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

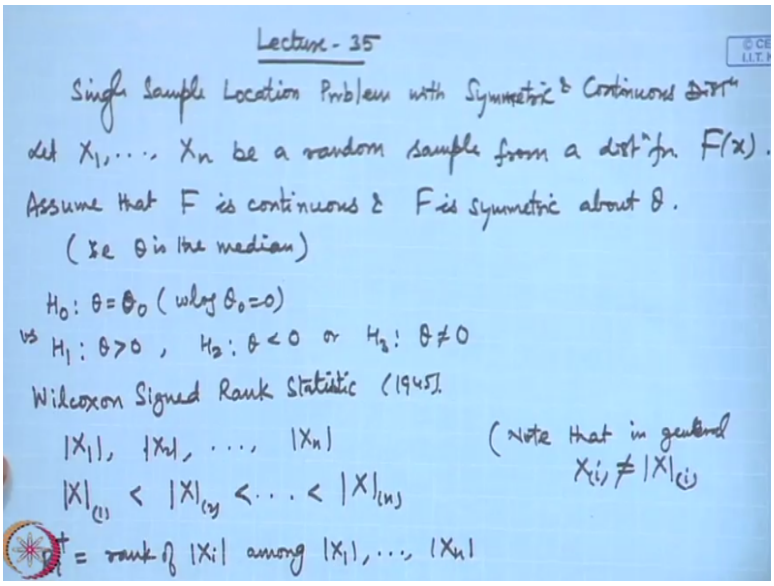
Lecture – 35
Nonparametric Methods - VIII

Friends, in this course till now we have introduced the order statics and their distributions. We considered probability integral transform, and therefore the distribution of the probability integral transforms of the order statistics, the distributions of one of them the joint distribution, their moment structure. Then we introduced the empirical distribution function and using the empirical distribution function if we consider the transformations of the random sample observations.

And their order statistics and we looked at their distributions, their joint distributions, and the moment structures. We saw that how these can be used in certain 2 sample testing problems. We discussed the goodness of a test by Kolmogorov and Smirnov and also the original one that why Karl Pearson. Now we can concentrate on the location problems, single sample location problem. We have seen that one of the raw test or a knife test is given by the scientist.

That means how many of the observations are above the median value which we want to test or below that so that is called the scientist. We have seen its all right it does not depend upon the measurement values. It is simply dependent upon the how many positive or how many negative values are there. Then there are certain other tests which are based on the observation, the ranks of the individual observations rather than just sign so one of the first one is the Wilcoxon signed rank test so let me introduce the problem.

(Refer Slide Time: 02:08)



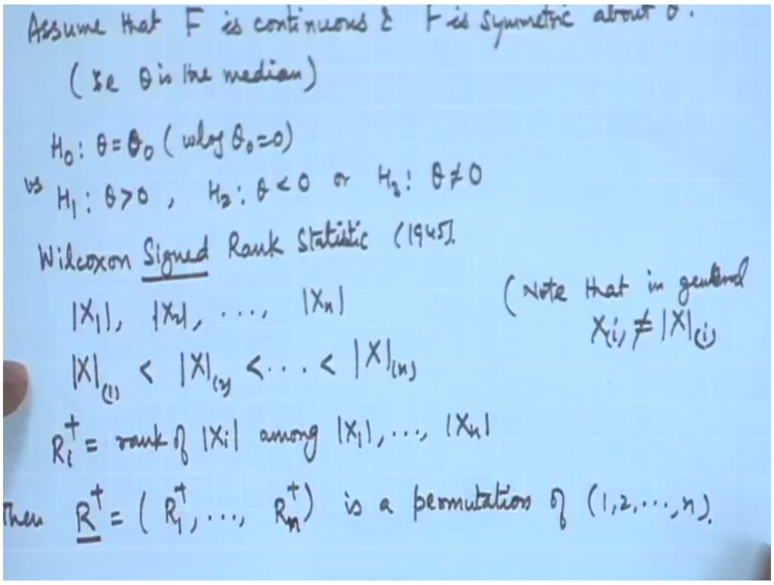
So we are considering single sample location problem with symmetric and continuous distribution. So let us consider suppose x1, x2, xn a random sample from a distribution function fx so this is the cumulative distribution function. Assume that f is continuous and f is symmetric about a point theta. See if it is symmetric about theta then of course we can say that theta is the median. In the scientist we have not assumed symmetry.

We just say that whether median is a given value or not of course the distribution of that thing we have found for the case of symmetric distribution also but in general it can be anything. So we want to test problems like this whether theta = 0. Basically we can test theta = theta not. See if we consider theta not so without loss of generality we can take theta not to be 0 as in previous problem.

I have already explained so again we can consider hypothesis of the time theta > 0 or H2 theta < 0 or H3 theta not = 0. So these could be alternatives. We will consider application of the sign rank statistics. This is called Wilcoxon signed rank statistics. This was given by Wilcoxon in 1945. Let us consider observations by taking their magnitude. So now the raw values have transformed to their magnitudes and consider their ordering.

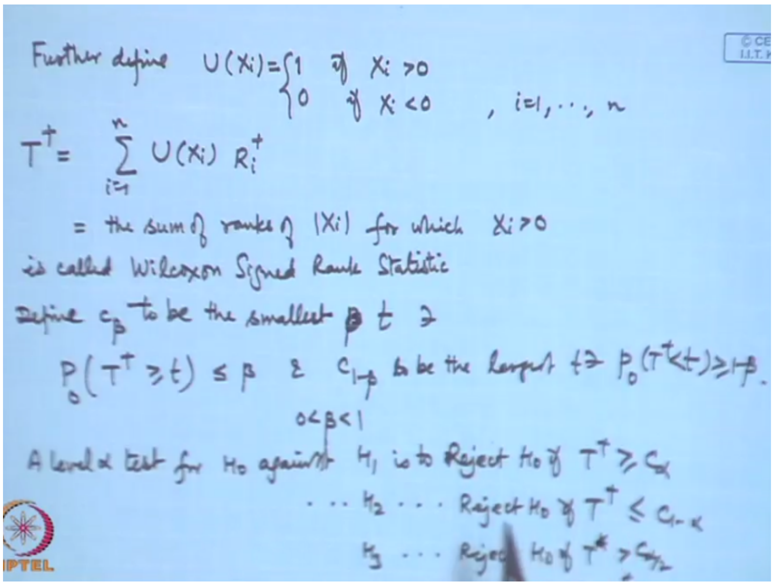
So let us consider say X1 among them. Now this is different. You note here firstly we are considering magnitude and then we are ordering. So these are different from that note that in general this xi will not be same as xi. If all the observations are positive, then this may be true. If all the observations are negative, then reverse of this may be true that means the ordering will be simply reversed. So this is different. We are looking at the magnitudes and let us consider say Ri + is the rank of absolute xi among x1, x2, xn.

(Refer Slide Time: 06:03)



Now if we are considering this, then if you consider the vector R+ that is R1 + and so on Rn +. So this will be simple a permutation of 1 to n is a permutation that is why it is called signed rank because we have considered modulus here. So we are bothered about the +, - sign is a permutation of 1, 2, n.

(Refer Slide Time: 06:42)



Now based on this we define uxi uxi = 1 if xi is positive and it is = 0 if xi < 0 of course = 0 case we are ignoring because we are dealing with the continuous random variables so probability of xi = 0 or will be 0. Now based on this we define T +. T + is summation of uxi

Ri + . I = 1 to n. Then actually it is nothing but the sum of ranks of modulus xi for which xi is actually positive. Because I am taking uxi * Ri + .

So if xi is negative then this term will not be counted. So it is the sum of the ranks of the modulus xi for which xi is positive. This is called Wilcoxon signed ranked statistics. Now you can understand that I am considering only the once which are positive and for those which are positive I am looking at the ranks of xi among the ordered modulus xi. So we then now you can easily see that.

What will happen that if theta is > 0 that means > theta not or something like that so here since we have taken without loss of generality 0 then there will be more values which will be positive. Therefore, this value will be somewhat larger. So if we consider the distribution of T + and we consider the percentage points of that and again see although the random variables are continuous, but this T + is discrete because this is simply the sum here.

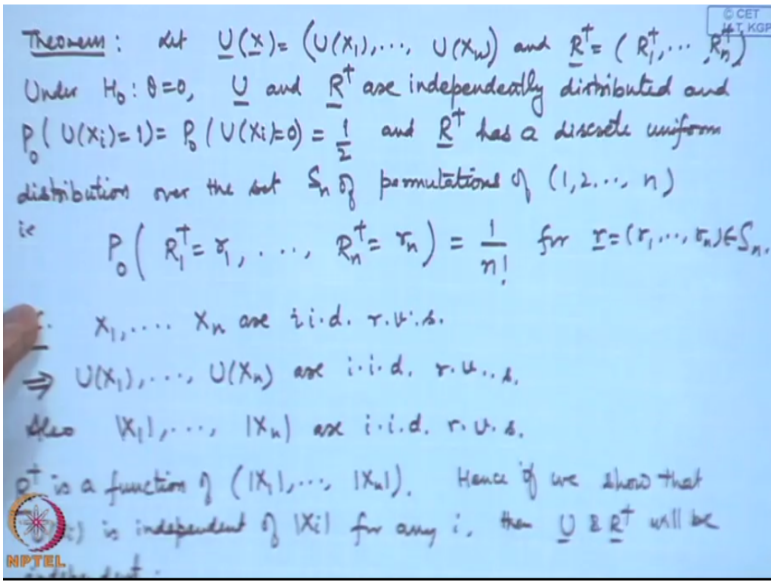
As in the signed rank as in the signed test statistics the Wilcoxon signed rank statistics is also having a discrete distribution. So therefore there is a possibility that a particular significant level may not be attain. So we then considered in the same say define c beta to be the smallest t such that probability of T + >= t under the null hypothesis that is median is 0 is <= beta and of course C1 - beta to be the largest t such that probability of T + < t is >= 1 - beta of course beta is some number between 0 and 1.

So you can consider basically that c beta is the actually if it is a continuous distribution then it will be simply the upper hundred beta percent point and this one will become the lower 1 01 - beta percent point here, but since the distribution of T + is discrete so we need to define in the terms of a smallest and largest here. So we can then consider that a level alpha test for H not against H1 is to reject H not if T + is >= to some c alpha against H2.

It will be to reject H not if T + is <= some C1 - alpha against H3 it will be reject H not if T + is either > some c alpha/2 or T + is < some C1 - alpha/2. Now the question comes about the determination of this c alpha values. Nowadays of course it is easy to look at the computer program and we can fix up this thing.

But let us look at a general result of this nature. actually since this is a random permutation in general because given observed values this R1 + R2 + Rn + 1 will be a random permutation of 1 to n and how many permutations will be there, there are n factorial permutations here. Therefore, each permutation will have a probability 1/n factorial under the null hypothesis. so let us write this as a result here.

(Refer Slide Time: 12:20)

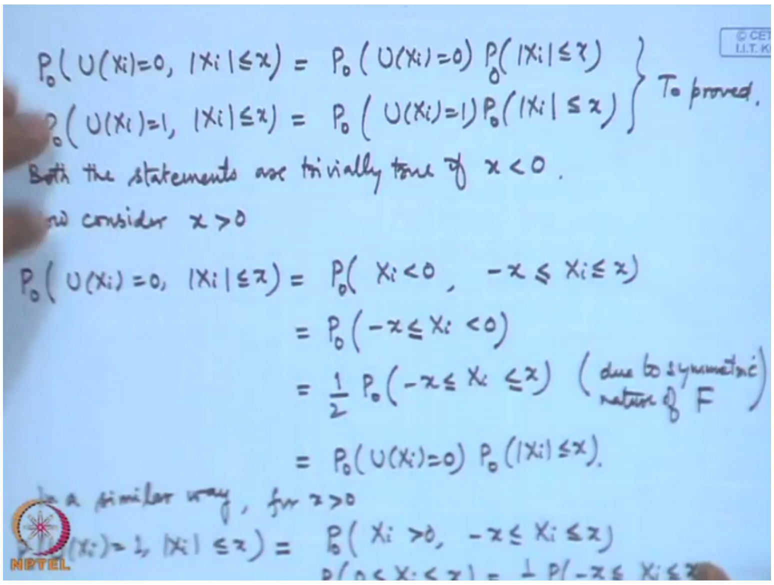


We have the following theorem. Let us consider ux vector to be the ux1, ux2, uxn that I sign of xi so we just collect them. So this is a collection of 1s and 0s and we consider the R+ as the vector of the signed ranks under H not that is theta = 0, u and R+ they are independently distributed and probability of uxi = 1 = P not of uxi = 0 that will be half and R+ has a discrete uniform distribution over the set Sn of permutations of 1 to n that is we are saying probability of R1 + = some R1 and so on.

Rn + = Rn that is = 1/n factorial for R = R1, R2, Rn, belonging to Sn. Sn is the set of all permutations of the number 1 to n. Let us look at a rough proof of this. So x1, x2, xn are independent and identifically distributed random variables. Now this implies that ux1, ux2, uxn they will be independent and identically distributed random variables. It will also mean that modulus of x1, modulus of x2, modulus of xn these are also iid. Now if we look at the R+ vector.

This ranks are functions of modulus x1, modulus x2, modulus xn is it not. Therefore, because how I have defined Ri. Ri is the rank of modulus xi among modulus x1, modulus x2, modulus xn that means this is entirely a function of the absolute values here. So here you look at R+ is function of modulus x1, modulus x2, modulus xn. Now you see here. u is a function of x1, x2, xn and this is a function of modulus so if we can show that uxi is independent of the modulus then we are through. So hence if we show that uxi is independent of modulus xi for any I then u and R+ will be independent.

(Refer Slide Time: 16:50)



Let us consider say probability of say uxi = 0 modulus xi <= x. then this is equal to probability of uxi = 0 * probability of modulus xi <= x. This is one statement I need to proof. I also need to proof probability of uxi = 1 modulus xi <= x = uxi = 1 modulus xi <= x. These are the things to be proved. now one thing you note, if we take this small x to be negative then certainly this term is 0 and this term is 0 and similarly in the second statement.

So both the results are satisfied. Both the statements are trivially if x < 0. Now let us consider x to be greater than 0. Now for greater than 0 let us consider one term here. Uxi = 0 modulus x1 <= x. Now this is xi < 0 because uxi is 0 if xi < 0 and the second part I write as - x <= xi <= x. Now this is nothing but if you combine these 2 it is becoming simple - x < xi < 0. we have assumed that xi has a symmetric distribution about 0.

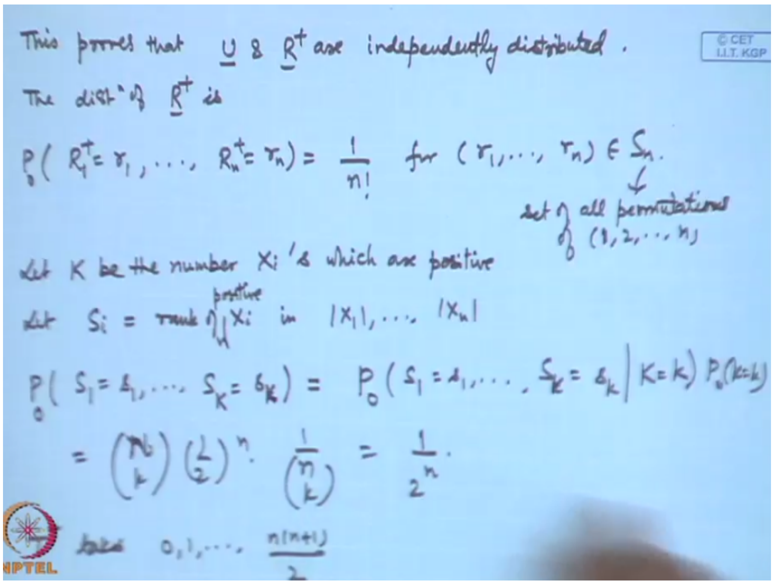
So this can be written as half times - x < xi <. So here of course <= is there so we can include that of course it will not make any difference if I by mistake do not put = because the probability of equality is actually 0. So this statement is due to symmetric nature of capital F. So therefore this is nothing but p not of Uxi = 0 * probability of modulus xi <= x. So you can see here.

I have proved this statement for x < 0 it is trivially true, for x > 0 now the proof is there. In a similar way if you consider P not uxi = 1 modulus xi <= x for x > 0. So that is equal to probability of xi > 0 - x <= xi < = x that = probability of now if I again combine these 2 statements it is reducing to 0 < xi <= x and as before due to symmetry this can written as half times probability of - x <= xi <= x which is nothing but probability of uxi = 1 * probability of modulus xi <= x.

So we have proved this second statement for x < 0 as well as for x > 0. so if you look at ux1 for example so it is certainly independent of modulus x1 and naturally it is independent of modulus x2, modulus x2 modulus xn. So in particular what I am able to prove is that u of x1 will be independent of the vector R1 + R2 + Rn + .

In a similar way, ux if it is consider it is independent of the vector R1 + R2 + Rn +. So if I look at the total vector because uxi are independent and identically distributed so if I look at the vector of u that is this one since each of them is independent of R+ if I look at the vector here which is obtained by simply by combining independent random variables therefore this is also going to be independent of R+. So this proves that.

(Refer Slide Time: 22:20)



This proves that u and R+ they are independent. They are independently distributed. Now since x1, x2, xn are independent therefore modulus x1, modulus x2, modulus xn are independent and also identical. Therefore, any ordering among them will be equally likely therefore the distribution of this will be simply. The distribution of R+ is R+ = R1 and so on. Rn + = Rn that = 1/n factorial for any permutation R1, R2, Rn belonging to Sn.

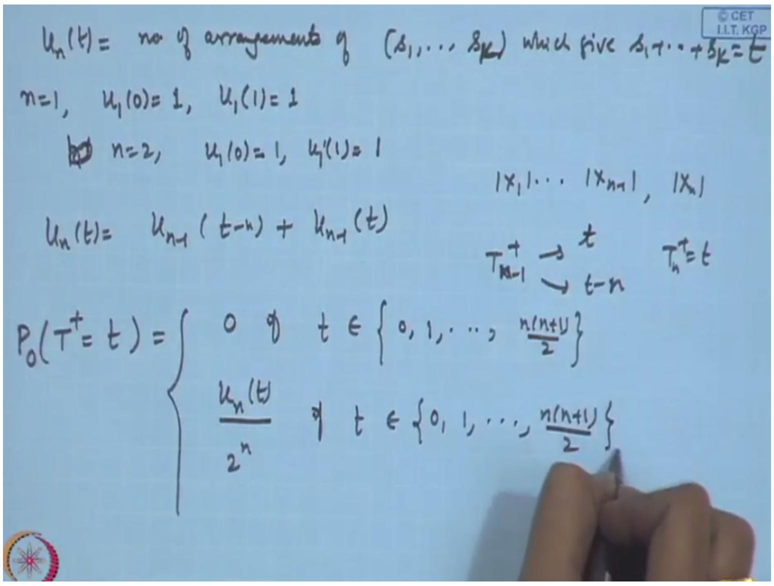
This Sn is denoting the set of all permutations of the numbers 1 to n. Now let us look at further the distribution of P +. I have been able to obtain separately the distribution of the terms which are involved in T +. Here the distribution of u is coming, the distribution of Ri + is coming also the independent is there. So now somehow we try to utilize this to derive the distribution of T + .

Let us look at it. Let K be the number of xi which are positive. Of course, we have seen the distribution of K that is binomial n half under the null hypothesis and also let us consider let Si be the rank of si in modulus x1, modulus x2, modulus xn. Now this is important here. When I am considering ordinary this one then the rank of xi is simply the ith one whatever term is coming.

Now I am looking at the rank of raw xi among the modulus x1, modulus x2, modulus xn. So this is only for positive xi. Rank of see originally it would have been that in the same order it would have come, but now because some of the negative xi will be placed in between because of the taking absolute value therefore these ranks will change. So let us consider say what is probability of say S1 = s1 and so on. SK = sk.

Let me put here K this is this number of positive xi. Let us consider the null hypothesis. so this is s1, Sk = sk. Now I am putting small k. Given k = k p not k = k. so that is equal to nck 1/2 to the power n, 1/nck. So this cancels out you are getting simply 1/2 to the power n. So you can see this number is 1/2 to the power n here. The reason that each of the xi can be positive or negative with probability half. Let us consider say T +. What are the values of T +. This takes values 0, 1, up to n * n + 1/2. If all of them are positive, then it will ben/n + 1/2 if all are negative then this will be 0. So now.

(Refer Slide Time: 27:20)

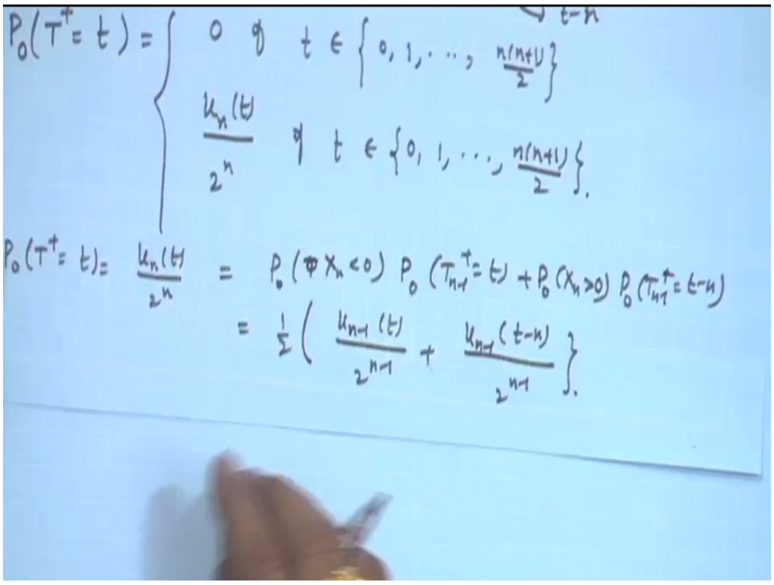


Let us consider u and t. It is the number of arrangements of s1, s2, sk which given s1 + s2 + sk = T. You can actually see suppose I have n = 1 that means only one observation is there then u10 that means how many arrangements will be giving you this is equal to 0 that will be simply 1. How many arrangements will give 1 only 1 because either x1 can be positive or negative. So if I consider say P10.

Now let me define and similarly if I look at say n = 2. For n = 2 u10 will be 1, u11 that will be equal to 1. Let us derive a recurrence relation here. It can be written like this. unt that will be equal to un - 1, t - n + un - 1 t. So this is the recurrence relation that will be getting because if you are looking at say ranks of x1, x2, xn -1 and then you add xn here then what will happen then this tn - 1 + so either it will remain t or it will become t - n. if tn + I = t because either it will be added by n.

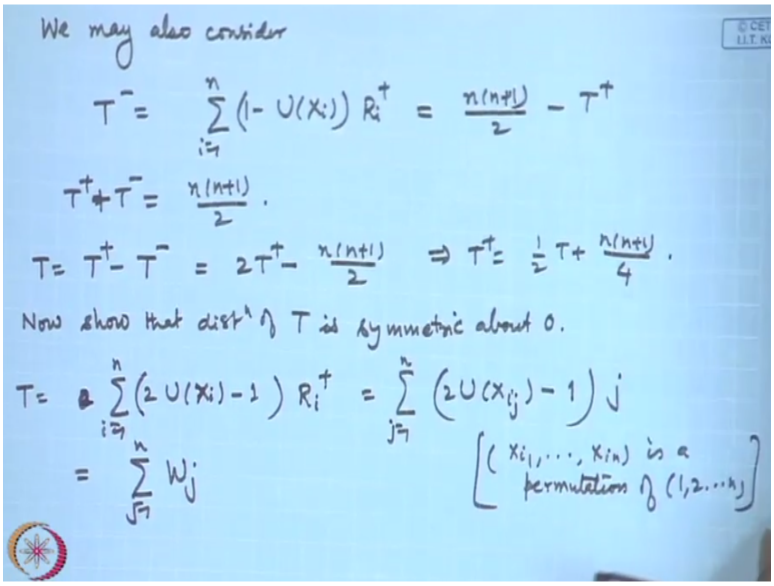
That means in case it is positive then all of them will be added by 1 if it is negative then no value is added here the previous ranks will remain the same so it will not change the value or in each one extension will be there so it will be there so it will become t - n + n. So if I consider probability distribution of t + then it = 0 if t is not in the interval. It does not take one of the value 0, 1, and so on n * n + 1 /2 and it = unt/2 to the power n if t is in the set 0, 1, 2 n * n + 1/2.

(Refer Slide Time: 30:10)



This recurrence relation actually gives you a method of calculation of this values of u and t because you are having say p not t + = t then that is unt/2 to the power n, but this we can also write as xn < 0 tn - 1 = t + xn > 0 tn - 1 = t - n. Now both of these are known that is half times un - 1 t/2 to the power n - 1 + u n - 1 t - n to the power. So actually this gives you a method of evaluating the probability distribution of t + at the nth stage. In a similar way one may consider t - also.

(Refer Slide Time: 31:20)

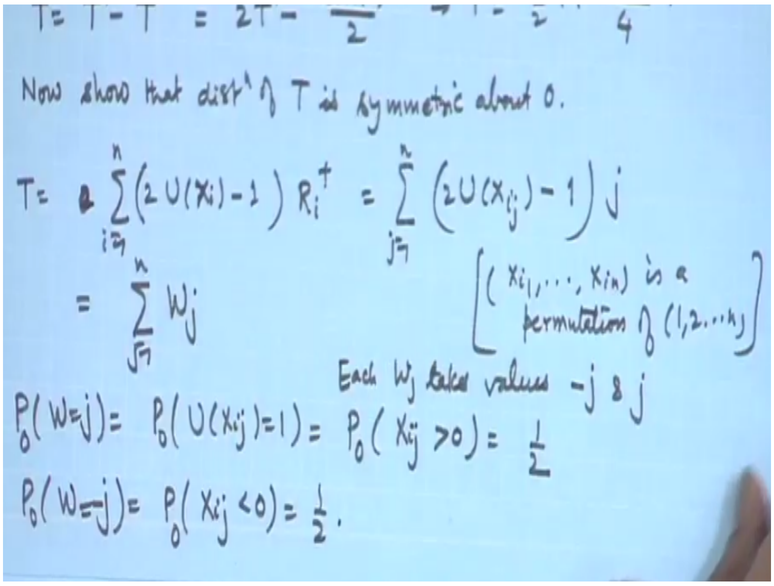


We may also consider t - that is 1 - uxi that means I am taking the ranks of negative one because when uxi is 0 1 - uxi will become 1. So that is actually n * n + 1/2 - t +. So this t - is directly related to that that is basically we are saying t + +t - = n * n + 1/2. If we consider say T = t + - t - which is of course = 2t + - n * n + 1/2. For 2 sided testing problem then the alternative is theta is not = 0.

If we are considering this alternative for this, this t gives more power than t +. So this actually implies t + = you can take it to the other side you get 1/2t + n * n + 1/4. Now we show that distribution of t is symmetric about 0. So t = 2 * sigma twice uxi - 1 * so this 2 I can write inside Ri + I = 1 to 2. So that is = 2 ux ij - 1 * j j = 1 to n. because each of this Ri + will take some values 1 to n so I am writing that then correspondingly this value will change here.

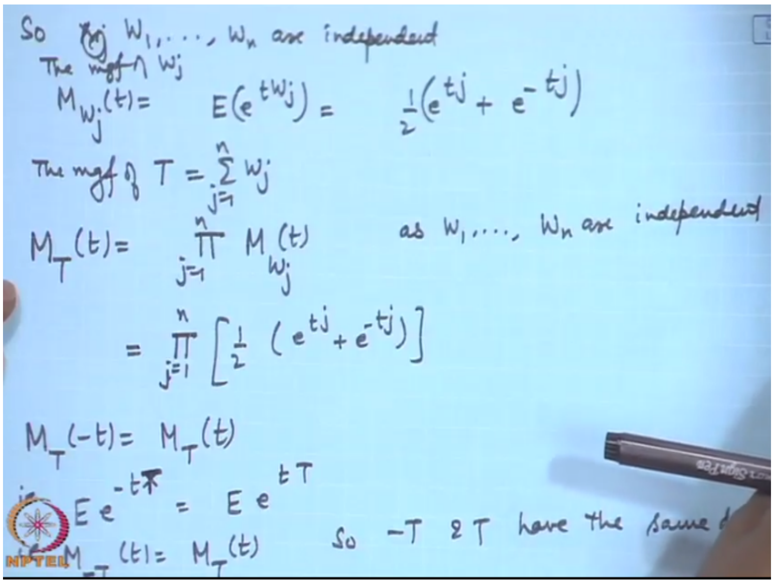
This xi1, xi2, xin this is a permutation of 1 to n. So this permutation is obtained in the way in which the ranks are distributed. So we get it a new name. Let us call it sigma wj is defined by this term. Now what are the values of the wj. It takes value either + j or - j. LEt us look at this.

(Refer Slide Time: 34:19)



What is the probability this each wj takes values - j and + j. What is the probability say w = j that is simply the probability of uxij = 1 that is probability of xij > 0 but under null hypothesis this is simply half and similarly if I consider - j then that is = probability of xij < 0 that is also half. So what we have proved that they are simply taking 2 values + j and - j each with probabilty.

(Refer Slide Time: 35:11)

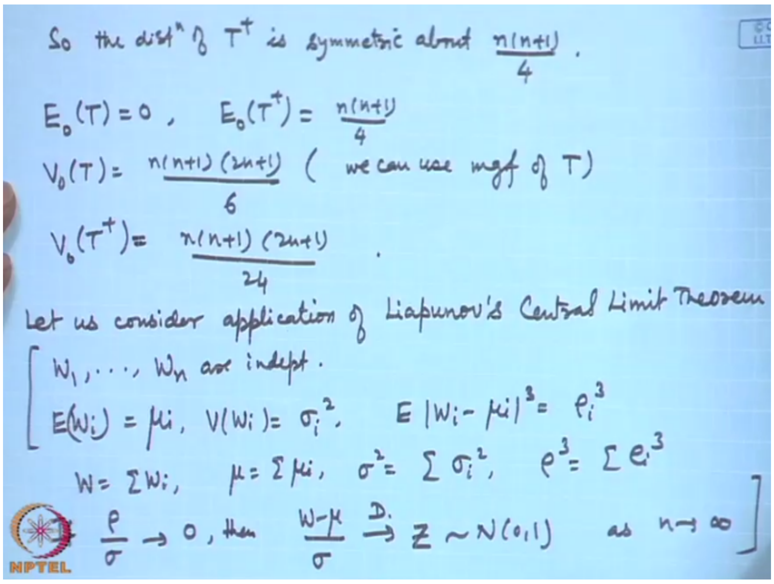


So w1, w2, wn they are independent of course we should not say identical because although they take 2 values of equal probability that those values are changing okay so this is wj here and if I look at the moment generating function of say wj mgf of wj that is expectation of e to the power twj = half e to the power tj + e to the power - tj because it is taking 2 values.

So if I consider the mgf of t that = sigma wj since they are independent it is simply becoming product of the mgfs of wj. So this is nothing but product of j = 1 to n half e to the power tj + e to the power - tj. Now if I look at m t of - t. Then it is same as mt of t that is expectation of e to the power - tx = expectation of e to the power tT. So this is same as saying m -t at t is same as mT of t. So - t and t have the same distribution.

So if the random variable and its negative has the same distribution it means that T has a distribution symmetric about 0. So this is interesting. We have obtained the distribution of t symmetric about 0 and what is T plus. We have expressed t + in terms of t. So if t is symmetric about 0, t + will be symmetric about n * n + 1/4. So these things actually give us more features about the test statistics that we are using here.

(Refer Slide Time: 38:36)



The distribution of T + is symmetric about n * n + 1/4. If I look at expectation of t that is 0 expectation of T + that will become n * n + 1/4 and variance of T that n * n + 1 * 2n + 1/6 well this, you can calculate from the mgf because we have the mgf. We can use mgf of t. because second moment we can obtain by see this is the product of the term so if I consider one derivative then I will get here in the product so each term will be coming here.

And they will become a minus sign here in each of them because there are n terms here so at ith level this term will be differentiated and other terms will be there, but the term which is differentiated will give me a minus value so that will cancel out. When we go for the second derivative now that term will become actually positive other terms will become 0, but that will happen with each of them.

So it is becoming basically sigma of j square because half half is there so that will be adding up so that is giving you simply n * n + 1 * 2n + 1/6 and if I consider variance of t + then simply because it is half times that so that is becoming n * n + 1 * 2n + 1/24. So this is interesting. We are able to find out the distribution of t + that is distribution of t and we are able to derive some of the first and second moment under the null.

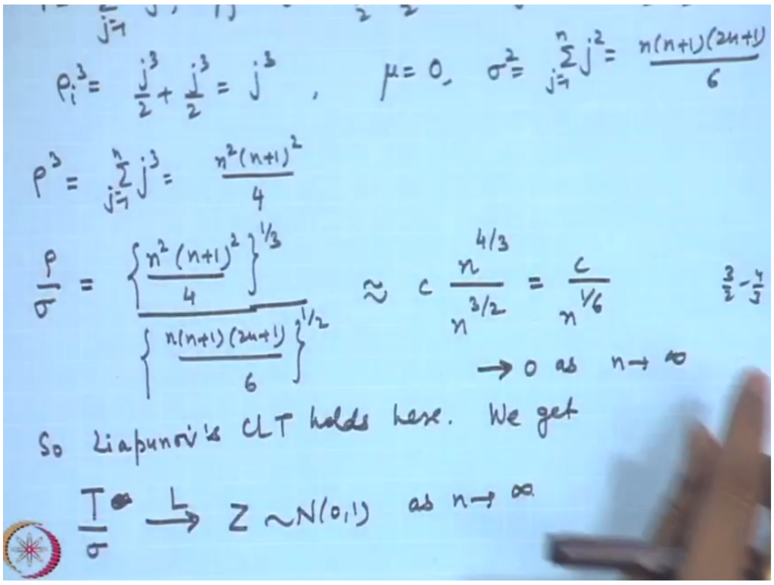
Now once that is there and we are expressing it as a summation we can actually consider the central limit theorem. Let us consider application of Liapunov's central limit theorem is applicable for independent but possibly nonidentical random variables. So w1, w2, wn they are independent expectation of wi so I am writing down the statement here. Let us consider expectation of wi = mui variance of wi = say sigma I square.

Let us consider the third central moment of wi. Let us call it say rho iq and if we are defining the terms like w = sigma of wi, mu = sigma of mui, sigma square = sum of sigma I square, rho cube = sigma of rhoi cube. Then if rho/sigma goes to 0 then the distribution of w - mu/sigma is as totally normal as n tends to infinity. So this is in convergence in distribution or convergence of law.

So this is actually the Liapunov's central limit theorem. See if you look at the original central limit theorem it is for the independent and identically distributed random variable which is also called I think Lindeberg Levy central limit theorem that is applicable when random variables are independent and identically distributed. We only assume that the variance is existing. So second moment existence is there. When the random variables are not identically distributed the Liapunov's central limit theorem gives a sufficient condition for the asymptotic distribution being normal.

Basically this is the central limit theorem here, but here we have to assume that third one here that means the third central moments must exist and then the condition is imposed upon that. Now if we look at our t it is exactly of that same form. Here w1, w2, wn are independent certainly they are not identically distributed. Their distributions are symmetric. Means are 0 but variance will be j square/2 + j square/2 so that is j square so let us use this.

(Refer Slide Time: 44:00)

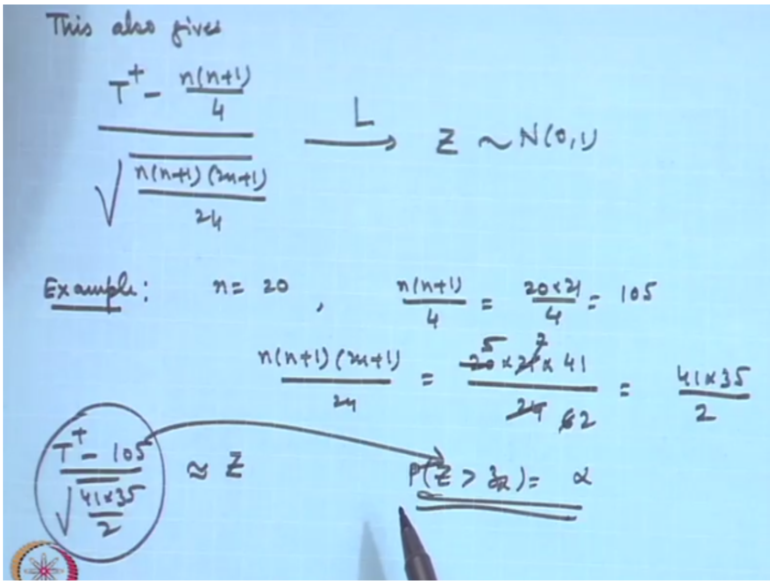


So t is sigma wj j = 1 to n. expectation wj that is mu j that = j/2 - j/2 that is equal to 0. If we consider say sigma j square that is expectation of wj square that will become = j square/2 + j square/2 = j square and if I consider the third central moment since mean is 0 it is simply equal to j cube/2 + j cube/2 because we have taken the absolute value here so this is j cube. So now we write all the terms here. mu is 0.

Sigma square = sigma j square j = 1 to n = n * n + 1 * 2n + 1/6. What is rho cube? Rho cube = sigma j cube for j = 1 to n. Then it is = n square * n + 1 square/4. So, if I consider rho/sigma so there will be some constant here because there is some constant coming here. Actually we can just write it is n square n + 1 square/4 to the power 1/3 divided by n * n + 1 * 2n + 1/6 to the power 1/2. So this is proportional to as n becomes large. So this is n to the power 4 so n to the power 4/3/n to the 3/2.

Some constant will be there. So this is 4/3 - 3/2 so that is coming in the denominator. So n to the power so 3/2 - 4/3 that is simply becoming 1/6. So this certainly goes to 0 as n tends to infinity. So Liapunov's CLT holds and we get asymptotic distribution of this t now mu is 0 so t/sigma. This sigma square root of this quantity. t/sigma this is converging to z in distribution as n tends to infinity. So asymptotic distribution of t is simply normal and since there is a direct relationship between t + and t so if I put it here then I get the asymptotic distribution of t + also.

(Refer Slide Time: 47:15)



This also gives the asymptotic distribution of t + - n * n + 1/4/ square root of n * n + 1 * 2n + 1/24 as asymptotical normal distribution. So this is basically giving a method that for large sample size we can straight away apply a normal test for testing the equality of the median to

0. Suppose n is really large for example let us take say some particular value suppose I take say n = say 20.

If I take n = 20 then what will become n * n + 1/4 that is becoming = 20 * 21/4 that is 105 and n * n + 1 * 2n + 1/24 that will become = 20 * 21 * 41/24. so this is 41 * 35/2. So in this case t + - 105/square root of 41/35/2. This will be approximated by normal 01. So if we are considering z greater than see T + greater than cl for then it is equivalent to z > z alpha and that we take to be alpha.

So we can actually consider the value based on this. So testing problem. So we calculate suppose some data set is given we calculate T + for that for n = 20 and then we compare with this value. Similarly, for the 2 sided testing problem we can directly use T itself. So we will look at the ty sigma whether it is large or small corresponding to z alpha/2. So you can see here the concept of this Wilcoxon signed rank test is.

Let me just look at the term here. I will explain once again. So from the original observations here one assumption is there of course that we are considering symmetry about the median that is if it is symmetric around some point that point becomes median and therefore we are checking actually symmetry about the median and now we are testing whether the median is equal to a specific value.

Without loss of general t we take that specific value to be 0 so then the testing problem becomes whether the median is 0 or it is > 0, < 0 or not = 0. For this Wilcoxon signed rank statistics considers the magnitude of the xi. Based on that we create the ranks of the absolute values and we look at those values which are positive from the positive ones we look at the ranks of modulus xi among this.

So once that is done so this the some of the ranks of modulus xi this is called the Wilcoxon signed rank statistics. So this can be used. We have shown that the distribution of this can be calculated using a recurrence relation which I gave that is the terms of un function here. This un, un is have a recurrence here so one can calculate and of course some tables of these are available, but even if we are not using the tables of that if the sample size is somewhat large then we can actually use this approximation because it is actually turning out to the some.

So t + is written as a sum. t is written as a some so therefore the distribution of this can be approximated by a normal distribution if the sample size is sufficiently large. Now based on this the problem becomes quite simple. Now whenever we are having large data sets we straight away use the normal test based on the tr t + therefore it is convenient to apply here. We will extend this concept further.

We will consider something called Walsh averages and we will consider these signed rank statistics in terms of that we will also define the general linear rank statistics. See here you see we are considering sigma of uxi * ri +. So we are actually adding the ranks linearly that means in multiplying by uxi uxi can take value 1 and 0. We will consider a general function of this nature. We will look at how it can be used for constructing some other test so, that we will be covering in the next lecture.