we can find out in terms of f the asymptotic distribution of rth order statistics from any distribution. Then the second condition was that when r tends to infinity and n tends to infinity, but r/n tends to p. That means basically we are fixing the position in a fixed proportion for example median it could be quantile etc.

(Refer Slide Time: 01:16)



In that case firstly when we consider from the uniform distribution then Ur is having asymptotically N p, p*1-p/n where r/n tends to p, r tends to infinity and n tends to infinity. Then this is the result that we had throughout. Now let me apply the following result if the asymptotic distribution of certain sequence of random variable is known then if I consider a function.

(Refer Slide Time: 01:52)



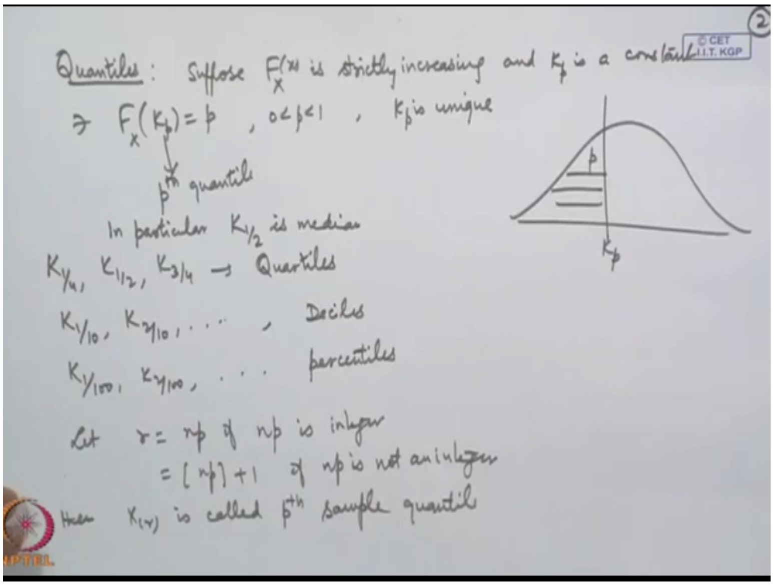
So we have the following result that let T have asymptotically normal mu, sigma square distribution and of course we assume that sigma square tends to 0 as n tends to infinity. This is an additional assumption and let g be a differentiable function such that g prime mu is not 0. Then the asymptotic distribution of gT that is g of T is asymptotically normal g of mu and sigma square*g prime mu square.

This is asymptotically okay. Now we have derived the asymptotic distribution of Ur here and we use the relation that Xr=Fx inverse Ur if we use this then we get. So when n tends to infinity r tends to infinity such that r/n tends to p then the asymptotic distribution of Xr is normal Fx inverse p, p*1-p/n 1/f F inverse p whole square. Sometimes one additional this thing is used if we use say Fx inverse p is say mu.

Then this is becoming mu and here I will get f of mu square, so that is an additional approximation okay. So this is the discussion about the asymptotic distribution of the order statistics and we have derived under 2 conditions. Now I discuss 1 next concept that is of quantiles. I already mentioned that in the case of non-parametric statistics, it is more convenient to handle the positions on the distribution.

Because capital F is there. We are not assuming functional form but capital F is there so making some you can say inferences based on quantiles, positioning etc is much more convenient, so let us look at this now.

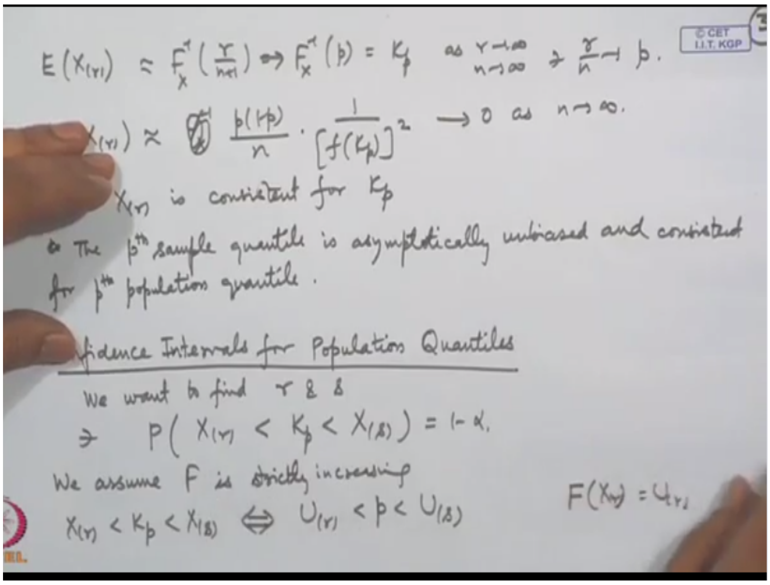
(Refer Slide Time: 04:49)



The concept of quantiles so suppose F is strictly increasing and Kp is a constant such that F of Kp=p then this is called and of course Kp is unique then we call Kp as pth quantile, in the case of continuous distribution you can think like this. So if the probability up to this point is p then this point is called Kp. So in particular you have K1/2 is median, K1/4, K1/2, K3/4 these are called quantiles.

K1/10, K2/10 and so on they are called deciles. K1/100, K2/100 etc these are called percentiles. So in general we are dealing with any type of quantile. We can consider here suppose r=np if np is integer and it is equal to integral part of np+1 if np is not an integer. Then Xr is called pth sample quantile. So it is the same thing for example if you consider p=1/2 then n/2 and n/2 integral part+1 so that is called the median for example.

(Refer Slide Time: 07:19)



Expectation of Xr that is approximately Fx inverse r/n+1 that is Fx inverse p=Kp as r tends to infinity, n tends to infinity such that r/n tends to p. Similarly, variance of Xr that is approximately I am considering the first hand approximations that we derived yesterday p*1-p/n 1/f of Kp square. So this will go to 0 as n tends to infinity. So Xr see it is asymptotically unbiased and the variance is going to 0.

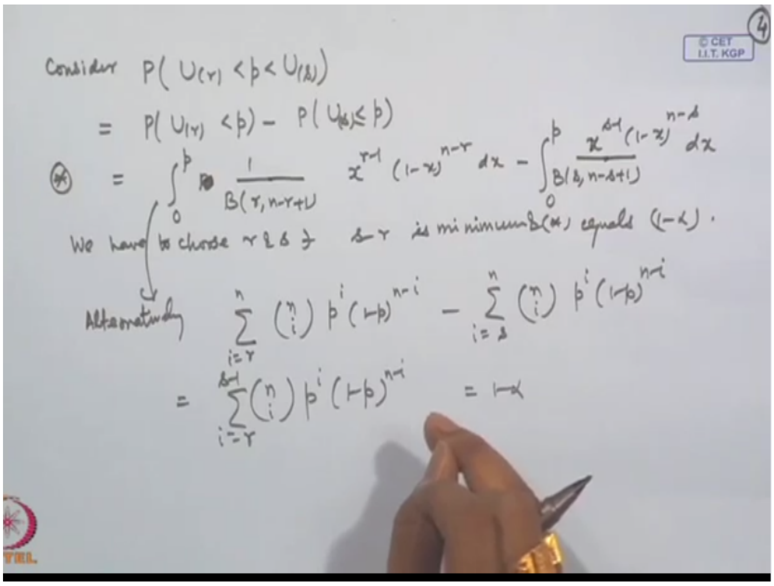
So Xr is consistent for Kp. That means the pth sample quantile is a consistent estimator of the pth population quantile. So we approved one result here. So like see when we have the known form of the distributions generally we consider mean so for the population mean we consider sample mean. We approve that it is unbiased and consistent estimator under of course certain conditions.

Then we also have the variance then for that we consider the sample variance, we have it as unbiased and consistent again under some mild conditions. Similarly, in the non-parametric case when we are considering quantiles then the corresponding sample quantile can be considered as a consistent estimator and it is asymptotically unbiased. So we can make this statement.

That is the pth sample quantile is asymptotically unbiased and consistent for pth population quantile. So we have something to like you can say to start with. Now we consider confidence intervals for population quantiles. Now already because if you consider say parametric form so when we consider the population mean then we start with the sample mean.

The reason is there, it is unbiased and consistent, but here for the quantile we have considered the sample quantile that means a natural choice would be to consider order statistics. So we can pose the problem like this. We want to find r and s such that probability of $Xr<Kp<Xs=1-$alpha. Now if we are assuming that F is strictly increasing then $Xr<Kp<Xs$ is equivalent to $Ur<p<Us$ where you are making the transformation by taking F here F of Xr=Ur.

(Refer Slide Time: 11:40)



So now let us consider this probability. So consider probability of Ur<p<Us, so we can write it as probability of Ur<p-probability of Us <= p but this is same as = p because these are continuous distributions. Therefore, keeping equality or not will not make any difference. Now the rth order statistics from the uniform distribution that has a beta r, n-r+1 distribution.

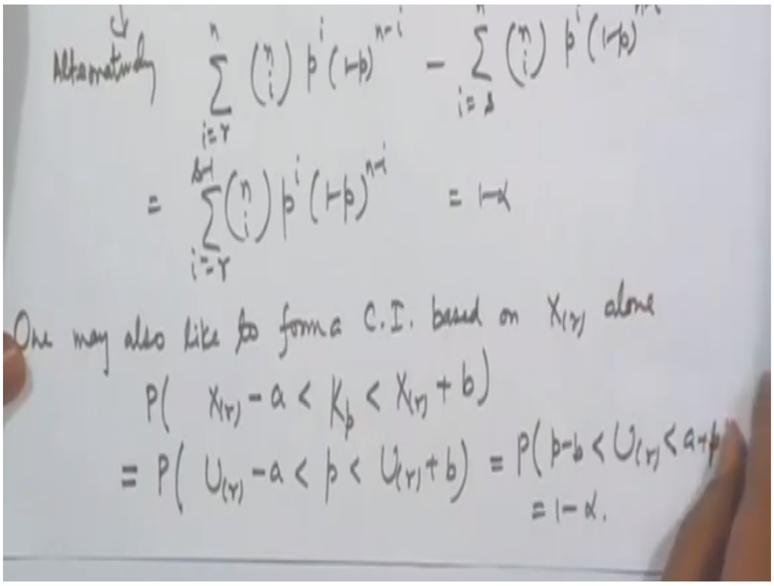
And similarly the sth order statistics will have a beta distribution with s and n-s+1. So this can be easily written like this 0 to p 1/B r, n-r+1 that is beta function x to the power r-1 1-x to the power n-r dx-0 to p 1/B s, n-s+1 x to the power s-1 1-x to the power n-s dx. Well we have to determine r and s so these both are incomplete beta functions, so we have to choose r and s such that s-r is minimum.

And if I call this quantity star and star=1-alpha. One can use the formula at the numerical integration for the incomplete beta function and we can calculate it. Another alternative is to write this incomplete beta function as the binomial expansions. So these things can also be written as one can also write alternatively this as sigma i=r to n n C I p to the power i 1-p n-i and this one becomes sigma i=s to n n C i p to the power i 1-p to the power n-i.

So that is = sigma i=r2 s-1 n C i p to the power i 1-p to the power n-i. So once again from the tables of the binomial distribution, we have to s and r such that s-r is minimum and this probability=1-alpha. Of course since now we have made it a discrete it is not necessary that we will achieve this early so it may be >= also. However, this methodology is quite clear.

Sometimes one may think of obtaining a confidence interval based on Xr itself where r is the pth sample quantile. Now in that case of course the distribution of Xr is known. So let me just mention that point also.

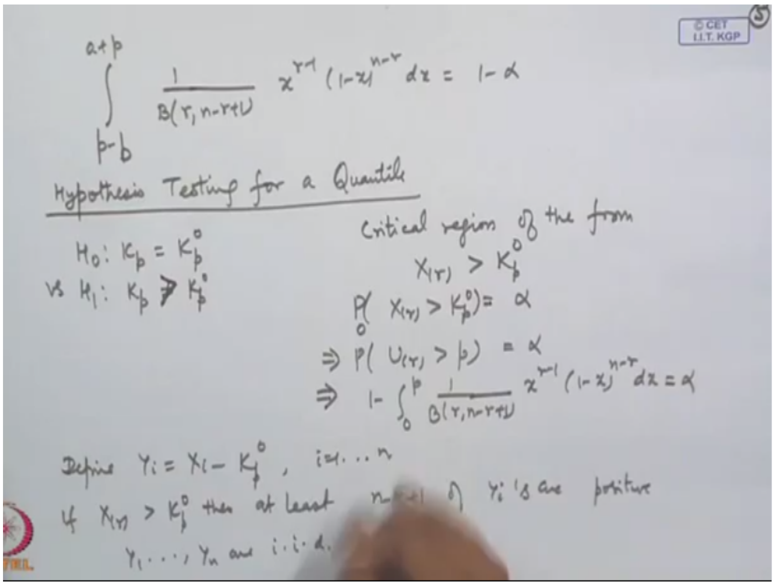
(Refer Slide Time: 15:25)



One may also like to form a confidence interval based on Xr alone. Of course, the distribution of Xr is not necessarily symmetric. Now you have 2 things, one is that one can basically if we consider say $Xr-a<Kp<Xr+$some b then this is equivalent to probability of $Ur-a<p<Ur+b$. So you can write it as after simplification see $Ur>p-a$ and Ur is also$<p+a$, so you can write it as probability of $Ur>p-b<Ur<a+p$ okay.

So we have to choose a and b such that this is = 1-alpha. The distribution of Ur is known.

(Refer Slide Time: 16:49)



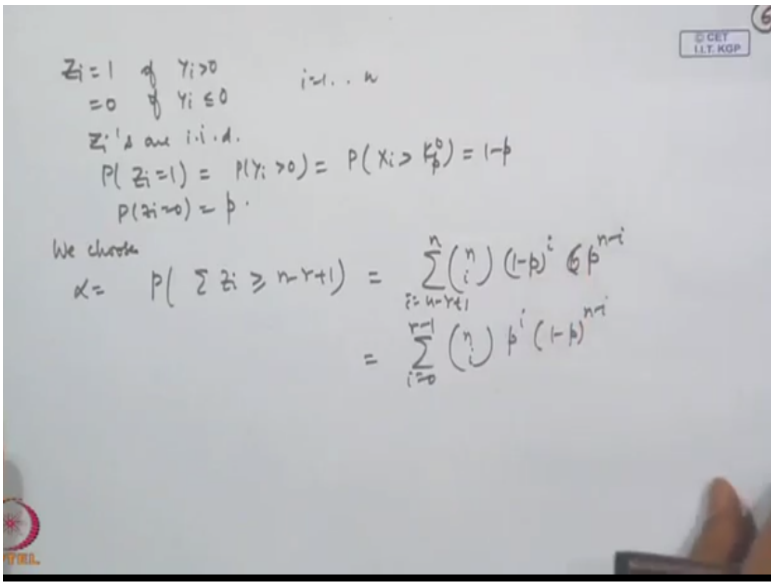
And therefore this is nothing but integral from p-b to a+p 1/B r, n-r+1 x to the power r-1 1-x to the power n-r dx. So basically you choose 2 values a and b of course here since the distribution is not necessarily symmetric. Actually, it is symmetric about 1/2 but we cannot actually consider 1/2-something to 1/2 because this is Kp.

So Kp is not necessary in the middle. If we are finding for middle, then it is a different matter but that is not so. Therefore, we take arbitrary choice but once again one can use the tables of the incomplete beta function to calculate this value. So let me move to the hypothesis testing now. Hypothesis testing for a quantile, so we formulate a hypothesis say H0=Kp=Kp0 against say Kp is not equal to or say greater than Kp0.

So we can have various like Kp>Kp0, Kp<Kp0, Kp != Kp0, 3 types of alternatives maybe there. Now as we have seen this Xr where r is the np or np+1 integral part is a consistent estimator for Kp so we can consider critical region of the form Xr greater than some constant. So let us say put say simply this one, but we should have the condition that probability of this region=alpha that is under H0.

Now this is equivalent to probability of Ur>p=alpha that means you are saying 1-0 to p, 1/B r, n-r+1 x to the power r-1 1-x to the power n-r dx=alpha. So one can easily find it out and I mean this is doable thing. We will give an alternative formulation of this. Let us consider say Yi=Xi-Kp0 for i=1 to n. Now if Xr<Kp0 then at least n-r+1 of Yi's are positive and Y1, Y2, Yn they are i i d.

(Refer Slide Time: 20:44)

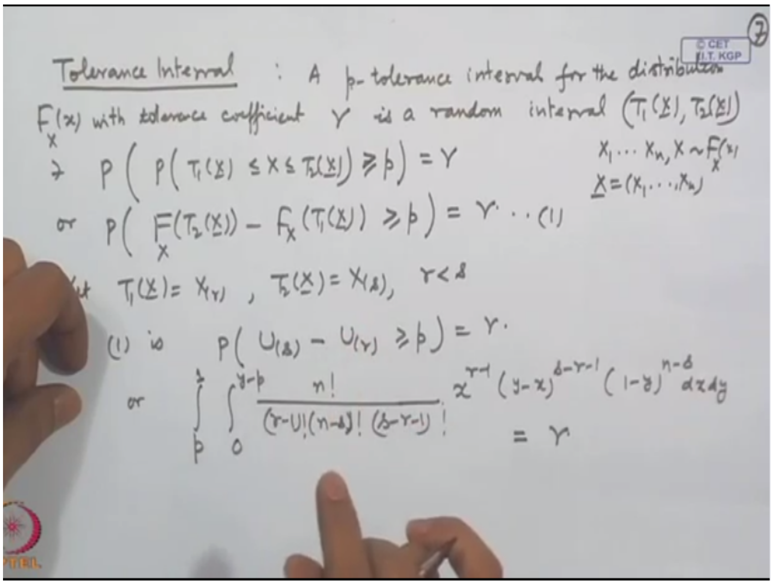


So we can further define, let us define say Zi=1 if Yi is positive, it is = 0 if Yi is <= 0 for i=1 to n. Then Zi's are i i d and we consider probability of Zi=1=probability of Yi>0. That is probability of Xi>Kp0=1-p and of course probability of Zi=0 that will become = p. So alpha=probability sigma Zi >= n-r+1.

Now this is simply coming from the binomial n C i 1-p to the power i p to the power n-i for i=n-r+1 to n. Of course, you can change here this i to j-n so this will become here = n C i p to the power i 1-p to the power n-i i=0 to r-1. So in either way one can actually obtain this here. One may think alternatively like we can consider Xr>c and then we choose c such that this probability=alpha.

So that can be another way of looking at this here. Then next we define what is known as tolerance intervals. See what we have discussed here is confidence interval. Now we are talking about tolerance interval.

(Refer Slide Time: 22:42)



So let me give a definition of what is known as tolerance interval. A p tolerance interval for the distribution Fx with so now note here I am having p tolerance interval and then I am introducing another one tolerance coefficient like you have confidence coefficient, this I call gamma this is a random interval T1 X to T2 X such that probability of T1 X <= X <= T2 X is >= p is equal to gamma.

See here we are having this X1, X2, Xn and X they all have the same cdf Fx here and this X is actually X1, X2, Xn here. So we want to find 2 statistics T1 and T2 such that the probability of X lying between these is >= p. Now if you look at the first statement in the first one we will consider the distribution of X here and in the second one, we will consider the distribution of this X here or we can do it in the reverse also.

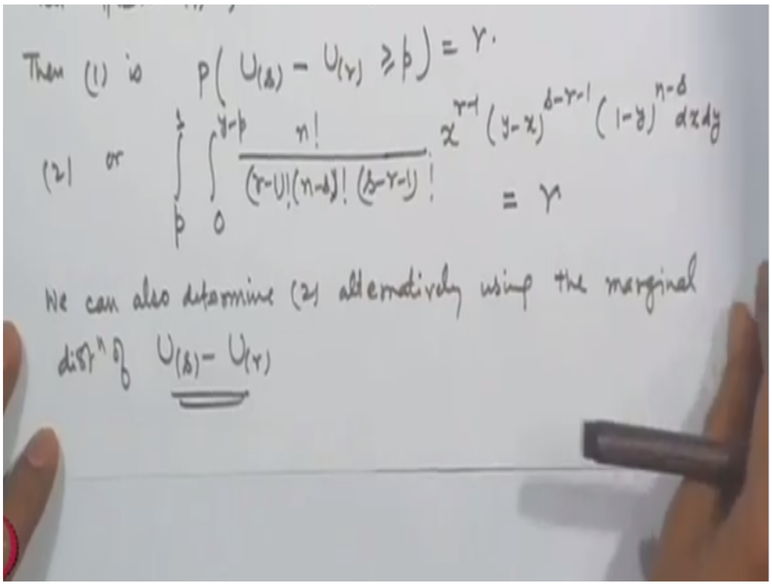
Firstly, we will consider the distribution of X and then we consider this. We can also write it as see if you write it in the terms of cdf then X <= something can be written as F of T2 X-F of T1 X, of course we assume continuous distribution here. This is >= p=gamma. If we replace this T1 and T2 by some order statistics say I take them to be rth and sth where r<s.

Then this is simply reducing to this condition let me call it 1. This is simply becoming probability of Us-Ur >= p=gamma. Now the distribution of the rth and sth statistics from the uniform distribution is very well known, so one can use this. If we write in the terms of joint distribution, then this is becoming n factorial/r-1 factorial n-s factorial s-r-1 factorial and then you have x to the power r-1 y-x to the power s-r-1 1-y to the power n-s dx dy.

Firstly, when we do is respect to x then we can go from 0 to y-p and then for this for y it can be from p to 1. So what we want to say that this should be = alpha. So this is a bivariate integral. Of course, one can also write down the direct distribution of Us-Ur also. In fact, I have earlier derived the distribution of the range from the order statistics of the uniform distribution.

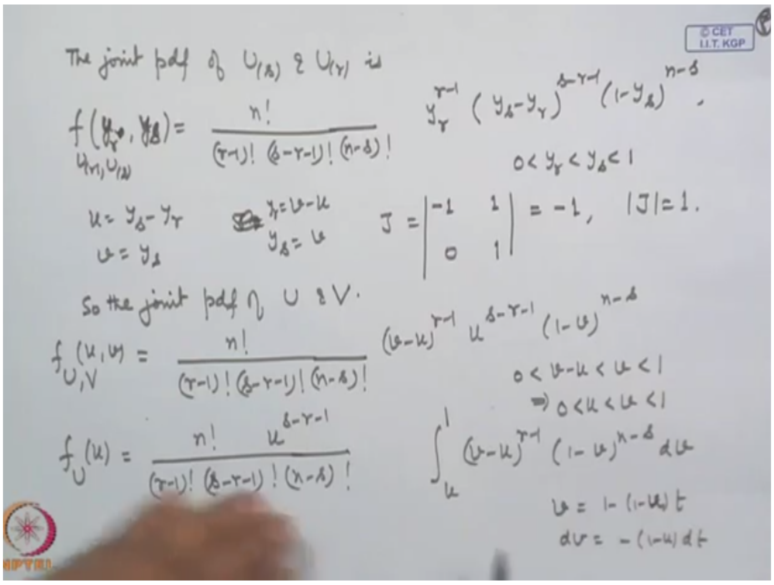
That was coming in a closed form because we are able to evaluate the integrals. In this case also this can be done let me just demonstrate that this can be done.

(Refer Slide Time: 26:39)



Let me call it 2, we can also determine 2 alternatively using the marginal distribution of Us-Ur so let me do this. We have this joint distribution, now you make the transformation here, let me write this joint distribution again.

(Refer Slide Time: 27:18)

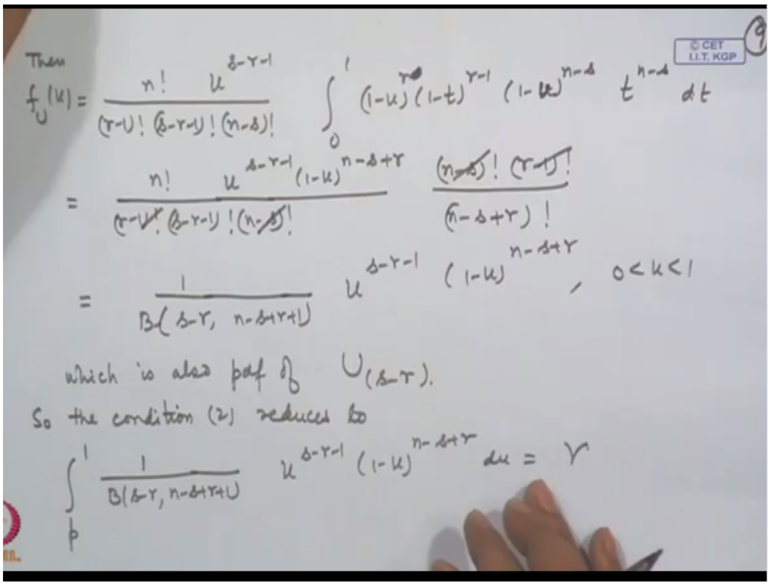


The joint pdf of Us and Ur that is given by F of yr, ys=n factorial/r-1 factorial s-r-1 factorial n-s factorial yr to the power r-1 ys-yr to the power s-r-1 1-ys to the power n-s 0<yr<ys<1. In this I make the transformation U=ys-yr and let v be ys itself. So the inverse transformation here is v-u that is yr=v-u and ys=v so if we calculate the Jacobian, del yr/del u that is -1, del yr/del v is +1, del yr/del u is 0, del ys/del v is 1.

That is = -1 so modulus of the Jacobian=1. So the joint probability density function of U and V=n factorial/r-1 factorial s-r-1 factorial n-s factorial. This will become v-u to the power r-1 u to the power s-r-1 1-v to the power n-s and 0<now yr is v-u<v<1 which is equivalent to saying see u<v, v is of course less than 1 and u is of course greater than 0 because ys>yr so this region can be written like this also.

So we ultimately need the distribution of u that is ys-yr so that is becoming n factorial/r-1 factorial s-r-1 factorial n-s factorial when we integrate with respect to u this term I can keep outside, u to the power s-r-1 and v-u to the power r-1 1-v to the power n-s dv from u to 1. Here we make the transformation say v=1-1-v*t so you are getting then 1-ut so dv=-1-u dt. When v=u then t is becoming = 1 and when v=1 t is becoming 0.

(Refer Slide Time: 31:18)



So this integral is transformed to fu u that is equal to this all n factorial/r-1 factorial s-r-1 factorial n-s factorial u to the power s-r-1 0 to 1. Now v-u will become 1-u*1-t, so this to the power r-1 and this to the power r-1 and there is 1-u again so this will go away and then we are having 1-u to the power n-s*t to the power n-s dt. So this is = n factorial/r-1 factorial s-r-1 factorial n-s factorial.

Now let us look at the terms that we are getting u to the power s-r-1 then 1-u to the power r and 1-u to the power n-s so this you combine so it is becoming n-s+r okay. Then you have a beta integral t to the power n-s*1-t to the power r-1 so it is becoming n-s factorial r-1 factorial/n-s+r factorial. So this term cancels out, this cancels out and you are left with 1/beta s-r, n-s+r+1 u to the power s-r-1 1-u to the power n-s+r, which is also the pdf of Us-r.

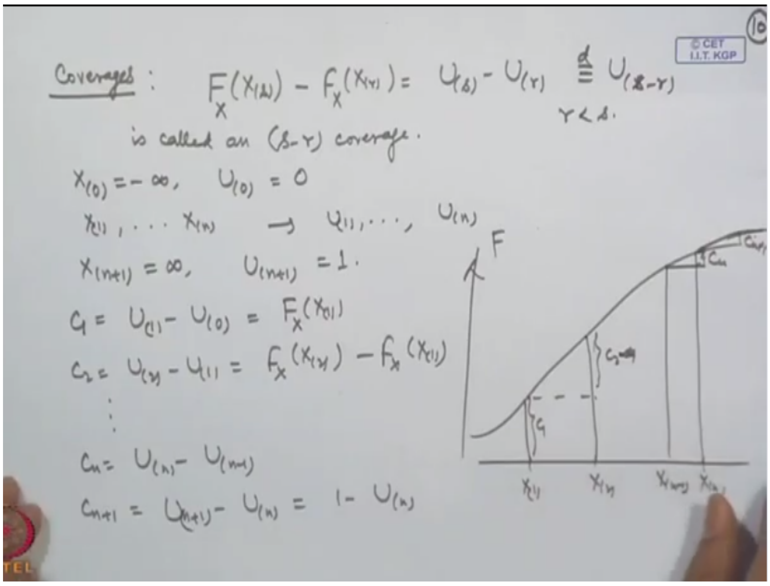
So this is interesting, we have obtained the distribution of Us-Ur, which is turning out to be this. It is the same as the distribution of Us-r that means in the sampling from uniform distribution on the interval 0 to 1 if I consider the distribution of the difference of 2 order statistics then the difference value so for example if I am looking at say U4-U2 then the difference is 2.

So if I consider the distribution of U2 it is the same as the distribution of U4-U2 so that is the very interesting phenomena about the order statistics from the uniform distribution. So if we look at this condition here then that I wrote that this double integral must be = gamma, now Us-Ur is a beta distribution then it is actually simply becoming a condition in a beta integral or incomplete beta function.

So the condition 2 reduces to p to 1 that is 1/beta s-r, n-s+r+1 u to the power s-r-1 1-u to the power n-s+r du=gamma or if you consider 0 to p then this is becoming 1-gamma. So this condition you can see it is similar to that for obtaining the confidence interval, but these are called tolerance interval, the reason being that I am considering the probability a particular confidence coefficient of X itself to be = gamma.

So this is a different thing than the usual confidence interval, but ultimately the solution is coming in terms of that and that role of p is coming here and gamma or you can say 1-gamma turns out to be the corresponding confidence coefficient here.

(Refer Slide Time: 36:17)



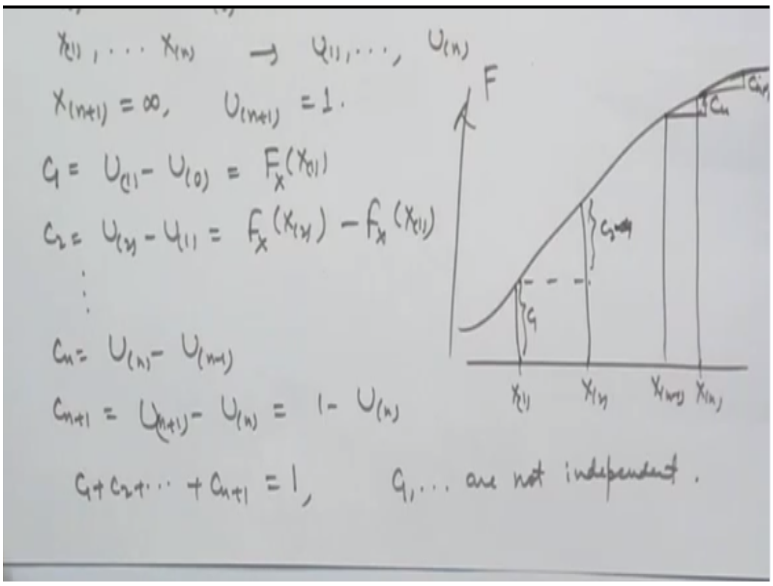
Next we consider the concept of coverages. What are coverages? Let us consider see we are having this Fx Xs-Fx Xr=Us-Ur okay. Of course, this we showed it is having the same distribution as Us-r okay where r<s, but if we are looking at this distributional thing then this is actually an s-r coverage that means it is covering the probability from the rth order statistics to the sth order statistics.

So this is called an s-r coverage. Let me define all coverages. So for example you may consider here say X0=-infinity correspondingly U0=0 and you have other order statistics where X1, X2, Xn corresponding to that you have U1, U2, Un then you consider Xn+1=infinity so the corresponding Un+1=1. So now we define the first coverage C1=U1-U0 okay.

So if you see in terms of this it is actually F of X1 simply because the second is 0 okay. Then C2=U2-U1=F X2-FX1 and so on. Cn=Un-Un-1, Cn+1=Un+1-Un=1-Un. If I consider say suppose this is cdf okay, this is F and these are the points say X1, X2 and so on Xn. Then Fx1 that is this is the C1 then Fx2-Fx1 that is this quantity will become C2 like that. Let us consider say Xn-1.

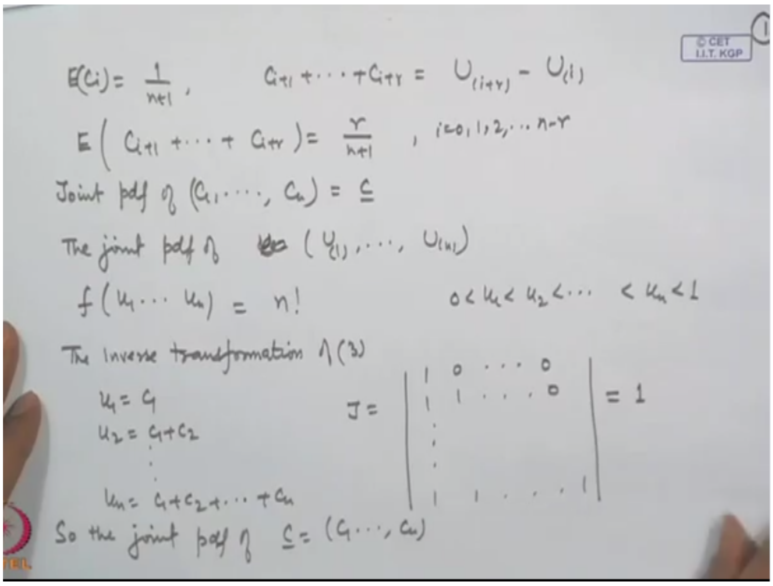
So this will become Cn and the last one is after this that means whatever remaining height is there that is = Cn+1. So basically what we are saying is that we are covering cdf that is the ordinate of the cdf that is why this is called the coverages, but it is based on the order statistics so they are no independent.

(Refer Slide Time: 39:42)



And another thing is that if you consider C1+C2+Cn+1=1, C1, C2 etc they are not independent. Since these are order statistics from the uniform distribution we know the moments here. For example, expectation of Ur is r/n+1. So in general then I can calculate all these differences will yield the expectation=1/n+1.

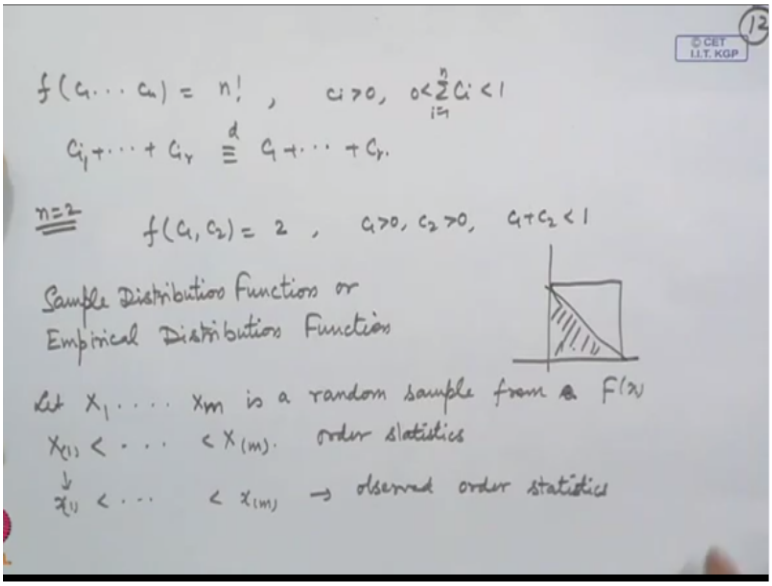
(Refer Slide Time: 40:31)



That is we are having expectation of ci=1/n+1 if I consider say ci+1 up to ci+r then we are considering the coverage from i to i+r here and if I consider say ci+1+up to ci+r then that is becoming = r/n+1 for i=0, 1, 2, etc up to n-r. Let us also talk about the joint distribution of c1, c2, cn. Joint pdf of c1, c2, cn okay so let me call it c vector. They are transformations from the Ui's and we know the joint distribution of Ui's.

The joint pdf of that is u1, u2, un that is f of that is n factorial 0<u1<u2<un<1 and the transformation that we are having here let me call it say 3 here the inverse transformation of 3 that is given by u1=c2, u2=c1+c2 and so on. u1=c1+c2+cn. So if I calculate the Jacobian here I will get 1 0 0 1 1 0 and so on 1 1 1. So it is a lower triangular matrix with the diagonal entries as unity. So if I take the determinant of that that is going to be = 1 only.

(Refer Slide Time: 43:09)



So the joint pdf of nomination c1, c2, cn it is simply n factorial again. However, the range is now different in the case of Ui’s now this is becoming c1, this is becoming c1+c2 so basically what you are saying is that c2>0, c1>0 and so basically all the ci’s will be greater than 0. At the same time the summation Ci will be between 0 and 1.

Now because of the symmetric nature of the Ci’s appearing in this one if consider say Ci1+Cir then the distribution of this is same as say c1+c2+cr because it is based on simply the differences and we have seen that the Us-Ur is having the distribution as Us-r that means only the difference matters. So therefore whether I start from any other point it will not make any difference. So this is the concept of coverage.

Let us look at say one particular case, suppose I take 2 here if I take 2 here then this will become f c1 c2=2 c1>0 c2>0 and c1+c2<1. So if we consider the distribution here how it is looking like? On this c1+c2=1 is basically this. So basically the distribution is here so the density value=2 in this region. If we consider say n=3, then this will become 6 and the reason then will become c1+c2+c3<1.

That means I consider the plane c1+c2+c3=1 and we are below that in the first quadrant. So that is the idea of the coverages here. So coverages are useful because they are telling that the corresponding distribution F how much area it covers between 2 successive order statistics or between any few of order statistics. So this is useful information and what is important here is that you can see.

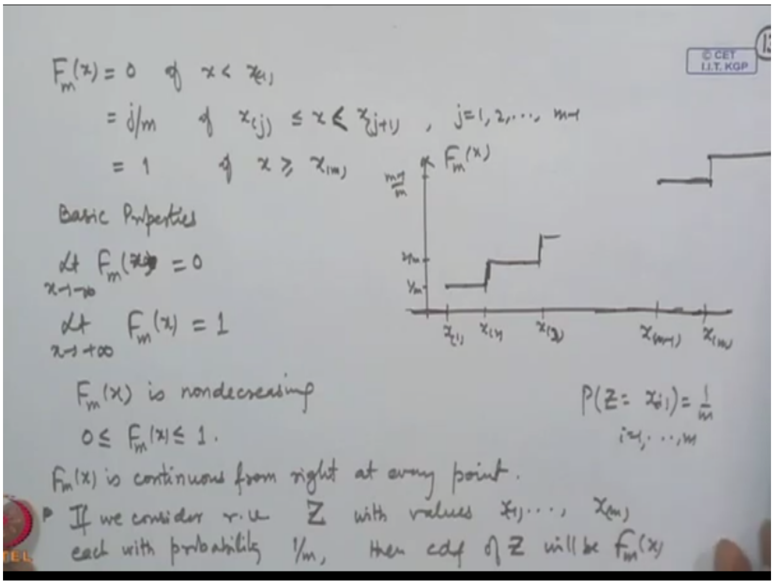
Basically, I started with any F here but now we are dealing with the uniform distributions, the distribution of c1, c2, cn they are free from what is the original distribution, so this is what is important here. When we do not pay enough attention on details of the exact model that means capital F is not known, but only we assume that it is a continuous distribution then we are able to talk about how much coverage is there etc without actually getting into the exact form.

So this is the advantage of the distribution free methods or the non-parametric method because the conclusions are independent of the original distribution. Another concept that will be of much use in fact it is one of the paramount importance that is actually empirical distribution function or the sample distribution function. As you can see, here I have proposed the estimations for population quantile as a sample quantile.

And as a consequence see for example variability is estimated by the sample range. In general, we consider any position and corresponding to that position we have an estimate here. Now if we consider the estimation of the distribution itself based on order statistics then you can define a function so that is what we call empirical distribution function or the sample distribution function.

So let us consider say X1, X2, Xn is the random sample from a distribution Fx okay. Now corresponding in place of n let me put m here because I will be using m and n interchangeably. Let us consider this order statistics X1, X2, Xm okay and now based on the observed values, I use with the small caps here these are the observed order statistics.

(Refer Slide Time: 49:22)



So if I consider Fmx=0 if x<x1=j/m if xj is <= x less than xj+1 for j=1, 2, up to m-1=1 if x is >= xm. So we can think of this function like this. Let us consider the plots here. So suppose this value is here x1, then x2 x3 and so on, xn-1 m-1 and xm. Then up to x1 this value is taken to be 0, then between x1 to x2 this value will be 1/m that means so suppose this is my 1/m okay this point is 1/m then you have 2/m and so on.

And between x2 to x3 this will be 2/m and so on. Between xm-1 to xm this will be = m-1/m. Suppose this is m-1/m and beyond that it is = 1. So we have made Fmx here. This is the function that we will be getting that means it is constant between 2 successive order statistics and at the end points it changes that means it has a jump at those points. So it is actually a step function.

So this is called the sample distribution function or the empirical distribution function of the order statistics here. Certain basic properties you can see for example if I consider Fm as x tends to –infinity then certainly this is = 0. If I consider as x tends to+ infinity then certainly this is = 1, then Fmx is non-decreasing function and the function is lying between 0 and 1.

Another thing that you observe that it is also continuous from right at every point. Fmx is continuous from right at every point. So if I consider the random variable say Z with values x1, x2, xm each with probability 1/m then cdf of Z will be Fm that means I am saying probability of Z=xi=1/m for i=1 to m. If we have this, then the distribution or the cdf of this it will be exactly this function here.

In the following lecture, I will discuss further applications of the empirical distribution function. I will prove some results based on that. We will define certain additional properties which will include 2 samples and based on that we will be able to derive some other results here, so that I will be covering in the next lecture.