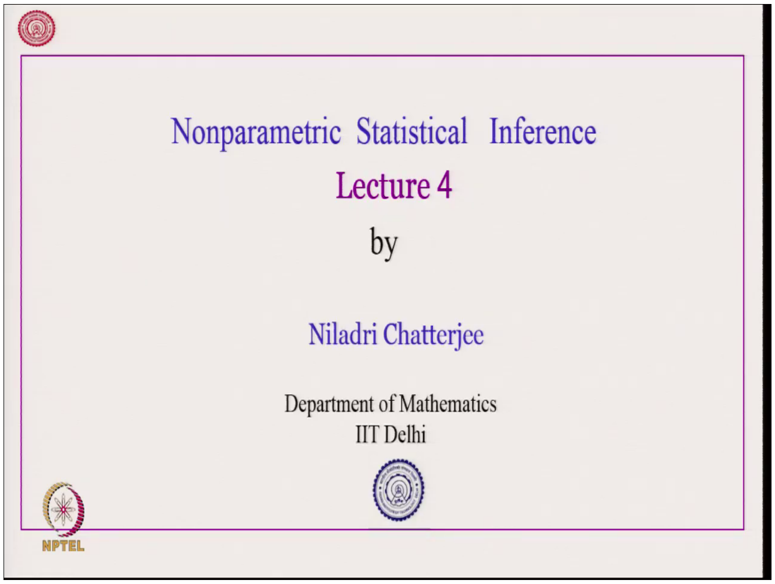
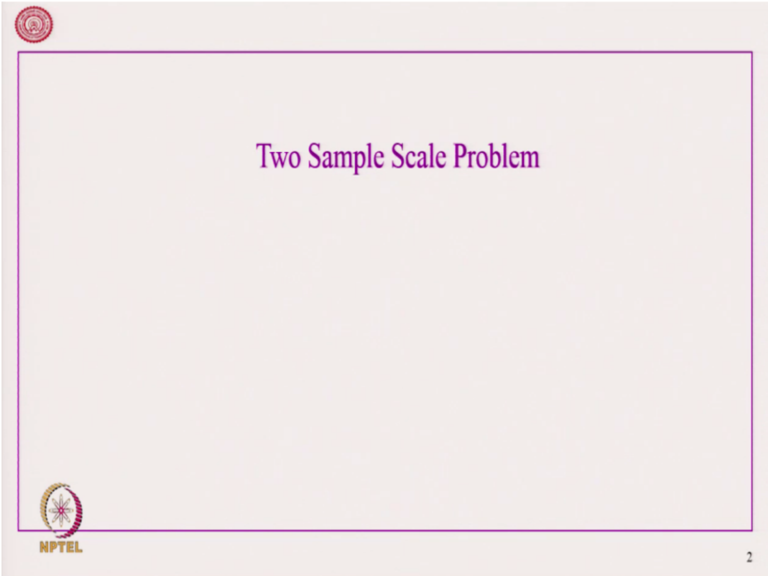
Nonparametric Statistical Inference
Professor Niladri Chatterjee
Department of Mathematics
Indian Institute of Technology, Delhi
Lecture 04
Nonparametric Statistical Inference

(Refer Slide Time: 00:17)



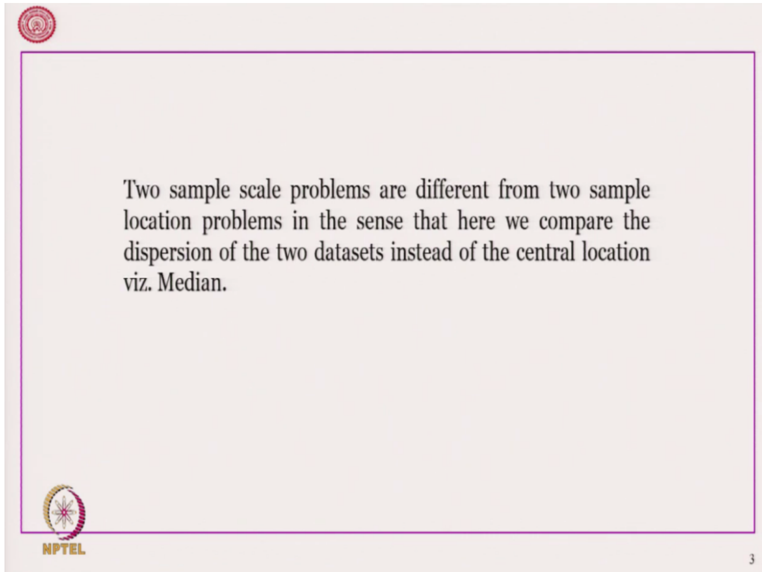
Welcome students to Moocs series of lectures on nonparametric statistical inference. This is lecture number 4.

(Refer Slide Time: 00:30)



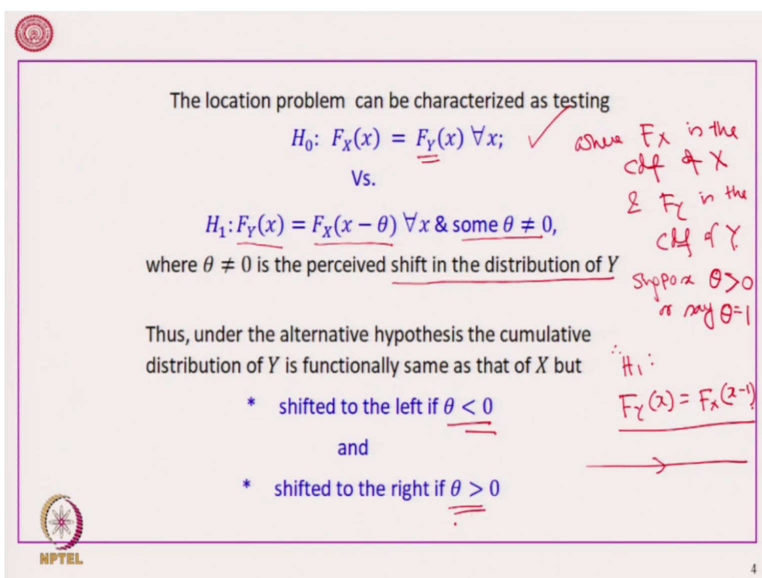
As I said at the end of the last class, that today, we shall be studying two sample scale problem.

(Refer Slide Time: 00:40)



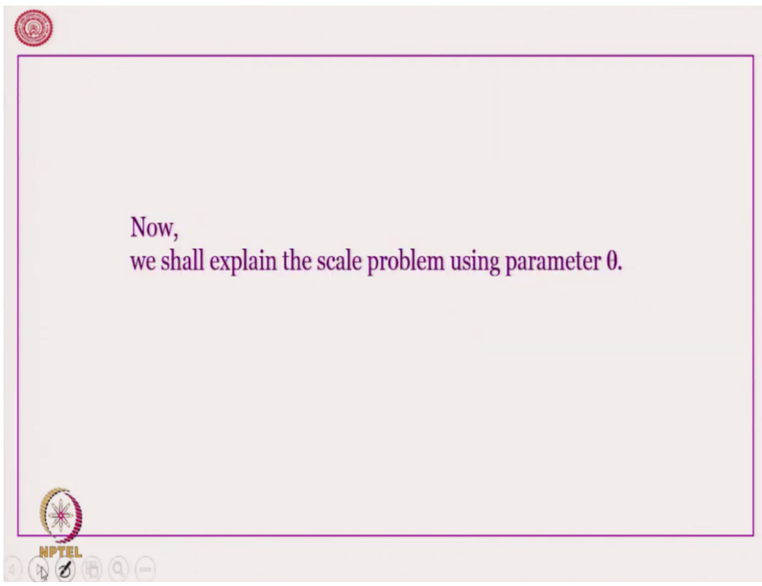
So, these are different from the two sample location problems in the sense that, here we compare the dispersion of the two data sets instead of the central location, namely, median. For median we have already seen Wilcoxon rank sum test and Mann Whitney U test which compares the central location of the two distributions.

(Refer Slide Time: 01:11)



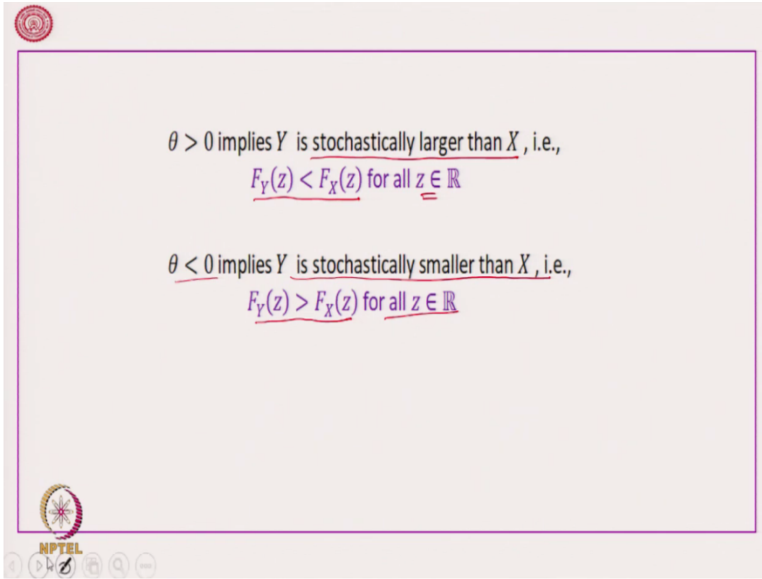
And therefore, if you want to characterize the location problem, we can use a parameter theta and we can say that the location problems can generally be characterized as testing, whether Fx at x is same as Fy at x where F_X is the cdf of x and Fy is the cdf of y against the alternative if Y of x is equal to F_X of x minus theta for some theta not equal to zero. So, suppose theta is greater than 0 or say theta is equal to 1. Therefore, the alternative H1, we are actually looking at whether Fy at x is equal to FX at x minus 1. That is we are looking at if there is a shift in the values of y from x. And hence, this is essentially testing if there is a shift in the distribution of y and that is why this is called a location testing. If the value of theta is less than 0, then we shall feel that the shift has been to the left side. If the value of theta is greater than 0, then the shift is going to be to the right side.

(Refer Slide Time: 03:21)



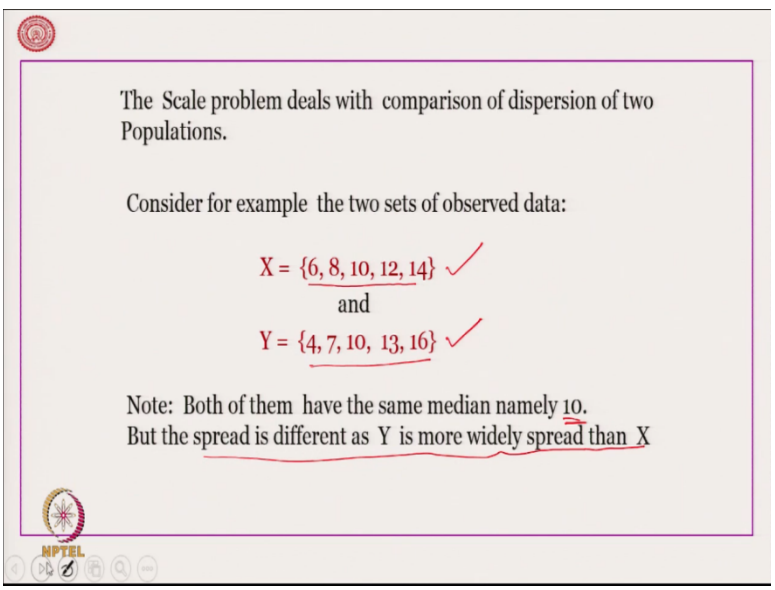
Now, we shall explain the scale problem using parameter theta.

(Refer Slide Time: 03:27)



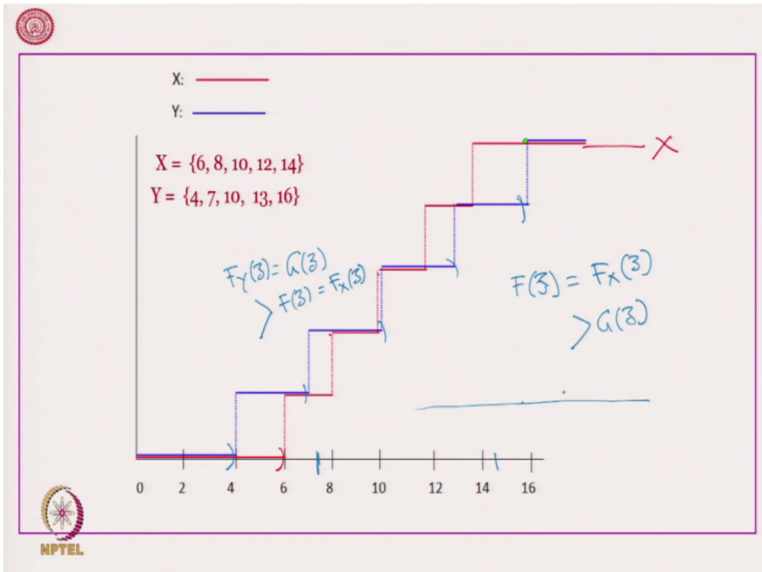
So, theta greater than zero implies Y is stochastically larger than X, that is $F_Y(z)$ is less than $F_X(z)$ for all z belonging to R and theta less than zero implies Y is stochastically smaller than x that is $F_Y(z)$ is greater than $F_X(z)$ for all z belonging to R. This I have already discussed in one of my earlier lectures, so, I am not going to illustrate this much.

(Refer Slide Time: 03:59)



So, let us look at the scale problem with an example first. Suppose I have two sets of observed data, X is equal to 6, 8, 10, 12, 14 and Y is equal to 4, 7, 10, 13 and 16. Note that both of them have the same median, namely 10. But the both of them have the same median namely 10. But the spread is different. As Y is more widely spread than X. You can see that the spread of X is from 6 to 14, but spread of Y is from 4 to 16.

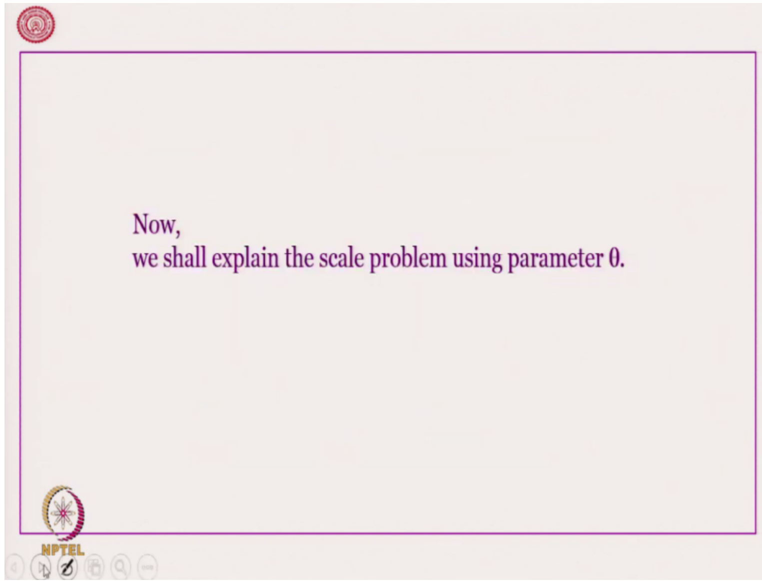
(Refer Slide Time: 04:44)



So let us plot them. Since X is 6, 8, 10, 12, 14, assuming they have equal probability, so till 6, the value is 0, at 6, there is a jump, then again another jump is at 8, then at 10 and then at 12 and then at 14. So, these red ones is giving me the cumulative distribution of the sample for X observations. Similarly, Y observation gives me this step function as its cdf. There are jumps at 4, then at 7, then at 10, then at 13 and then at 16 and then it becomes 1.

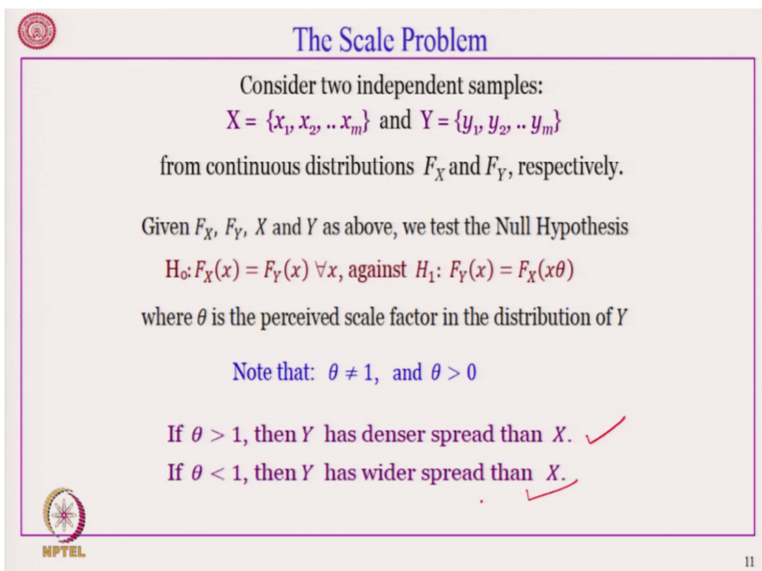
Now, if we look at the two curves, we see that below 10 the G(z) which is the cdf for y, let me write it also $F_Y(z)$ is equal to G(z), which is greater than F(z) which is F_X of z. On the other hand, if we look at an arbitrary point, say here we can see that F(z) that is the cdf for X that is greater than G(z). Therefore, we cannot say that X is stochastically larger than Y or Y is stochastically larger than X; rather from here we can see that the value of Y is more spread than the value of X. So, this diagram helps us to visualize this.

(Refer Slide Time: 06:56)



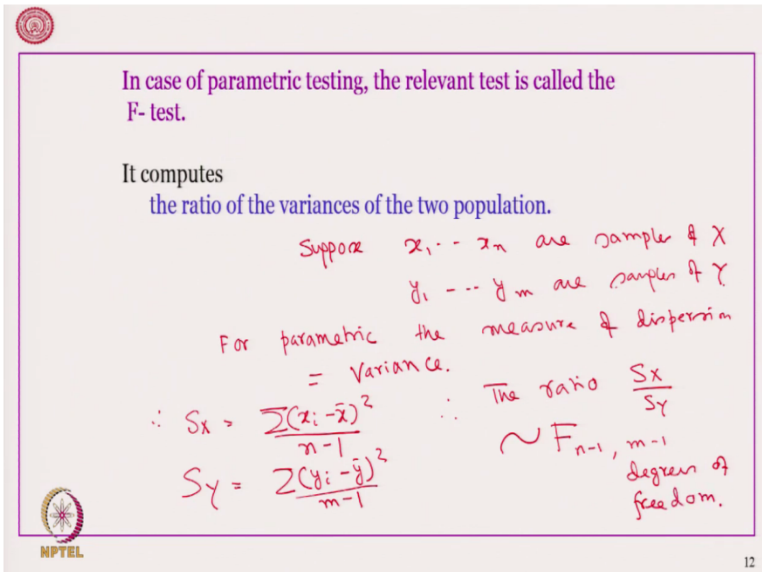
Now, we shall explain the scale problem using parameter theta.

(Refer Slide Time: 07:02)



So, what we are doing in a scale problem, we consider two independent samples, X is equal to x1, x2, xm and Y is equal to y1, y2, yn. So from continuous distributions Fx and Fy respectively. Given Fx, Fy x and y as above, we want to test if Fx(x) is same as Fy of x for all x, which is the null hypothesis against Fy of x is equal to a Fx of x theta, where theta is the perceived scale factor in the distribution of Y. Note that theta is not equal to 1 and theta is greater than 0. Do not get confused greater than 0 may mean theta is less than 1, but greater than 0 or 1 less than theta, say less than infinity. That means theta can be greater than 1 or theta can be less than 1. If theta is greater than 1, then Y has a denser spread compared to X, if theta is less than 1, then Y has a wider spread in comparison with X.

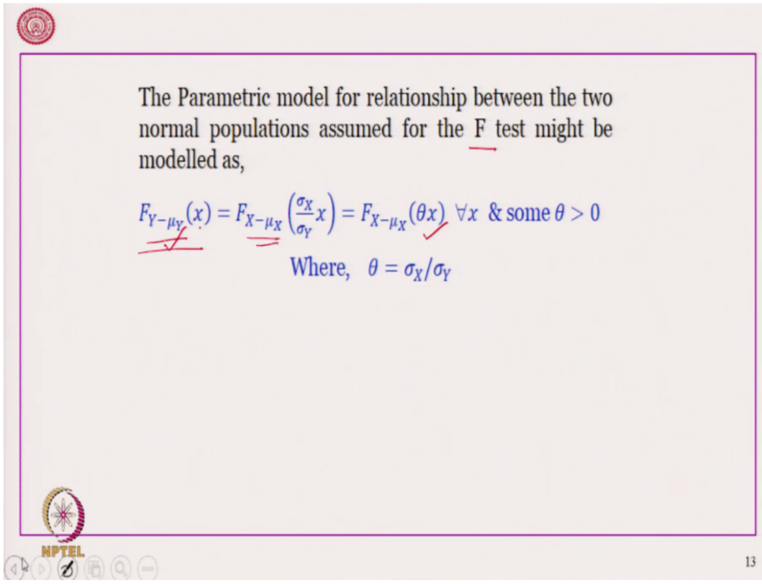
(Refer Slide Time: 08:36)



For illustration, in case of parametric testing, the relevant test is called F-test. I hope all of you are familiar with the F test. But for example, suppose x1, x2, xn are sample of X and y1, y2, ym are samples of Y. For parametric, the measure of dispersion is equal to variance, all of us know it. Therefore, sample variance of X is equal to sigma xi minus x bar whole square divided by n minus 10. This is the unbiased estimate for the population variance, similarly for y it is sigma yi minus y bar whole square divided by m minus 1. Therefore, the ratio Sx upon Sy is going to be Fwith n minus 1 comma m minus 1 degrees of freedom.

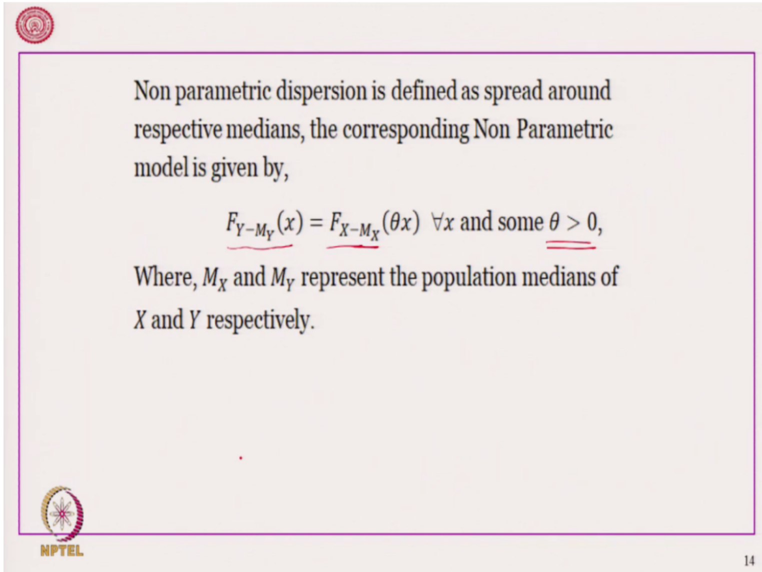
Therefore, when you compare the variances of 2 continuous distributions in a parametric setup, we use F-test for testing the hypothesis whether they are equal or not.

(Refer Slide Time: 10:53)



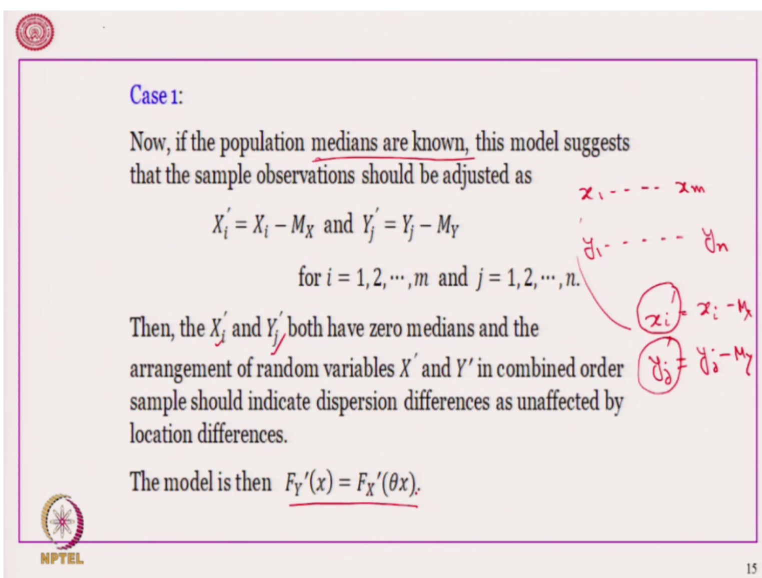
So, the same test now, we can actually look at it with the help of theta, that the F test can be written as what we are doing, we are standardizing y minus mu y that is we are subtracting the mean from the observation and we are looking at the cumulative distribution function. Similarly, corresponding to x, we are doing the same thing and we are checking if sigma x by sigma y by which this is the standardization of that one and that is what is called theta x, then we can see that the F test can be written in the form of F of y minus mu at x is equal to F of x minus mu x at theta x.

(Refer Slide Time: 11:58)



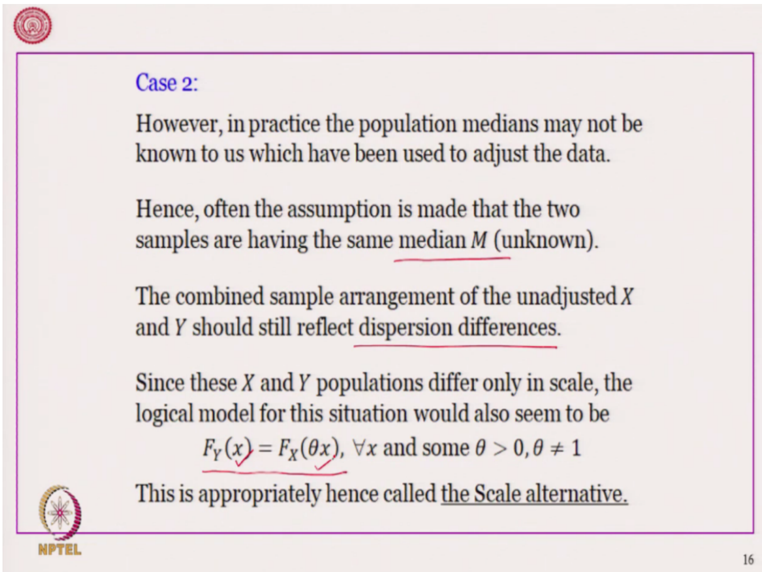
Why we have used this particular expression, it is only to correlate it with the nonparametric method and in a nonparametric also the dispersion is defined as spread around the respective medians. And the corresponding model is given by FY of minus MY, where MY is the median of y is equal to f of x minus MX at theta x, where theta is greater than 0. So, this is what we want to test.

(Refer Slide Time: 12:30)



Now, two things can happen. Case one is that the medians are known, we know the actual median of x and the actual median that is the population median of y. Therefore, when we have the observations x1, x2, xm and y1, y2, yn, we can subtract the median and we can get from there xii is equal to xi minus Mx and yj prime is equal to yj minus M of y. Thus the transformed observations are x1 prime, x2 prime, up to xm prime and y1, prime y2 prime up to yn prime. And we are looking at if the spread of the Xi prime and the Yi prime are actually following this, that Fy prime x is same as Fx prime theta X.

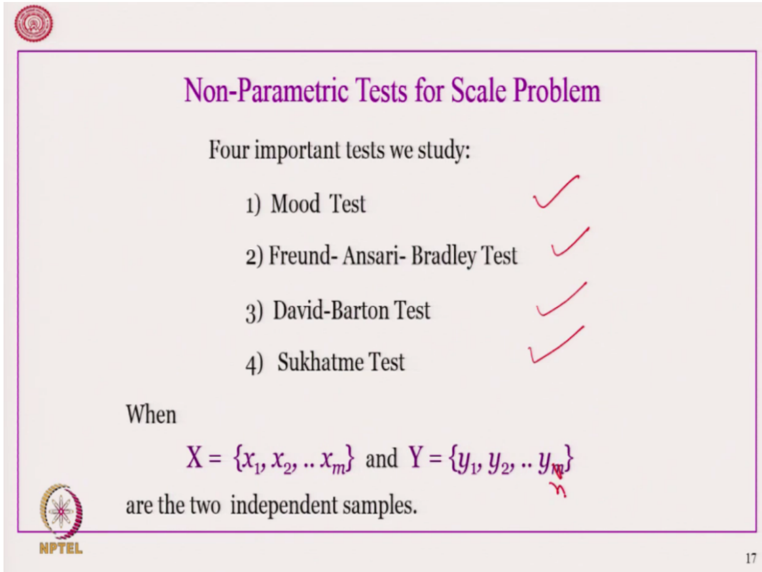
(Refer Slide Time: 13:49)



But suppose we do not know the actual medians of the two populations. This is generally true hence often the assumption is made that the two samples are having the same median because if we do not standardize it with respect to the central location, we cannot just measure the dispersion. So what we assume that they have the common median M and then the combined sample arrangement of the unadjusted x and y should still reflect that dispersion differences.

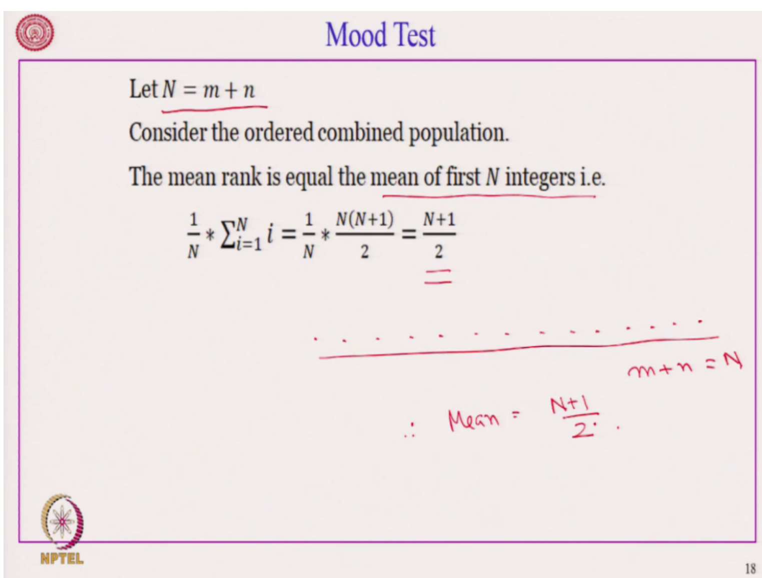
And therefore, again we go back to the model, Fy at small x is Fx at theta x or not. So, this is a little bit of theoretical interpretation to go inside what is happening in a scale problem test.

(Refer Slide Time: 14:50)



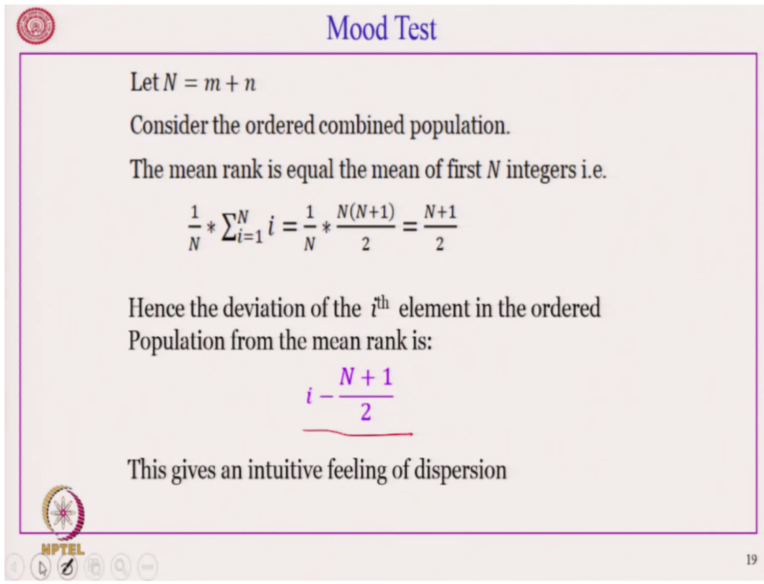
Now, let us look at which tests are available to us for testing such hypothesis in particular we shall look at the following four tests Mood test, Freund Ansari Bradley test, David Barton test and Sukhatme test. When X is equal to x1, x2, xm and y is equal to y1, y2, yn are the two independent samples. So, this is n.

(Refer Slide Time: 15:50)



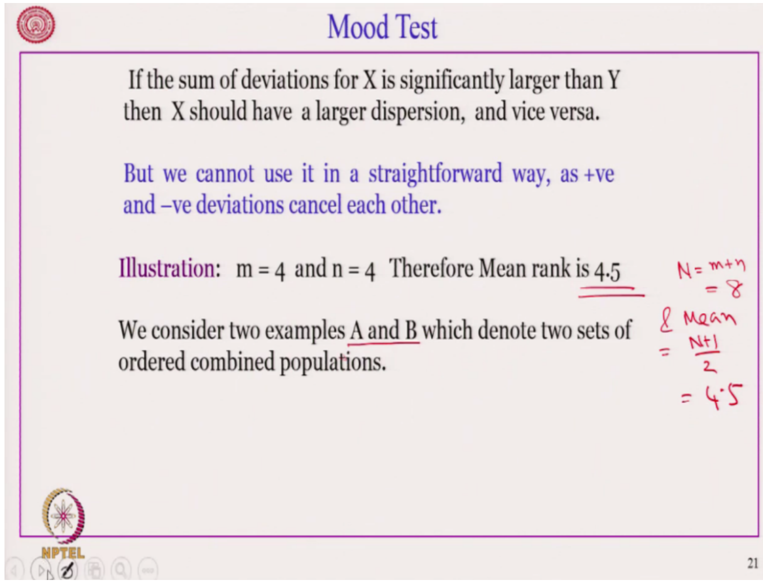
We have already seen that we pull together the two sets of observations and let capital And be m plus n consider the ordered combined population, the mean rank is equal to the mean of the first N integers that is N plus 1 by 2. So, if we consider all the x and y observations together and order them, the number of observation is going to be m plus in is equal to capital N. Therefore, the mean is equal to N plus 1 by 2. That is what I have written here.

(Refer Slide Time: 16:14)



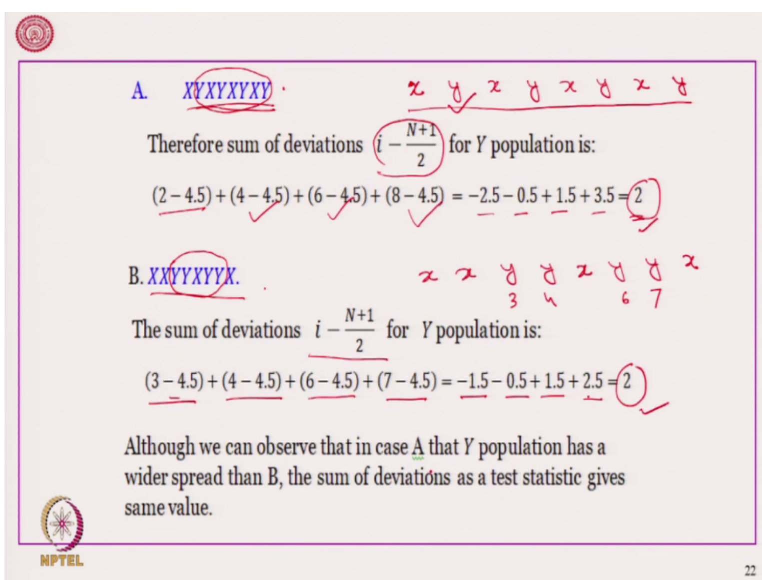
Hence the deviation of the ith element in the ordered population from the mean rank is going to be i minus N plus 1 by 2. So, this will give us an intuitive feeling of the dispersion, because we are now subtracting the mean to sort of normalize it. Therefore, the, we will look at the difference of the rank of the x observations from the mean. And if they are very close to the mean, then their sum is going to be smaller. If they are very far from the mean, their sum is going to be larger and from there we shall try to infer about the dispersion of x, similar thing can be done with respect to dispersion of Y.

(Refer Slide Time: 17:07)



So, let me illustrate that the mood test the sum of deviations for X, if that is significantly larger than Y, then X should have a larger dispersion, and vice versa. But the problem is we cannot use it in a straightforward way, because the positive and negative deviations cancel each other. So, let me illustrate it, suppose I have taken m equal to 4 and n is equal to 4, both of them are 4. Therefore, the mean rank is going to be 4.5 because N is equal to m plus n is equal to 8 and mean is equal to N plus 1 by 2 is equal to 4.5.

(Refer Slide Time: 18:14)

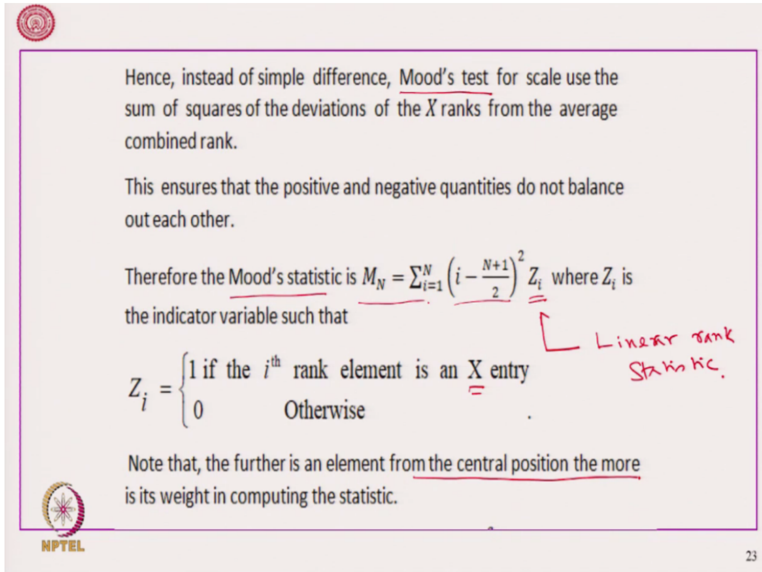


Suppose, we consider two examples A and B with the following arrangements, A is XYXYXYXY that means, the first observation is X, second observation is Y, third observation is X, fourth one is Y, fifth one is X, sixth one is Y, seventh one is X, eighth one is Y, this is in the ordered sample, I am not talking about the order in which samples are taken. Therefore, sum of deviations i minus N plus 1 by 2. For Y population is 2 minus 4.5 plus 4 minus 4.5 plus 6 minus 4.5 plus 8 minus 4.5, which is equal to minus 2.5, minus 0.5, plus 1.5, plus 3.5 is equal to 2.

Similarly, now consider B is equal to XXYYXYYX. Therefore, what we got smallest one is X, second smallest is X, then Y, Y, X, Y, Y, X. Therefore rank for the Ys are 3, 4, 6 , 7. Therefore, the sum of deviations of i minus n plus 1 by 2 is going to be 3 minus 4.5, plus 4 minus 4.5, plus 6 minus 4.5, plus 7 minus 4.5, which is equal to minus 1.5, minus 0.5, plus 1.5, plus 2.5. And as we can see, it is still coming out to be 2.

Thus, if this is the ordered sample or this is the order sample, we get the same value for the dispersion. This happens, because the negative values are cancelling against the positive values, but if we look at these two distributions, we can easily see that the Y population has a wider spread in this whereas in this white population is very concentrated in a middle part of the sequence of observations, here it has a much larger spread. So, that is the difference between these two that although that their sum of deviations from the central location is same. Actually their spreads are differing.

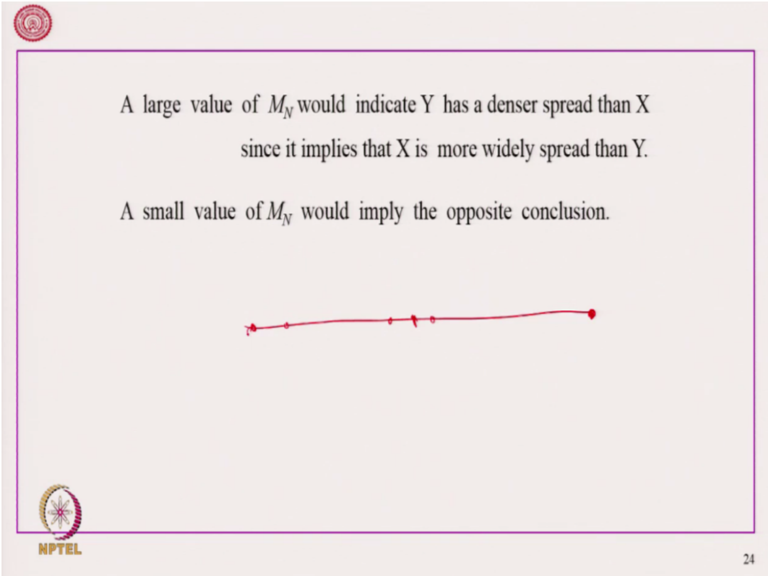
(Refer Slide Time: 21:07)



Therefore, instead of simple difference, we consider the square of the difference and their sum and that is the statistic that we are going to use and the corresponding test is called the Mood's test. Therefore, Mood's statistic can be written as M_N, which is sigma i minus N plus 1 by 2 whole square i is equal to 1 to capital N Zi, where Zi is equal to the indicator variable. We can take it corresponding to x or corresponding to y that is not very important.

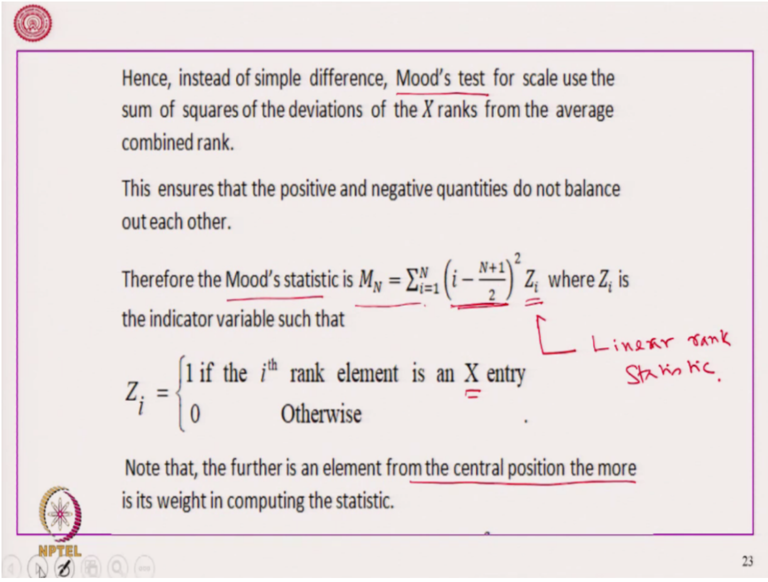
In this I have taken it corresponding to the X entry. Therefore, Zis are going to be 1 if the ith rank element is an X entry. Zi is going to be zero otherwise and we are already familiar with this type of thing as in the last class, we have seen that this is called a linear rank statistic. Note that the further an element from the central position the more is its weight in computing the statistic.

(Refer Slide Time: 22:28)



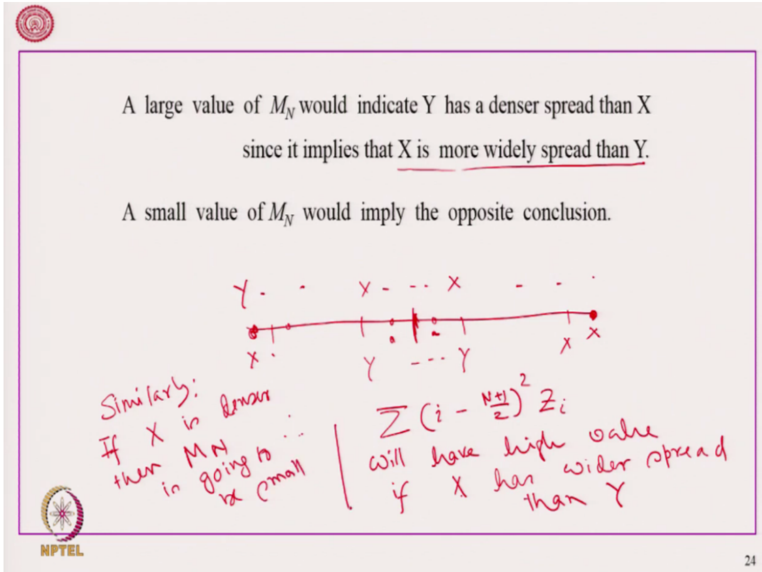
That is, suppose this is the ordered sample and this is the central location. So, these are distant observations from the central location and these are closer observations from the central location.

(Refer Slide Time: 22:47)



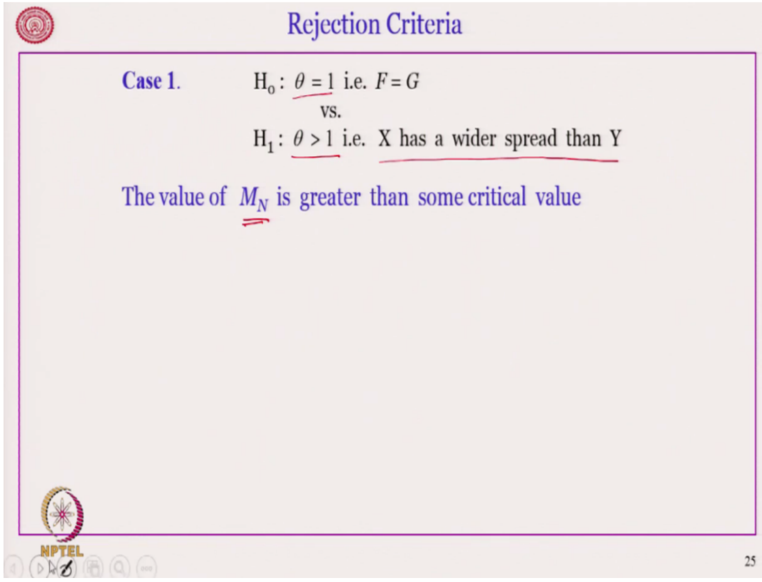
The weight given is i minus N plus 1 by 2 whole square.

(Refer Slide Time: 22:54)



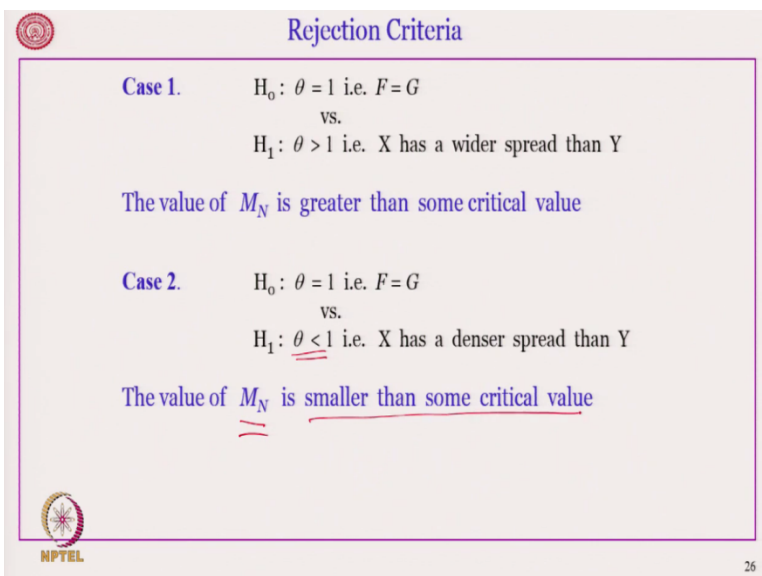
Therefore, corresponding to these distant elements the weight is going to be much more than the corresponding to the elements which are very close to the central location. Now, suppose X is more widely spread than Y. Therefore, it is expected that these values are more likely to be from X and these values are more likely to be from Y. Therefore, the statistic sigma i minus N plus 1 by 2 whole square times Zi will have a high value if X has wider spread than Y. Similarly, if X is denser that means, if these values are closer X and the farther values of are Y, then the value of the statistic M_N is going to be small. So, this is the philosophy on which Mood's test works.

(Refer Slide Time: 24:41)



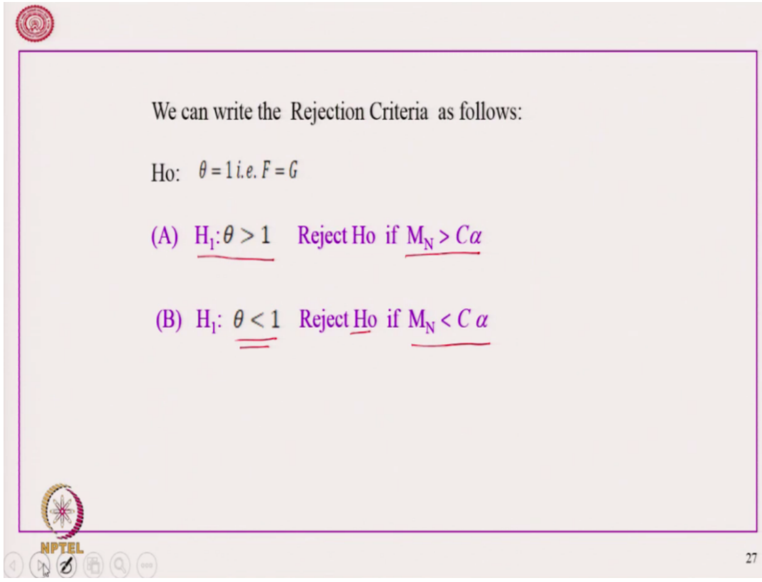
So, when we are rejecting the null hypothesis that is theta is equal to 1 in favor of theta is greater than 1, if the mood statistic M_N is greater than some critical value. As we have already seen that it is expected to be greater. Therefore, if it crushes some threshold, then we are going to say that X has a wider spread than Y.

(Refer Slide Time: 25:12)



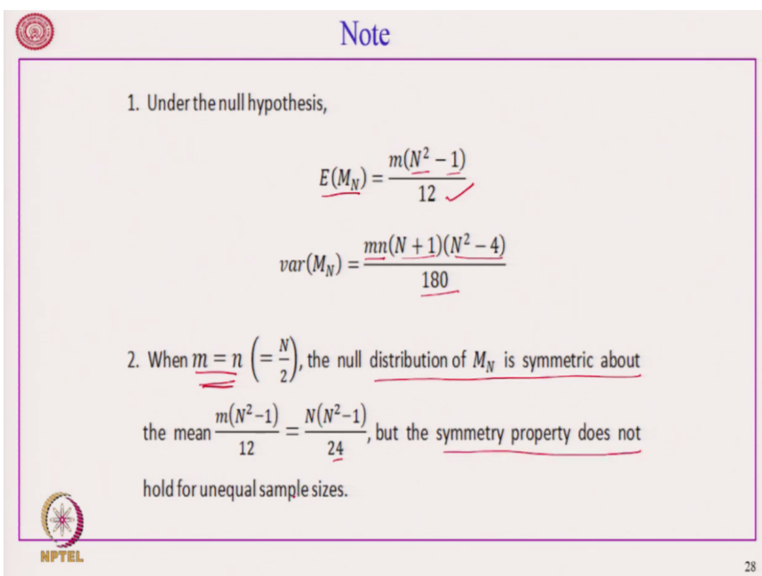
Similarly, if the null hypothesis is theta less than 1 that is x has a denser spread, then the value of MN is smaller than some critical value, which is very obvious.

(Refer Slide Time: 25:30)



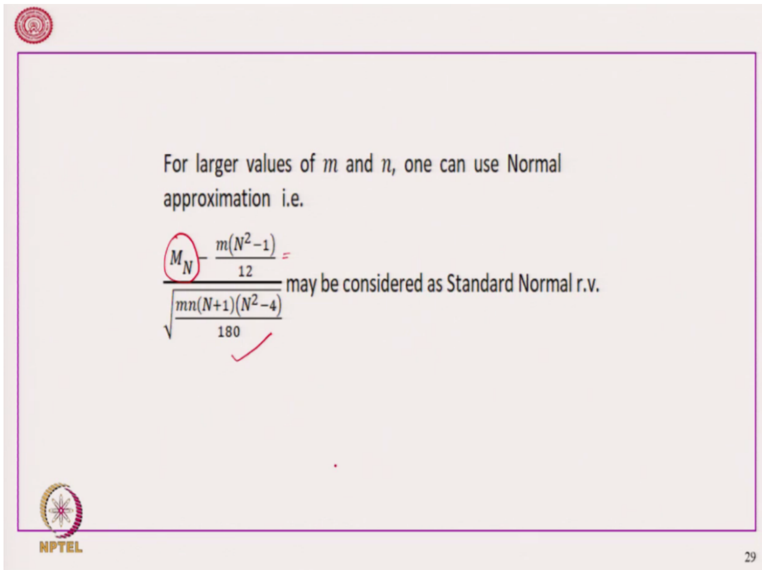
And similarly, we can write that, if alternative is theta is greater than 1, then reject H nought, if M_N is greater than some critical value at significance level alpha on the other hand if alternative is theta less than 1, then reject H nought if M_N is less than the critical value, at the significance level alpha.

(Refer Slide Time: 25:57)



Now, note that under the null hypothesis, expected value of M_N is equal to m times capital N square minus 1 by 12 and variance of M_N is equal to small m small n times N plus 1 into N square minus 4 divided by 180. I shall prove these in my lecture five, when I shall do some mathematics with linear rank statistic. For the time being let us assume these two in particular, if m is equal to small m is equal to small nthen the null distribution of capital M_N is symmetric about the mean, which is going to be m into N square minus 1 upon 12, which is nothing but N capital N upon 24 into N square minus 1. But if this condition does not hold that small m is equal to small n, then this symmetric property will not hold. So, this is a special case.

(Refer Slide Time: 27:13)

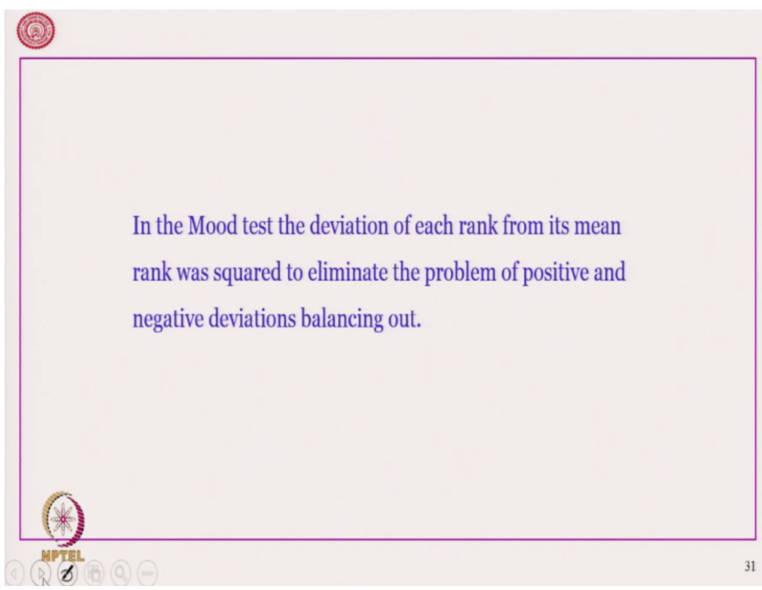


When m and n tend to be large then one can as before use normal approximation. And by now, you already know how to standardize it, it is M_N minus the expected value of M_N which is small m into N square minus 1 upon 12 divided by the standard deviation of that, which is this large expression and that is going to be standard normal. Therefore, the way we have done corresponding to Mann Whitney, or Wilcoxon rank sum test in a similar way, we should test it from the normal table. We shall check the value and decide whether to accept or reject the null hypothesis. So, that is about the mood test.

(Refer Slide Time: 28:04)

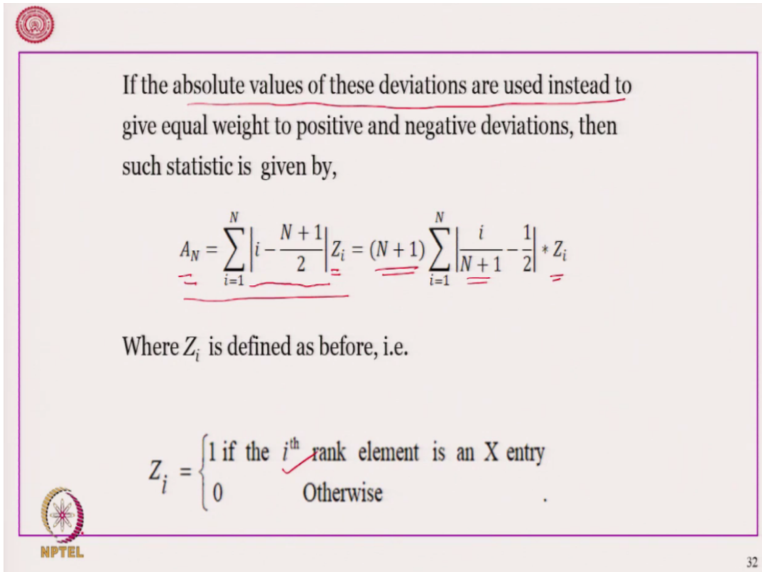
Let me consider a very similar test, which is called Freund Ansari and Bradley test.

(Refer Slide Time: 28:16)



So, in the Mood test, the deviation of each rank from its mean rank was squared to eliminate the problem of positive and negative deviations that they are balancing out each other. This we have already seen.

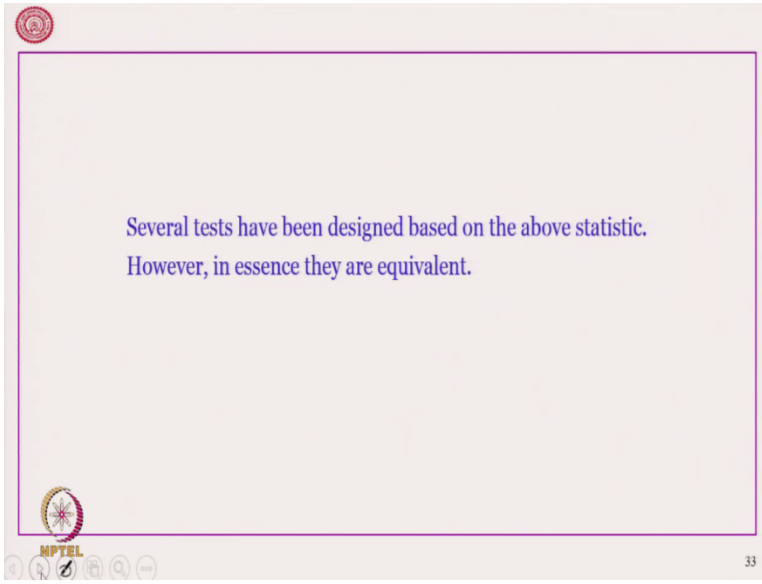
(Refer Slide Time: 28:36)



But in Freund Ansari test what we do instead of squaring it, we consider the absolute deviation from the central location. Therefore, the corresponding statistic is called A_N, which is now you look at it, it is the mod value of i minus N plus 1 by 2. It is the absolute deviation from the central value multiplied by the Zi, which is the indicator variable. So that is going to be the statistic for Freund Ansari test.

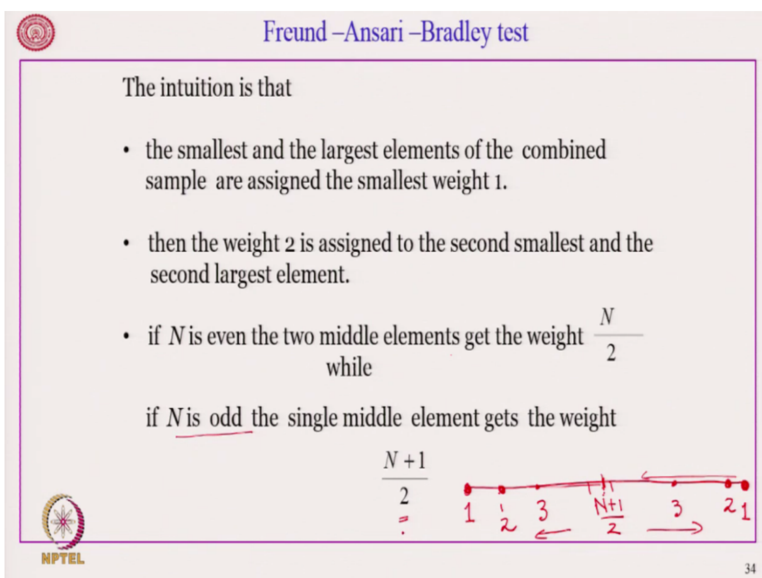
Now, this can be written, if we take out capital N plus 1 as common, then we can write it as i upon N plus 1 minus half multiplied by the indicator variable that is Zi is equal to 1 if the ith rank element is an X entry and it is 0, if ith rank element is from Y.

(Refer Slide Time: 29:41)



So, same principle has been used in designing several statistic, but in some sense they are equivalent. However, there are subtle differences, which is worth studying because to understand how these people have thought about making their nonparametric tests based on only a few samples.

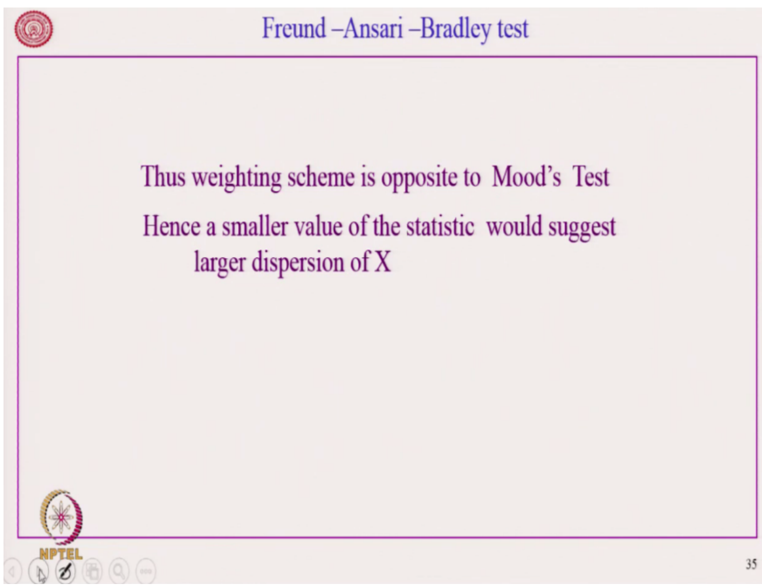
(Refer Slide Time: 30:08)



So, the intuition for Freund Ansari test is that the smallest and the largest elements of the combined sample are assigned the smallest weight 1. So, if we consider these to be the ordered sample, and suppose, this is the mean N plus 1 by 2, if this is the smallest and this is the largest element in the ordered sample, they are going to be given the value 1, the weight. Suppose, this is the second smallest, this is the second largest, they are going to be given the values 2. Similarly, this the third one and this is the third one, then they are going to be given the values 3. As we come to the center, if N is even that two middle elements get the weight And by 2, which is very clear. While if N is odd, the single middle element gets the weight N plus 1 by 2.

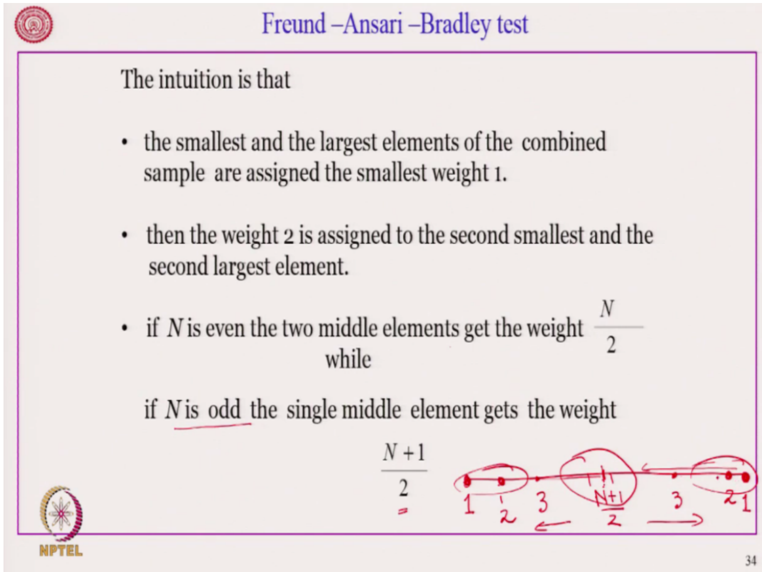
So, all the values are given numeric integer weight. The difference from Mood's test is that in Mood's test, the elements which are farthest from the central location have been given maximum weights and that was decreasing as you are coming to the center. But in Freund Ansari test the maximum value is given to the central elements and as we go towards the ends we can see that weights are decreasing. What is the effect?

(Refer Slide Time: 32:07)



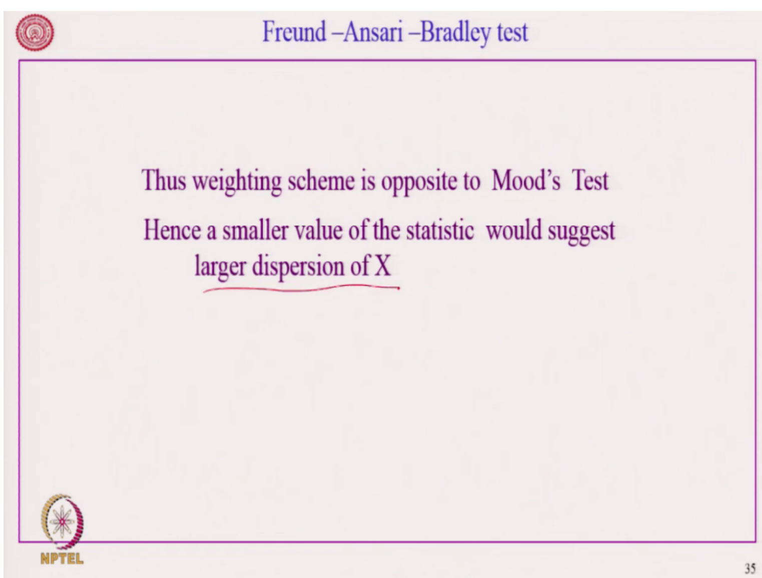
The effect is that it is because it is opposite to Mood's test a small value of the statistic would suggest that there is a larger dispersion for X.

(Refer Slide Time: 32:22)



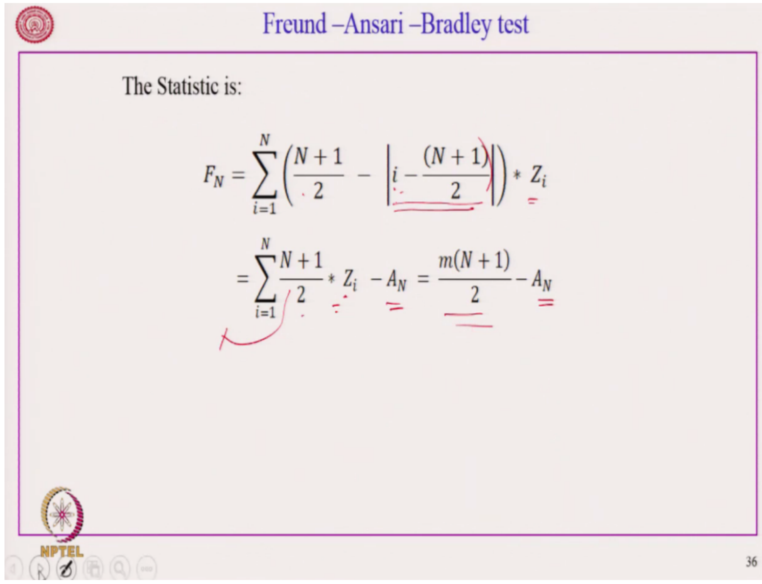
Because if X is having larger dispersion then these values are more likely to be coming from x and these values are more likely to be coming from y and as a consequence, the weighted sum of the x values is going to be smaller.

(Refer Slide Time: 32:42)



Hence, the smaller value of the statistic would suggest that X values are more dispersed in comparison with Y.

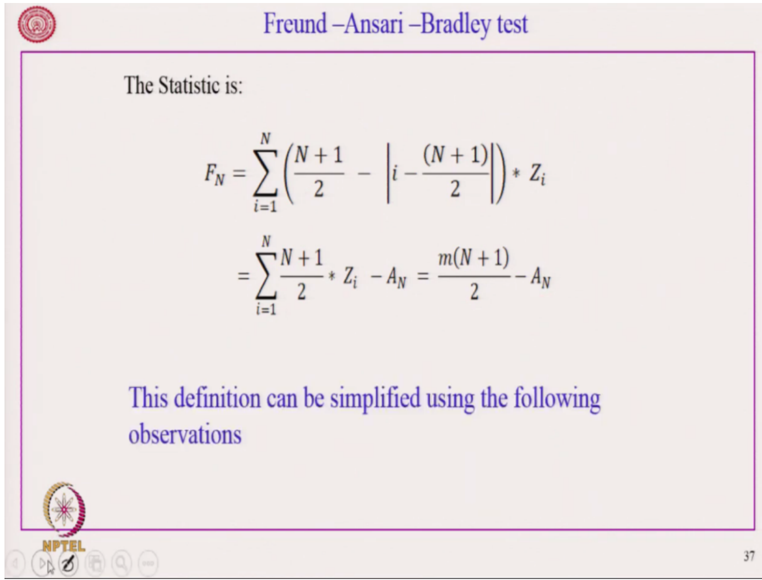
(Refer Slide Time: 32:52)



Now, the statistic is N plus 1 by 2 minus modulus of i minus N plus 1 by 2 whole thing multiplied by Zi. Therefore, you understand that for the extreme values, this is high and therefore, as I am subtracting it from N plus 1 by 2, this value is going to be small. On the other hand for the central values, suppose it is odd then the most central value the i is going to be N plus 1 by 2, therefore, this whole thing is going to be 0 and therefore, it will get the weight N plus 1 by 2.

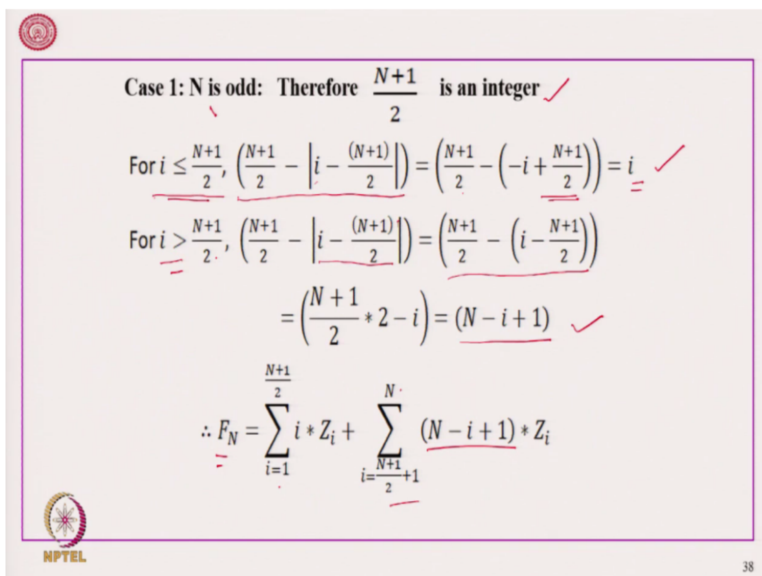
I hope the concept is clear. Now, let us simplify it. So, let us write it as N plus 1 by 2 times Zi minus A_N, because this is the element we have already computed to be A_N is equal to m times N plus 1 by 2 minus A_N why, because N plus 1 by 2 you can take out of the summation and Zi will be 1 only for aim of them, which are coming from x thus this sum is going to be m into N plus 1 by 2.

(Refer Slide Time: 34:16)



Now, we further simplify it in the following way.

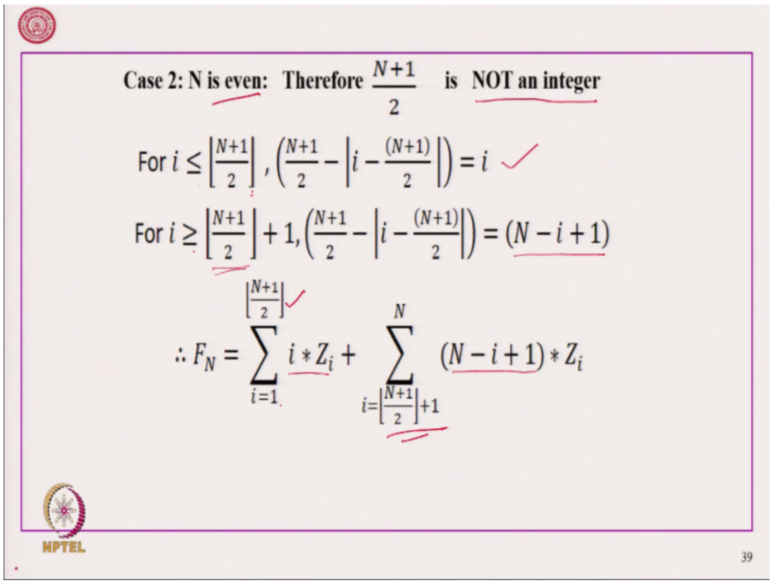
(Refer Slide Time: 34:20)



Consider N is odd, therefore, N plus 1 by 2 is an integer. Therefore, for i less than equal to N plus 1 by 2, this entire quantity, we can write it as N plus 1 by 2 minus because i is less than N plus 1 by 2. So, this is going to be negative of that one, therefore, minus i plus N plus 1 by 2 which is equal to i. But if i is greater than N plus 1 by 2 then this quantity is positive. The absolute value won't change any sign. Therefore, this is written as N plus 1 by 2 minus i minus N plus 1 by 2, which comes out to be N minus i plus 1.

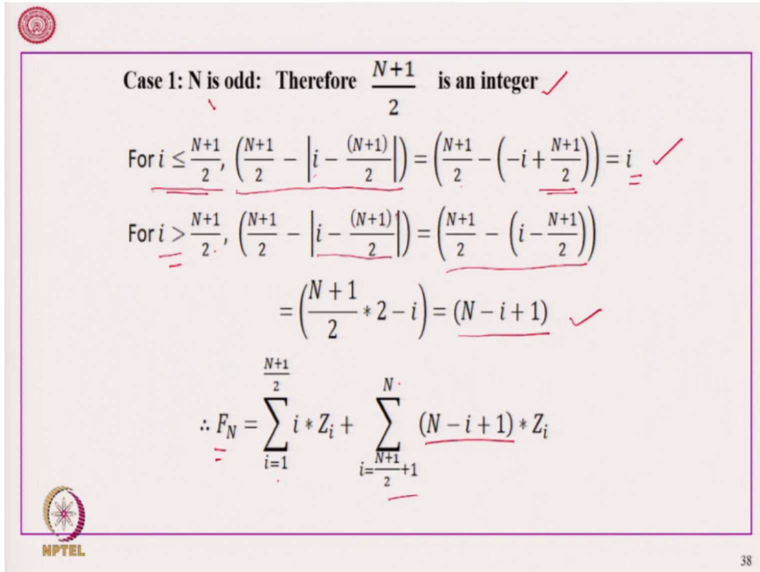
Thus, if i is less than equal to N plus 1 by 2, then we can simplify it to i. If i is greater than N plus 1 by 2, then this is equal to N minus i plus 1. Therefore, the statistic can be written as i is equal to 1 to N plus 1 by 2 i times Zi plus i is equal to N plus 1 by 2 plus 1 to N, N minus i plus 1 times Zi. This is when N is odd.

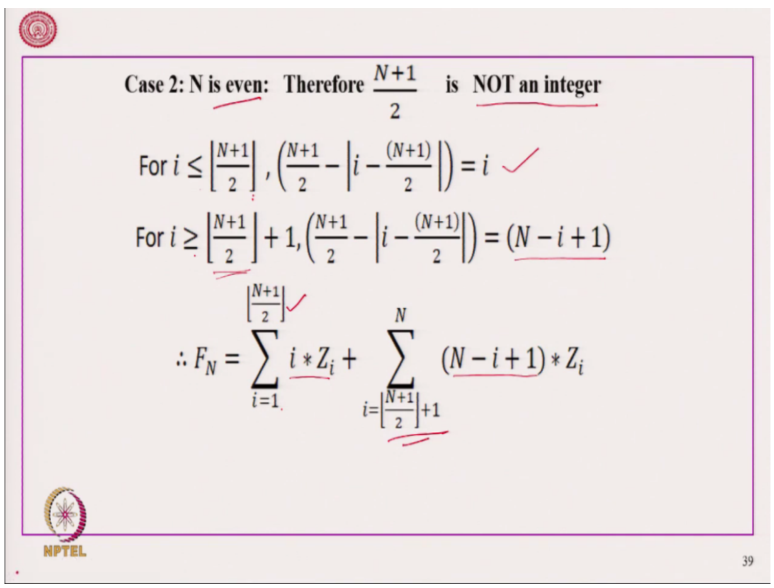
(Refer Slide Time: 35:52)



But, if N is even, then N plus 1 by 2 is not an integer. Therefore, we have to consider equality in both the cases and we are going to take the floor function. And in a very similar way, we can see that for i less than equal to the floor of N plus 1 by 2, we get the value i and i greater than equal to the floor value of N plus 1 by 2, we get N minus i plus 1. Therefore, the Freund-Ansari statistic for N is even is equal to F_N is equal to i is equal to 1 to N plus 1 by 2 floor summation of i star Zi plus i is equal to N plus 1 by 2 plus 1 to N, N minus i plus 1 times Zi. So that is the statistic when N is even.

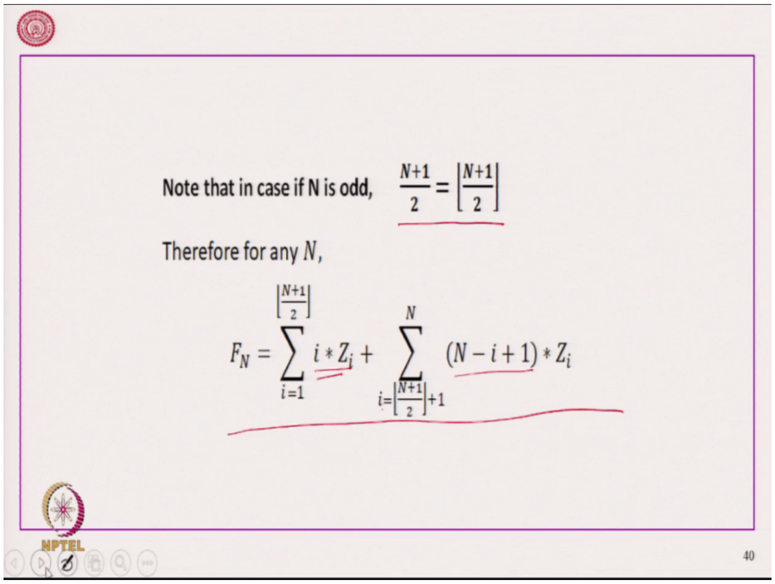
(Refer Slide Time: 37:01)





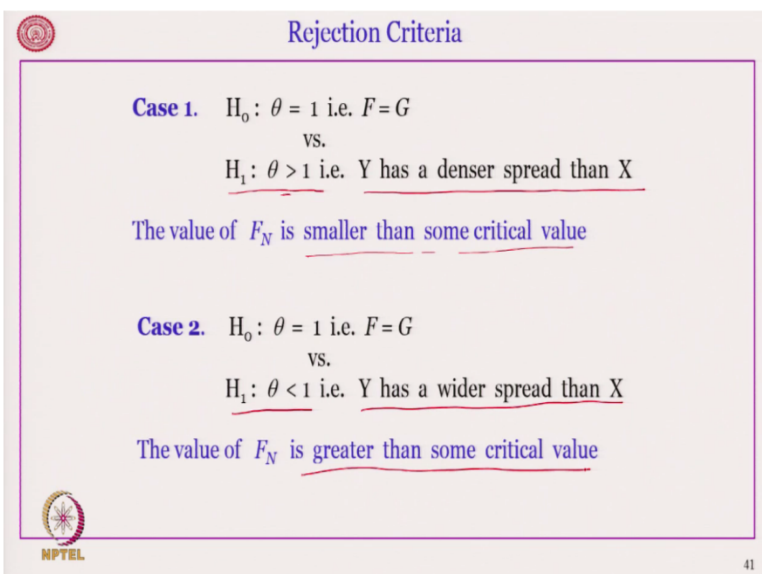
So apparently there is a small difference between the two statistic here it is like this, but in case of even it is slightly more complicated because of the floor function.

(Refer Slide Time: 37:12)



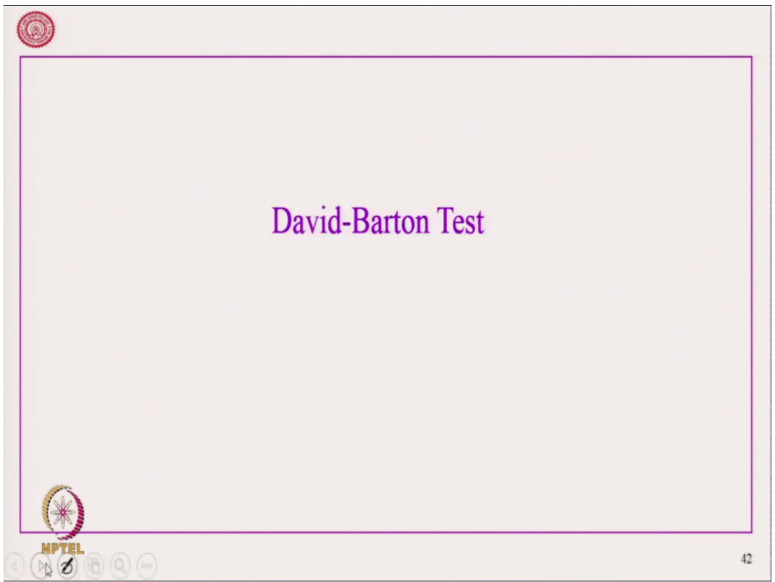
But, we notice that when N is odd N plus 1 by 2 is same as the floor of N plus 1 by 2. Therefore, we can write both odd and even by a single function FN is equal to 1 to floor of N plus 1 by 2 i times Zi plus i is equal to floor of N plus 1 by 2 plus 1 to N, N minus i plus 1 star Zi. So, this is what the Freund-Ansari statistic which is used for measuring or comparing the dispersion between the two populations.

(Refer Slide Time: 38:02)



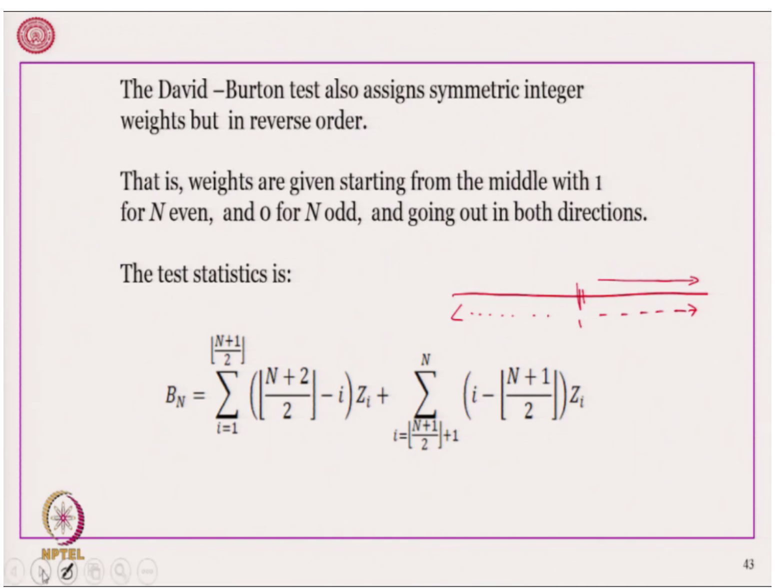
So, the rejection region for the null hypothesis is as follows. If the alternative is H1 is that theta is greater than 1, that is Y has a denser spread than X, then the value of F_N is smaller than some critical value. Because, we have said it is going to be opposite of the Mood's test. On the other end if the alternative is that H1 is theta less than 1 that is Y has a wider spread than X, the value of FN is going to be greater than some critical value.

(Refer Slide Time: 38:50)



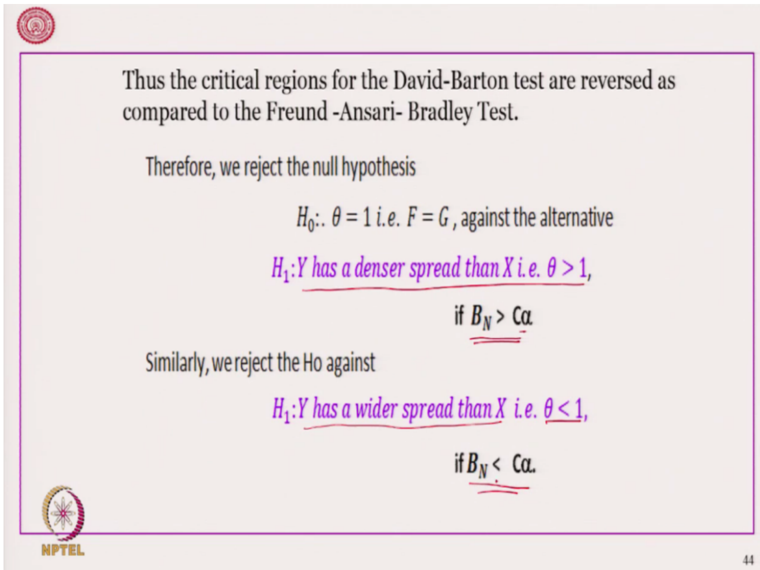
Let me talk about another test of very similar nature, which is called the David-Barton test.

(Refer Slide Time: 38:57)



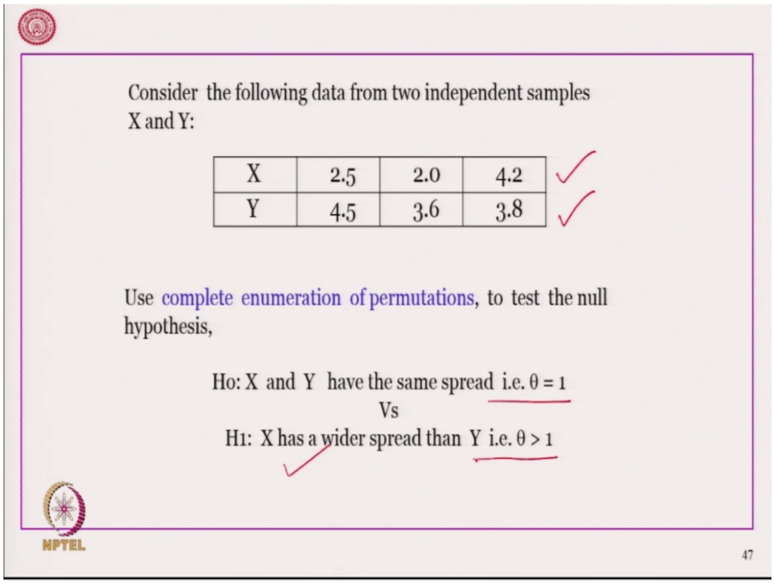
It is also very, very similar to the Freund-Ansari test, but there is a subtle difference here the widths are given starting from the middle with 1 and as it goes farther to the end, then the weights are increasing. So, the middle weights are going to be 1, but weights is going to increase as we are going farther from the central region. Therefore, what is going to happen, everything will remain the same, but it will alter the Rejection Criteria.

(Refer Slide Time: 39:41)



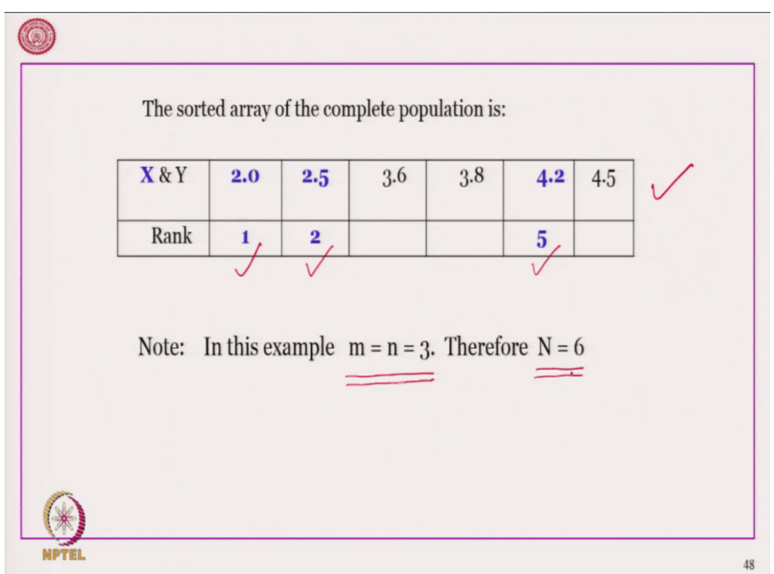
What is going to happen if Y has a denser spread than X that is theta is greater than 1, then the rejection criteria is going to be the value of statistic is greater than the critical value and similarly, if the alternative is Y has a wider spread than X that is theta is less than 1, then the condition is going to be BN is less than some critical value. So, let me illustrate the above algorithms with an example.

(Refer Slide Time: 40:20)



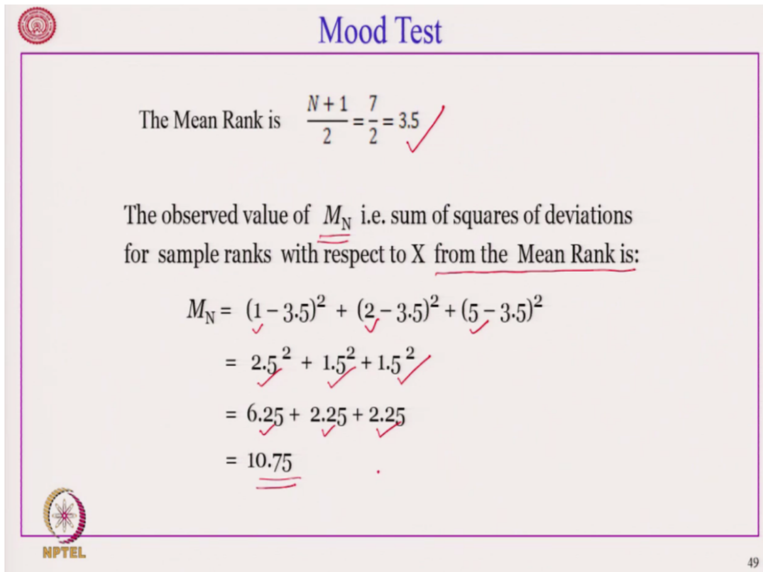
Consider the following data from two independent samples. So, there are three observations from X and three observations from Y. And we want to check if both of them have the same spread that is the theta is equal to 1 that is our null hypothesis and our alternative is theta is greater than 1. Why? Because if we look at this, we feel that X has more spread it is going from 2 to 4.2. On the other hand the spread of Y is less. Therefore, the valid alternative to test is whether X has a wider spread than Y or not?

(Refer Slide Time: 41:10)



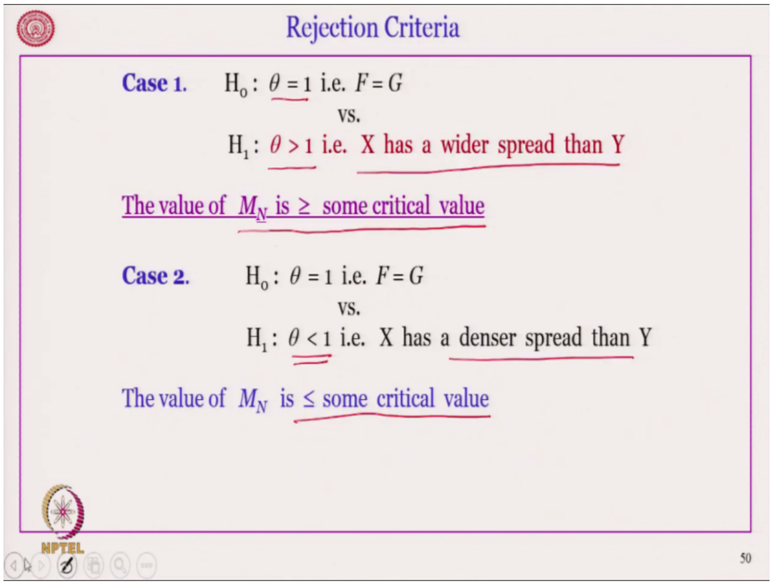
So, what we do, we pull together all the observations and make a sorted array of the complete population. In this, now, we obtain the rank of the X observations. Therefore, we get XR, 1, 2 and 5th position in the sorted array. Now, note that here m is equal to n is equal to 3, because number of observations for both X and Y is 3. Therefore, total number of observations is 6.

(Refer Slide Time: 41:48)



Let us first apply Mood test, what we are doing? First we obtain the value of the Mean Rank which we know is equal to N plus 1 by 2 is equal to 3.5. The observed value of the M_N, the statistic is going to be the sum of squares of deviations of the sample ranks with respect to X from the Mean Rank. Mean Rank is 3.5 and for X, the sample ranks are 1, 2 and 5. Therefore, the value of MN is equal to 1 minus 3.5 that is 2.5 square 2 minus 3.5 that is 1.5 square and 5 minus 3.5 which is again 1.5 square. Therefore, their sum is 6.25, 2.25 plus 2.25, which is equal to 10.75. So this is the observed value of the statistic M_N.

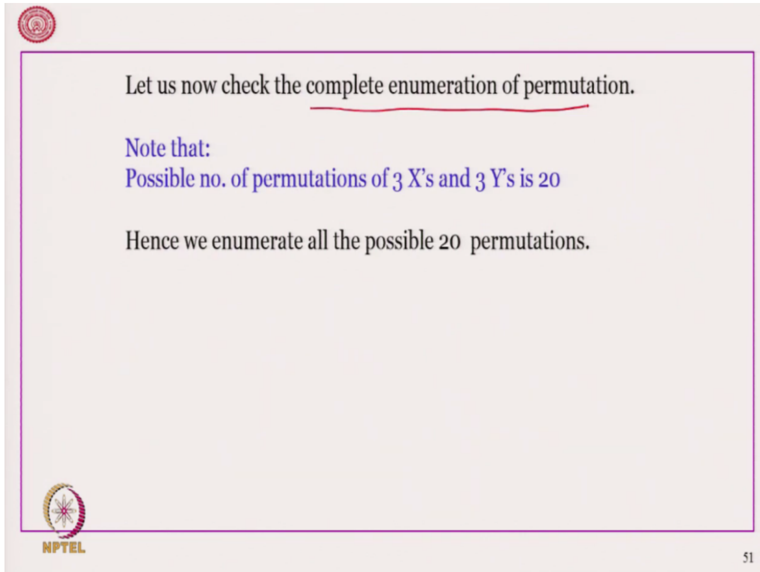
(Refer Slide Time: 43:01)



Now, what is the Rejection Criteria, we have already seen that if we are testing theta is greater than 1, as opposed to theta is equal to 1, then what we are checking is that X has a wider spread than Y and in that case, the critical value is going to be if MN is greater than some critical value. So in this case the test criterion is going to be that MN is greater than equal to some critical value.

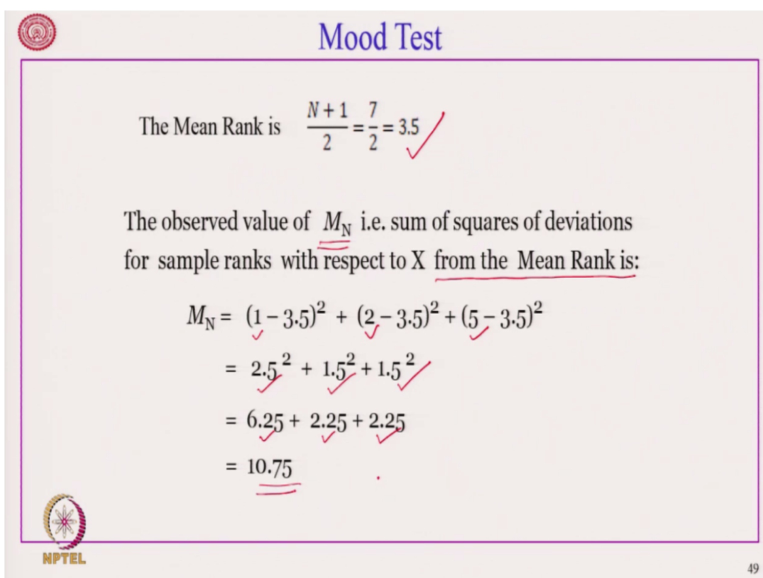
However, if the case would have been that theta is less than 1, that is X has a denser spread than Y, then M_N had to be less than equal to some critical value. Since we are testing theta is greater than 1, we will look at if M_N is greater than equal to some critical value.

(Refer Slide Time: 43:49)



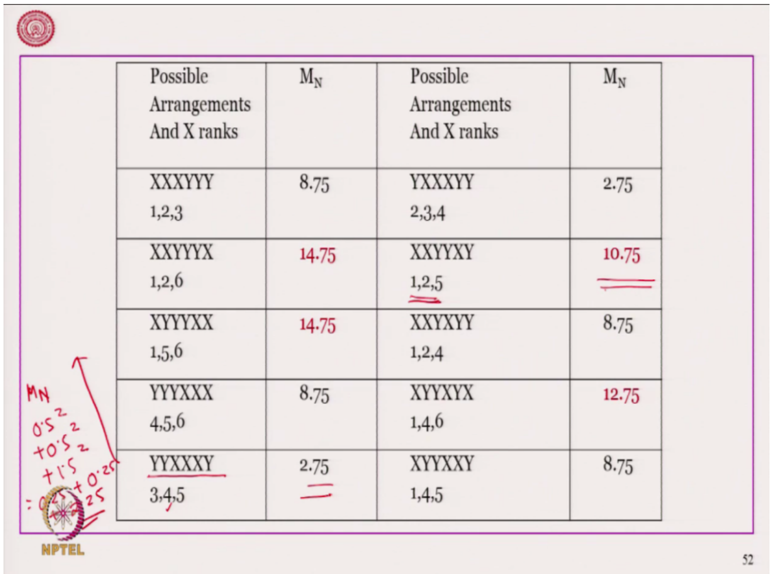
Now tables are very rare for these tests. So what we are doing, we are looking at a complete enumeration of all the possible permutations. And for each one of them, we will calculate the value of M_N. And then we will see what can be the probability of getting the value that…

(Refer Slide Time: 44:22)



We have obtained is equal to 10.75.

(Refer Slide Time: 44:27)



Possible Arrangements And X ranks	M_N	Possible Arrangements And X ranks	M_N
XXXYYY 1,2,3	8.75	YXXXYY 2,3,4	2.75
XXYYYX 1,2,6	14.75	XXYYXY 1,2,5	10.75
XYYYXX 1,5,6	14.75	XXYXYY 1,2,4	8.75
YYYXXX 4,5,6	8.75	XYYXYX 1,4,6	12.75
YYXXXY 3,4,5	2.75	XYYXXY 1,4,5	8.75

So, we have 6 elements, three Xs, and three Ys, there are 20 possible permutations. So for each one of them, we are computing the value of MN, our values were XXYYXY that means that the first second and fifth observations were X observations and we obtain the value 10.75. Let me illustrate with some other say for example, this one X is occupying the positions 3 comma 4 comma 5.

Therefore, the value of M_N in this case is going to be 3 minus 3.5 that is 0.5 square plus 4 minus 3.5 that is another is 0.5 square plus 5 minus 3.5 is equal to 1.5 square is equal to 0.25 plus 0.25 plus 2.25, which is, is equal to 2.75. So, this is a bit laborious, but we have done it for your sake. And for these 10 permutations, we have computed the value.

(Refer Slide Time: 45:57)

Possible Arrangements And X ranks	M_N	Possible Arrangements And X ranks	M_N
XXXYYY 1,2,3	8.75	YXXXYY 2,3,4	2.75
XXYYYX 1,2,6	14.75	XXYYXY 1,2,5	10.75
XYYYXX 1,5,6	14.75	XXYXYY 1,2,4	8.75
YYYXXX 4,5,6	8.75	XYYXYX 1,4,6	12.75
YYXXXY 3,4,5	2.75	XYYXXY 1,4,5	8.75

52

For another 10 permutations, we have these values computed, as I said, it is laborious, but since we are doing a complete enumeration, we had to do it. Now, let us going to check what is the probability that MN is greater than equal to 10.75. In this page we are finding two of them.

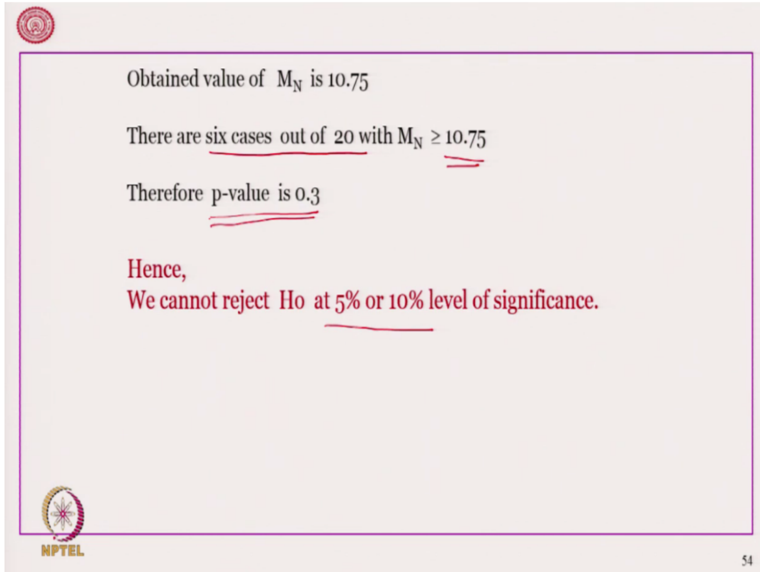
(Refer Slide Time: 46:27)

Possible Arrangements And X ranks	M_N	Possible Arrangements And X ranks	M_N
XXXYYY 1,2,3	8.75	YXXXYY 2,3,4	2.75
XXYYYX 1,2,6	14.75	XXYYXY 1,2,5	10.75
XYYYXX 1,5,6	14.75	XXYXYY 1,2,4	8.75
YYYXXX 4,5,6	8.75	XYYXYX 1,4,6	12.75
YYXXXY 3,4,5	2.75	XYYXXY 1,4,5	8.75

52

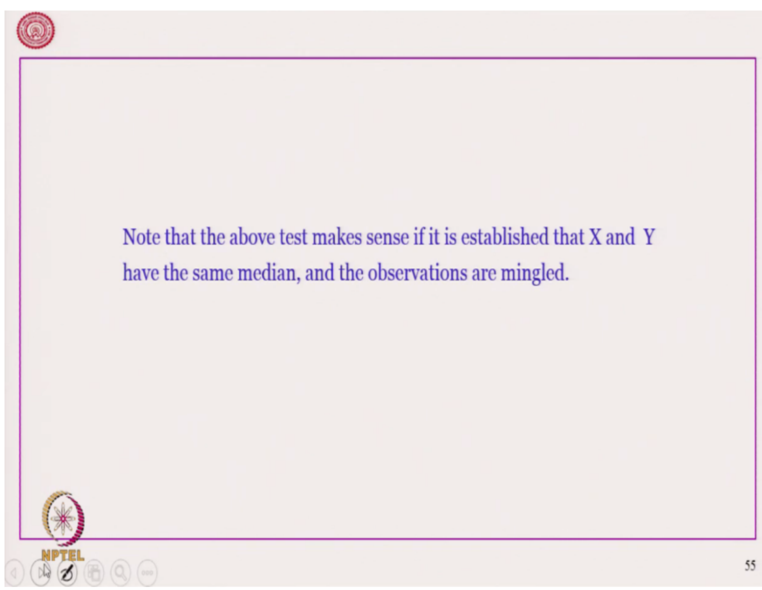
On the previous page there are 1, 2, 3 and 4, four of them.

(Refer Slide Time: 46:35)



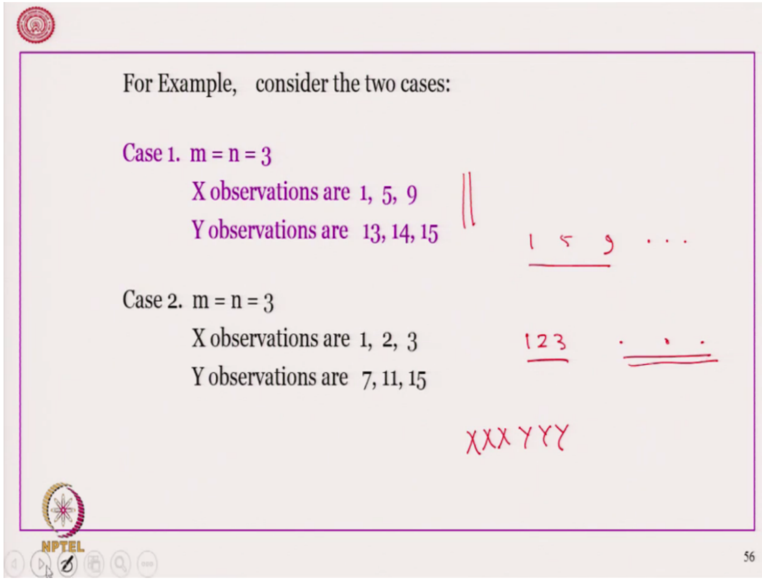
Therefore, what we will find is that 6 out of 20 cases, we are having the value greater than equal to 10.75. Therefore, the p value is equal to 0.3. Therefore, we cannot reject the H nought at 5% or 10% level of significance. So, that is with respect to Mood test.

(Refer Slide Time: 47:08)



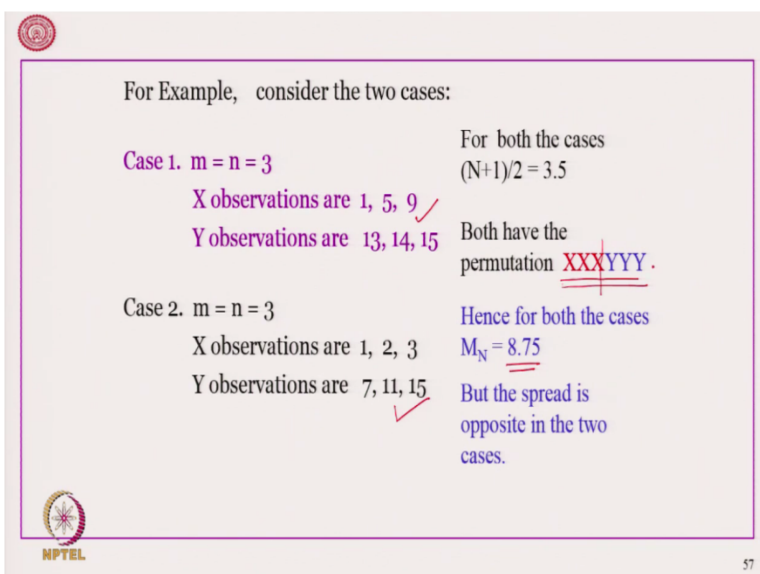
Note that the above test makes sense, if it is established that X and Y have the same median and the observations are mingled, what I mean by that, that X and Y observations are mixed. Suppose that is not the case.

(Refer Slide Time: 47:26)



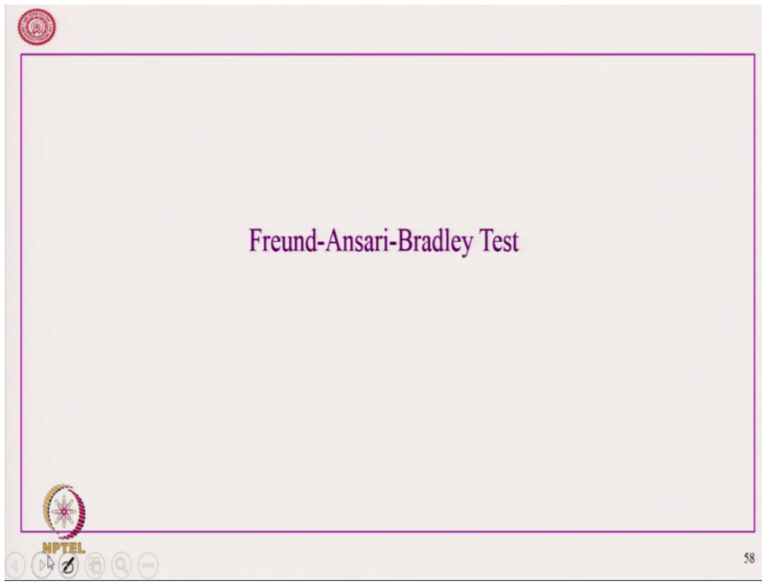
For example, consider this case 1, we have X as the observations are 1, 5 and 9 and Y observations are 13, 14 and 15. In the second case, our X observations are 1, 2, 3 and y observations are 7, 11 and 15. When we are computing the M_N, what is going to happen? In both the cases we are getting XXX YYY, because the smallest threes are X and the bigger three are Y observations.

(Refer Slide Time: 48:10)



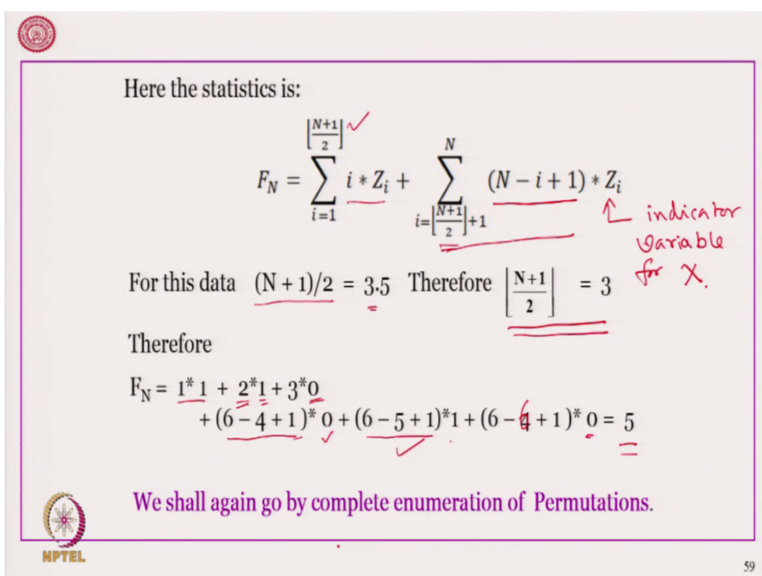
So, for both the cases, the Mean Rank is going to be in N plus by 1 by 2, as we have calculated both have the permutation XXXYYY. Hence, for both the cases, the value of M_N is coming out to be 8.75 this we have already seen in our table. But if we look at that, we can see that in the first case X has much wider spread, but in the second case, Y has much wider spread, but that we cannot detect using these tests. This is because X and Y are not at all mingled here. We have got the X observations on the left side and Y observations are on the right side.

(Refer Slide Time: 49:03)



Now, for the same data, we are going to apply the Freund-Ansari-Bradley test.

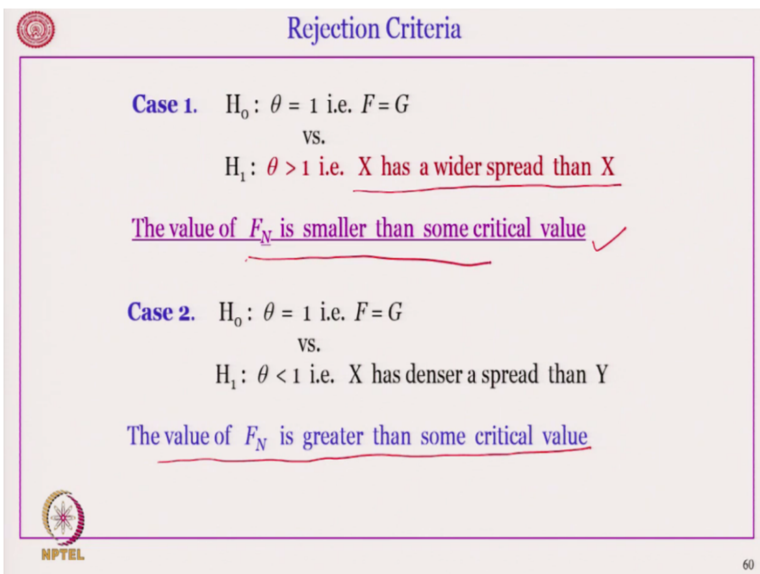
(Refer Slide Time: 49:11)



We already know the test statistic, which is F_N which is I equal to 1 to floor of N plus 1 by 2 multiplied by i star Zi plus i is equal to N plus 1 by 2 floor plus 1 to N, N minus i plus 1 star Zi where Zi is the indicator variable for X. Therefore, for the given data N plus 1 by 2 is equal to 3.5 and therefore floor of that is equal to 3. Therefore, FN is equal to when i is equal to 1, I am getting the value 1 and indicator variable Z1 is equal to 1, when i is equal to 2 I am getting 2 multiplied by the indicator variable, which is again 1, but for i is equal to 3, the indicator variable is taking the value 0, because that is a Y observation.

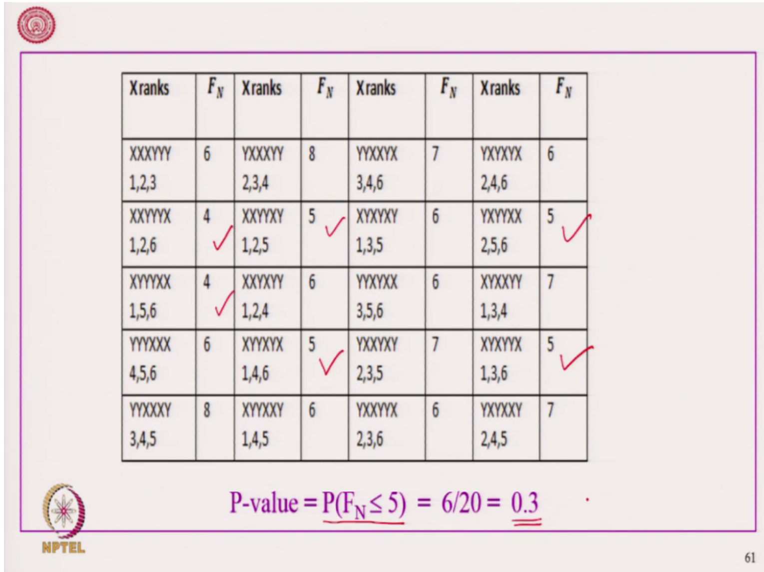
Now let us look at this part for 4, 5 and 6. When i is equal to 4, N minus i plus 1 is equal to 6 minus 4 plus 1 multiplied by 0 because Z4 is equal to 0. Similarly, 6 minus 5 plus 1 when i is equal to 5 multiplied by 1 because Z5 is 1. We know that the fifth observation in the sorted array is an X observation and for i is equal to 6. This is going to be 6. So, it is 6 minus 6 plus 1, but it is multiplied by 0. Therefore, what we are getting is that the value of FN is equal to 1 plus 2 plus 2 is equal to 5. Again, we will go for a complete enumeration of permutations.

(Refer Slide Time: 51:20)



We know that the Rejection Criteria is that X has a wider spread, which is our alternative if FN is smaller than some critical value. However, if X has a denser spread, then our rejection criteria would have been that the value of FN is greater than some critical value. But since this is our case, we are looking at FN is smaller than some critical value.

(Refer Slide Time: 51:20)



Xranks	F_N	Xranks	F_N	Xranks	F_N	Xranks	F_N
XXXYYY 1,2,3	6	YXXXYY 2,3,4	8	YYXXYX 3,4,6	7	YXYXYX 2,4,6	6
XXYYYX 1,2,6	4	XXYYXY 1,2,5	5	XYXYXY 1,3,5	6	YXYYXX 2,5,6	5
XYYYXX 1,5,6	4	XXYXYY 1,2,4	6	YYXYXX 3,5,6	6	XYXXYY 1,3,4	7
YYYXXX 4,5,6	6	XYYXYX 1,4,6	5	YXXYXY 2,3,5	7	XYXYYX 1,3,6	5
YYXXXY 3,4,5	8	XYYXXY 1,4,5	6	YXXYYX 2,3,6	6	YXYXXY 2,4,5	7

P-value = $P(F_N \leq 5)$ = 6/20 = 0.3

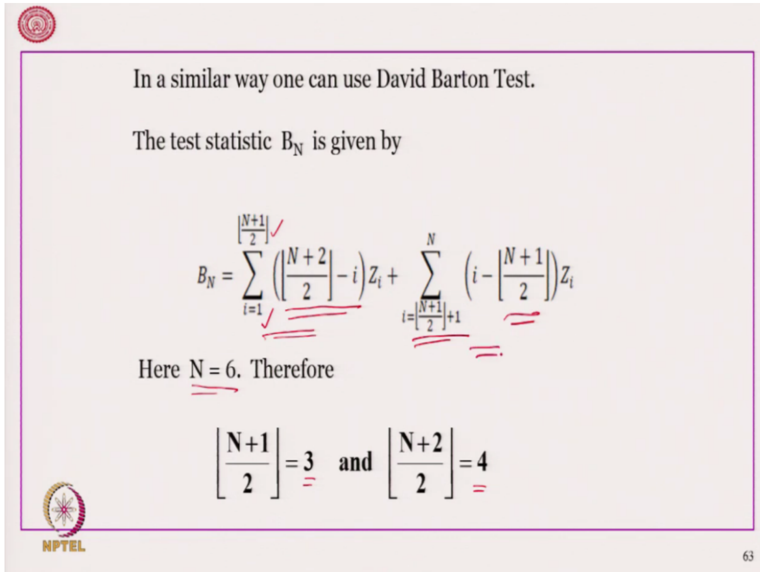
61

So what we do for all the 20 permutations, we compute the value of FN, so we have all those values. Therefore, probability FN is less than equal to 5, how many cases are there 1, 2, 3, 4, 5, 6. Therefore, again, there are 6 cases out of 20, where the statistic is less than equal to 5. Therefore, the sample P value is 6 by 20, which is, is equal to 0.3.

(Refer Slide Time: 52:29)

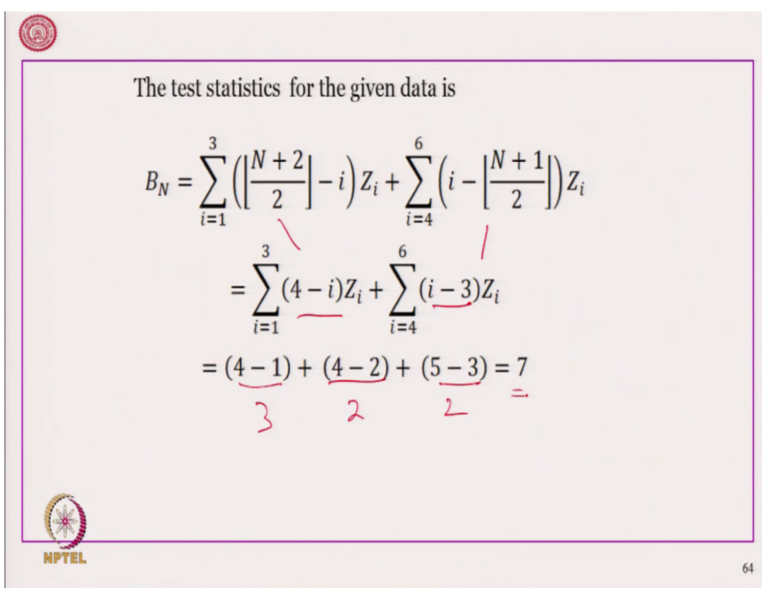
And therefore, we cannot reject H nought, because this is much larger than 0.05 or 0.1, which are the typical values for alpha.

(Refer Slide Time: 52:44)



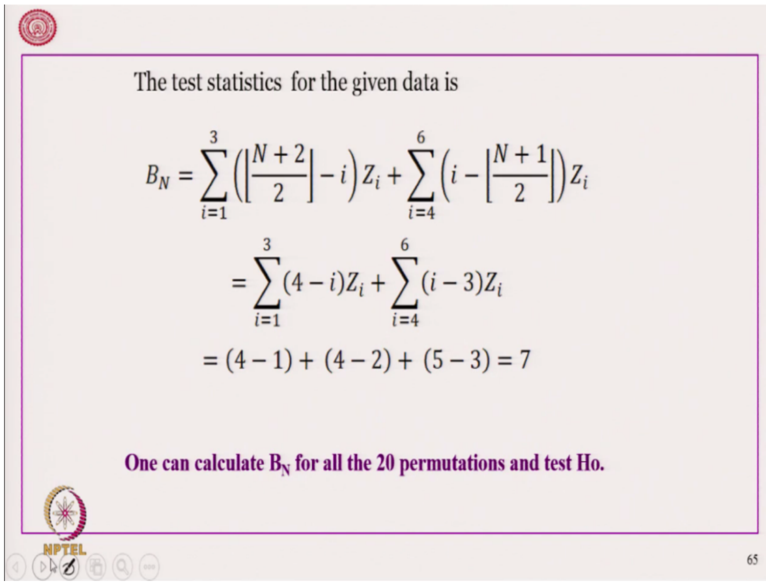
In a similar way, one can use the David Barton test where the test statistic B_N i is equal to i is equal to 1 to N plus 1 by 2 floor of N plus 2 by 2 minus i multiplied by Zi and plus i is equal to N plus 1 by 2 plus 1 to N, i minus capital N plus 1 by 2 floor, multiplied by Zi. This we know. Now, for this case, since capital N is equal to 6, N plus 1 by 2 is 3 and N plus 2 by 2, its value is 4. Therefore, this sum will take for 1, 2, 3, and this will take for 4 to 6.

(Refer Slide Time: 53:35)



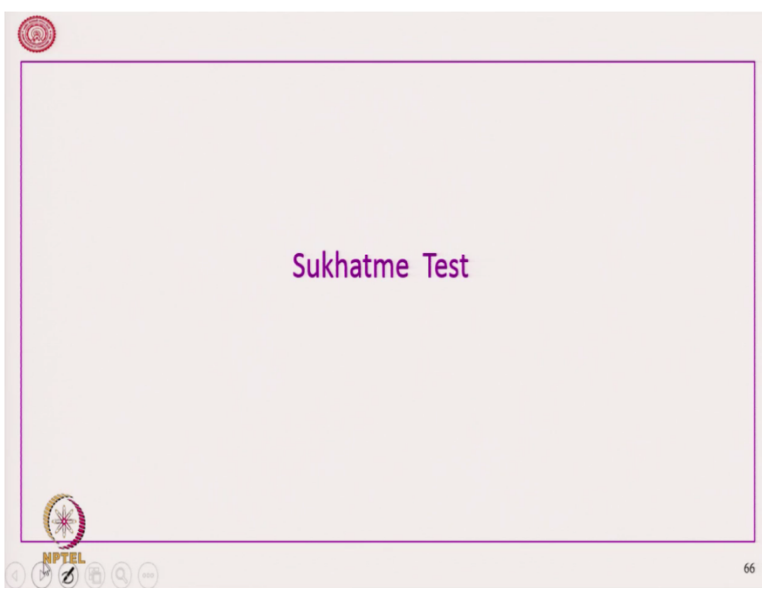
Therefore, B_N is going to be 4 minus i multiplied by Zi plus i minus 3 multiplied by Zi for 4 to 6 on this side and 1, 2, 3 for this side. Since we know that Zi is equal to 1 for 1, 2 and 5 therefore, we are getting 4 minus 1 plus 4 minus 2 plus 5 minus 3, which is, is equal to 3 plus 2 plus 2 is equal to 7. Therefore, the computed value of the statistic B_N for this data is seven.

(Refer Slide Time: 54:19)



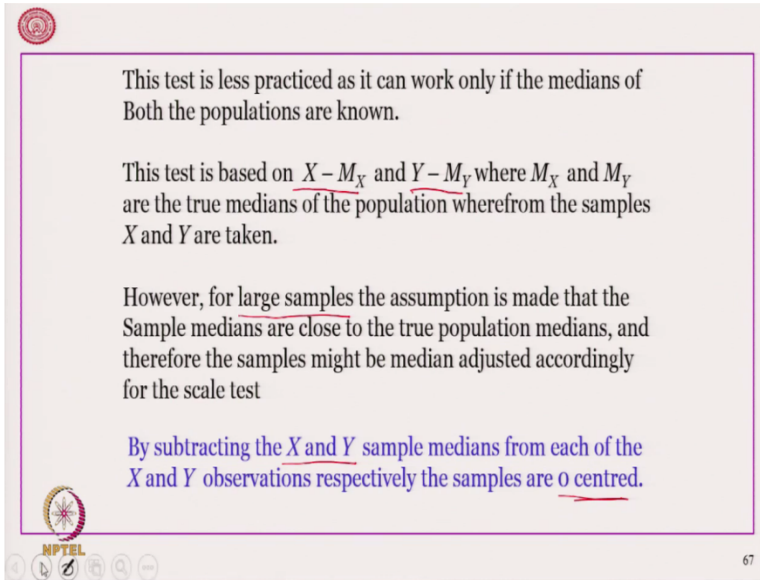
In a similar way, one should calculate the value of B_N for all the 20 permutations, and one should test H nought the way we have done for the 2 earlier cases. I leave it as an exercise for you. It is laborious, but it gives insight into how data works.

(Refer Slide Time: 54:19)



Let me now talk about another test, which is called Sukhatme Test.

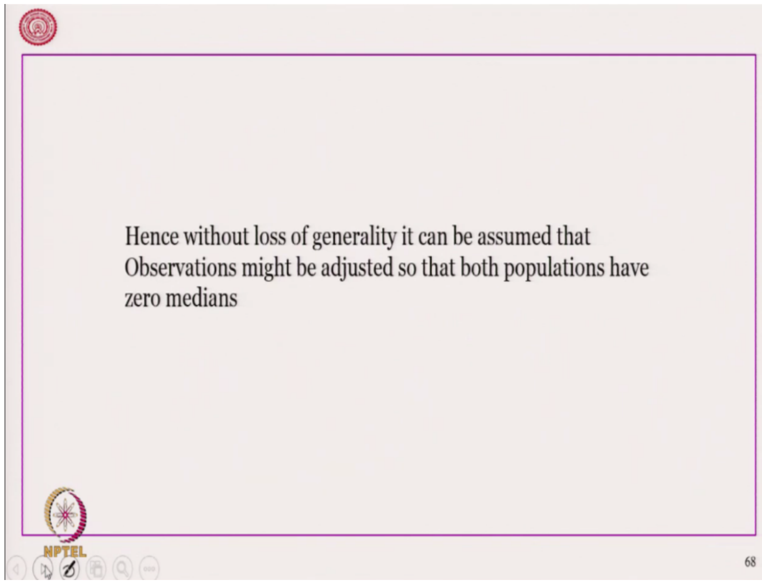
(Refer Slide Time: 54:52)



This test is less practiced as it can work only if the medians of both the populations are known. Therefore, it is very important, why? Because the data will be transformed to X minus M_X where M_X is the median for X and Y minus M_Y, so, that they can spread on both the sides of 0. Therefore, if F_X and M_Y are not known, truly we cannot use this, but typically what happens when the sample size is large, one can assume that the sample median is same as population median.

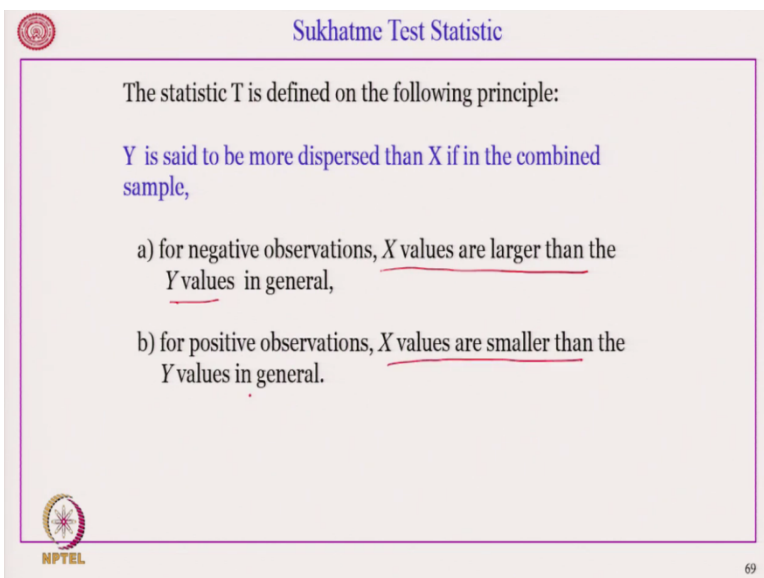
And therefore, we can use the sample median corresponding sample median from X and Y and so that they become 0 centered. And once they are like that, we can use Sukhatme test for testing of the difference in spread of X and Y. Or if X and Y have the same spread or not.

(Refer Slide Time: 55:55)



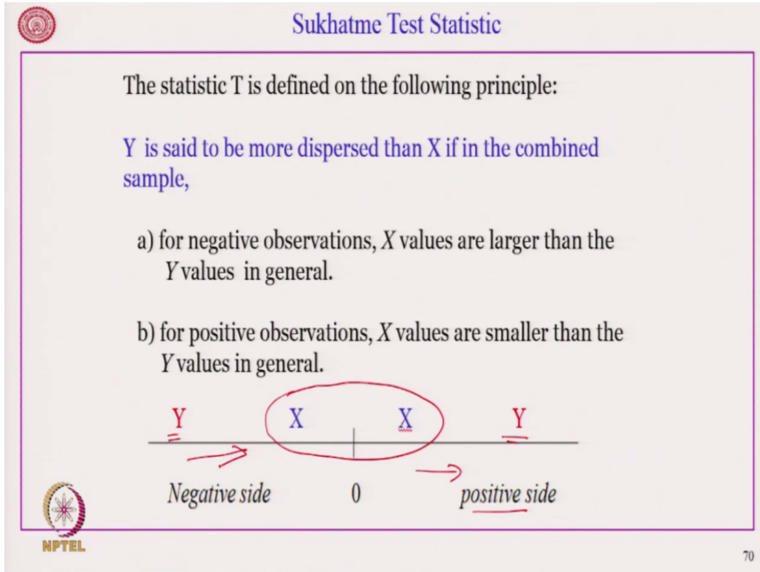
So, without loss of generality, we assume that the observations will be adjusted so that both the populations have 0 median.

(Refer Slide Time: 56:05)



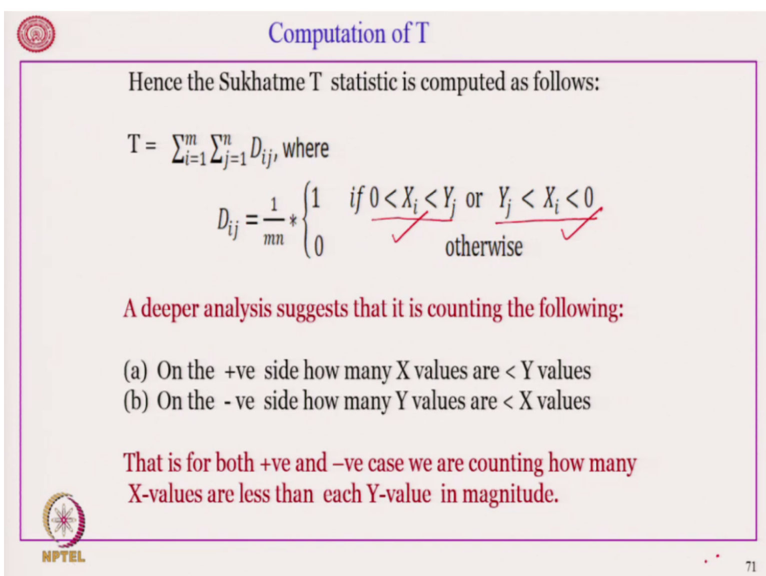
This statistic T is defined on the following principle, Y is said to be more dispersed than X, if in the combined sample, for negative observations, X values are larger than the Y values, and for positive observations, X values are generally smaller than Y values.

(Refer Slide Time: 56:29)



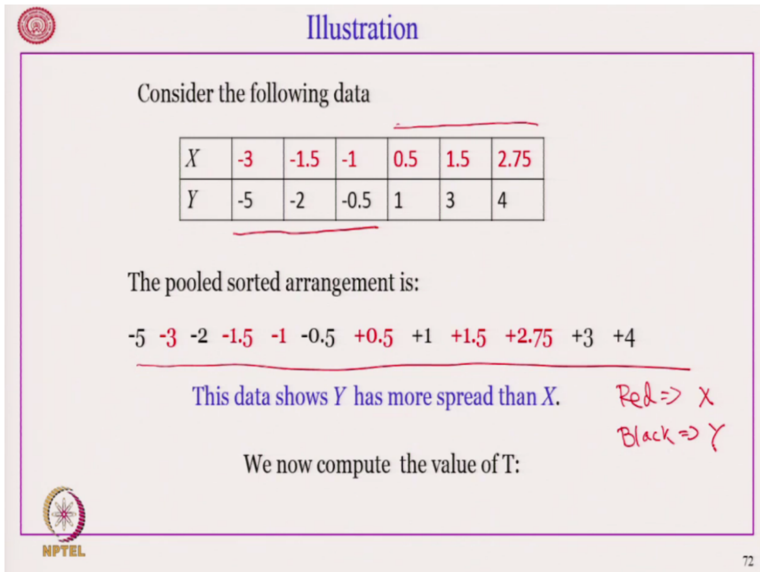
That means, if we look at the diagram, values are spread on the negative side and positive side of 0, Y is said to be more dispersed or higher spread, because the more outer values are Y, but the central values are from X. And we notice that for positive side, X values are smaller than Y, but on negative side X values are bigger than Y.

(Refer Slide Time: 57:02)



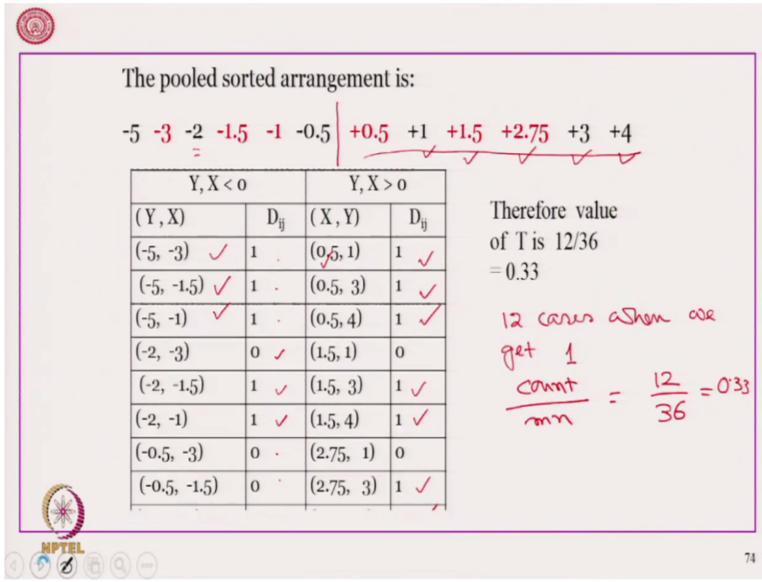
So, this Sukhatme T statistic is computed as follows. T is equal to sigma over i is equal to 1 to m, sigma over j is equal to 1 to n Dij, where Dij is equal to 1 upon mn and it is 1 if 0 less than Xi less than Yj or Yj less than Xi less than 0 and it is 0 otherwise. So a deeper analysis suggests that, it is counting the following. On the positive side how many X values are less than Y values and on the negative side how many Y values are less than X values. So together, we can see that for both positive and negative cases, we are counting how many X values are less than each Y value in magnitude. So let me illustrate with one example.

(Refer Slide Time: 58:23)



Consider the following data. X is equal to minus 3, minus 1.5, minus 1, 0.5, 1.5 and 2.75 and Y is equal to minus 5, minus 2, minus 0.5 and 1, 3, 4. So, these are on the negative side and these are on the positive side. When you pool them together and sort it, the arrangements look like this where red implies X and black implies Y. So looking at this, we can feel that Y has more spread than X. So let us compute the value of T.

(Refer Slide Time: 59:24)

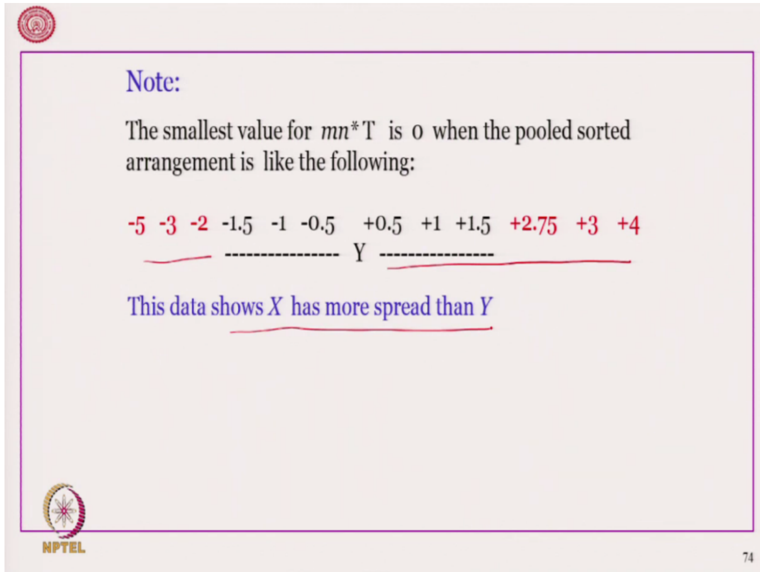


So corresponding to each Y value on the negative side, we are checking how many X values are greater than this. So let me make a partition here. Now, if we consider minus 5, then minus 3, minus 1.5, minus 1, all three X values are greater than minus 5, therefore, we get 1, 1, 1 here. For Y is equal to minus 2, we find that minus 3 is less than this therefore, it is 0, but minus 1.5 and minus 1 are greater than this therefore, they get 1.

With respect to minus 0.5 we find that all three are less than minus 0.5 therefore they are getting 0. Now, let us come to the positive side. We now look at for each X, it is smaller than how many Ys. So for 0.5 we find that it is smaller than 1, 3 and 4. So for all three of them we get positive values, for plus 1.5 it is smaller than only 3 and 4, so we get 1 here and for plus 2.75, again it is smaller than both, so they get plus one here.

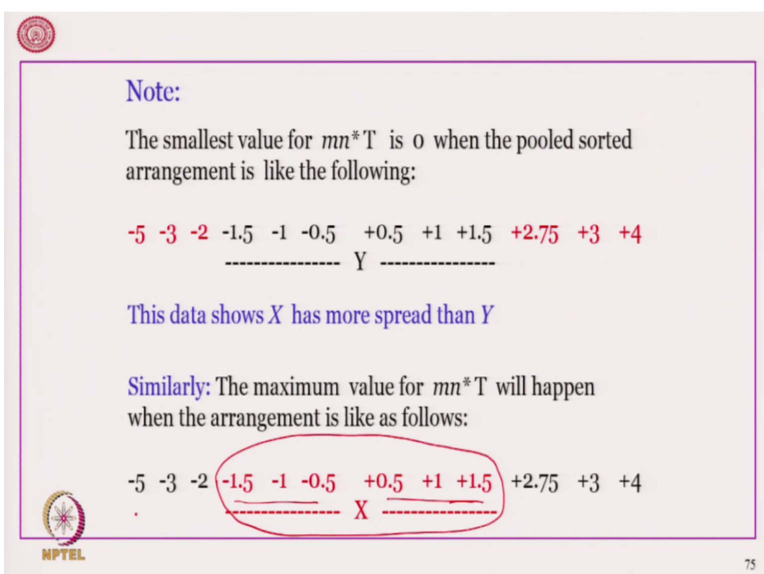
Therefore, there are 12 cases when we get 1. Therefore, as we have said the value of T is this count divided by mn, where m is equal to 6 and n is equal to 6 therefore 12 by 36 is equal to 0.33. So, that is how the T statistic is computed.

(Refer Slide Time: 01:01:42)



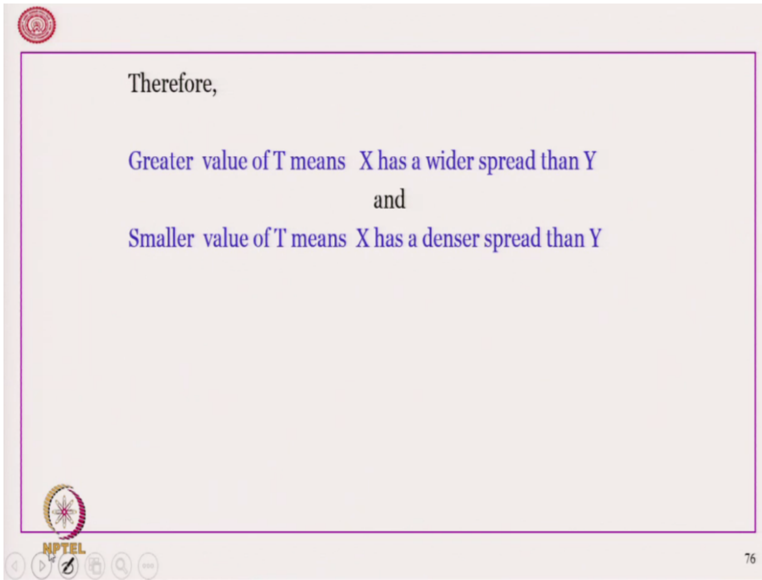
Now, the smallest value for mn times T that means, you are looking at the counting is 0 when the pooled sorted arrangement is something like this, all Y values are at the center. Therefore, no X is less than Y on the positive side and no X is greater than Y for negative side. Therefore, the value of the count is equal to 0. So in this case, the data shows that X has more spread than Y. Therefore, for lower values of T, we can see that the X will have more spread than Y.

(Refer Slide Time: 01:02:27)



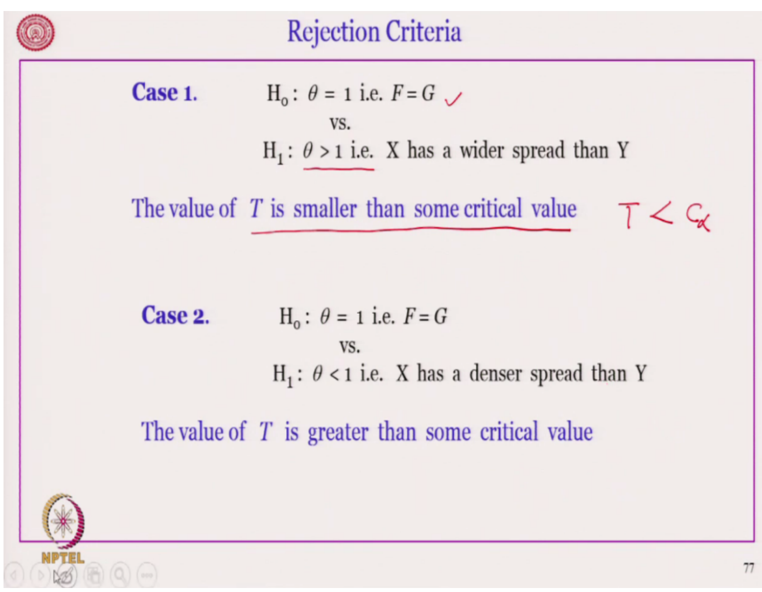
On the other hand, if we have an arrangement like this, where all the X values are at the center, then we can see that we will get the maximum count for T, because all X values are less than each Y value on the positive side and all X values are greater than each Y value on the negative side.

(Refer Slide Time: 01:02:55)



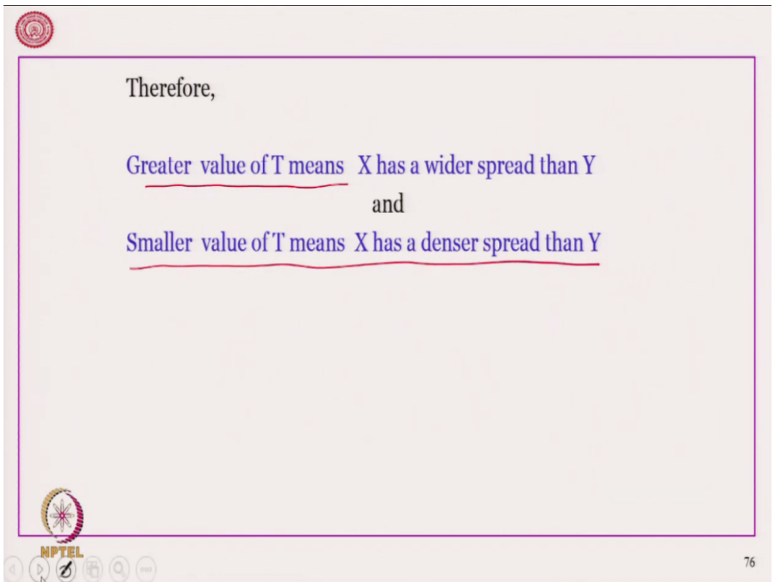
Therefore, greater value of T means X has a wider spread than Y and smaller value of T means X has a denser spread than Y.

(Refer Slide Time: 01:03:12)



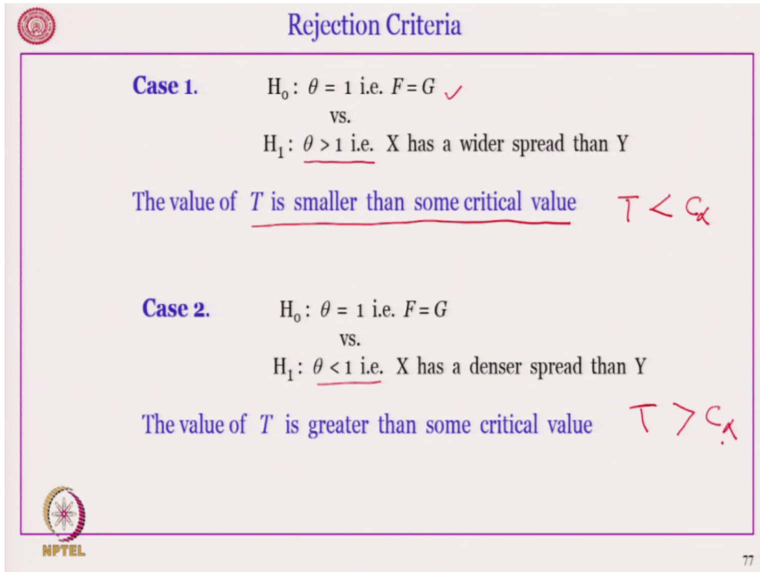
Therefore, what is going to be the Rejection region, suppose the alternative is theta greater than 1 that is X has a wider spread then in that case, we have seen that greater value of T means X has wider spread. Therefore, we expect the value of T should be big, but it need not be the maximum, it can be somewhat less, but if it is too small, then we will have to reject the null hypothesis in favor of the alternative. Therefore, the critical value is going to be that T is less than some C alpha. In a very similar way, when X has a denser spread than Y, then we know that the value should be small.

(Refer Slide Time: 01:04:07)



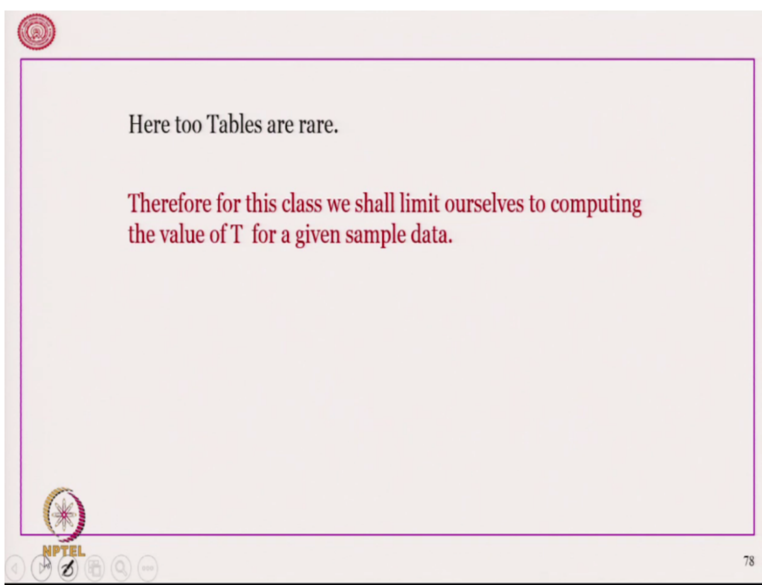
Smaller value of T means X has a denser spread than the Y.

(Refer Slide Time: 01:04:13)



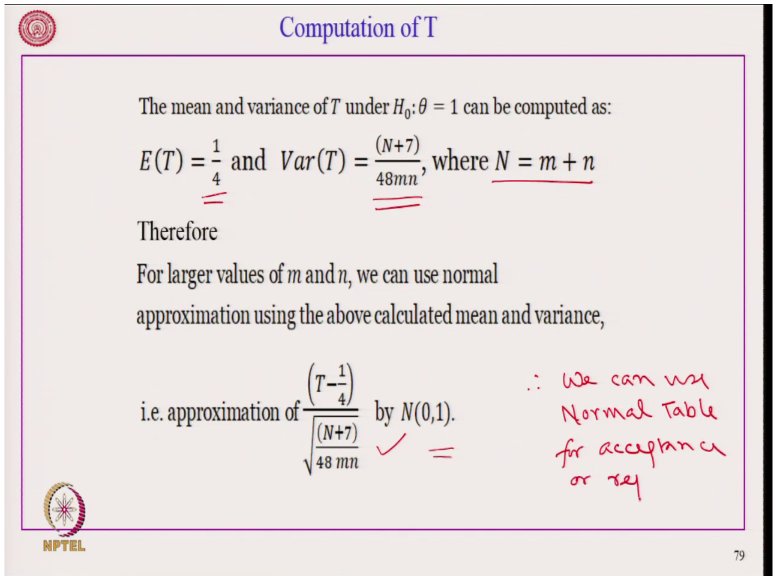
It need not be 0, it can be a little bit more, but if it crosses a threshold, then we are going to reject the null hypothesis in favor of the alternative. Therefore, T has to be greater than some critical value at alpha level.

(Refer Slide Time: 01:04:35)



But for Sukhatme Test also the tables are rare. So, we are not going to test the hypothesis the way we have done for others, but we limit ourselves to the computing of the T value for a sample data.

(Refer Slide Time: 01:04:54)



However, when the number of sample is large that is N is equal to m plus n is large, then it can be computed that expected value of T is going to be 1 by 4 and variance of T is going to be N plus 7 upon 48 mn. I am not going to calculate this, but what it suggest that, therefore, for larger values of capital N T minus 1 by 4 upon square root of N plus 7 divided by 48 mn is going to be distributed as normal 0, 1. Therefore, we can use normal table for acceptance or rejection.

Okay friends I stop here today. In the next class, I shall prove some mathematical results involving linear rank statistics. Thank you.