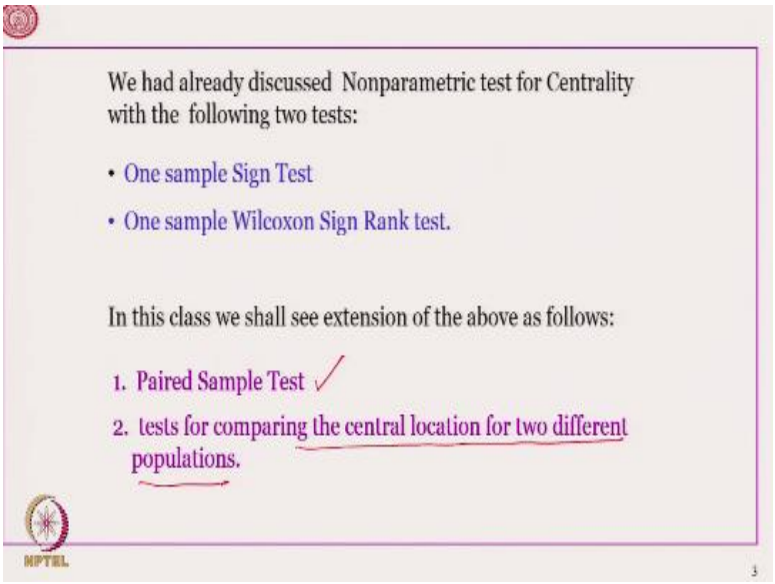
Nonparametric Statistical Inference
Professor. Niladri Chatterjee
Department of Mathematics
Indian Institute of Technology – Delhi
Lecture – 03
Nonparametric Statistical Inference

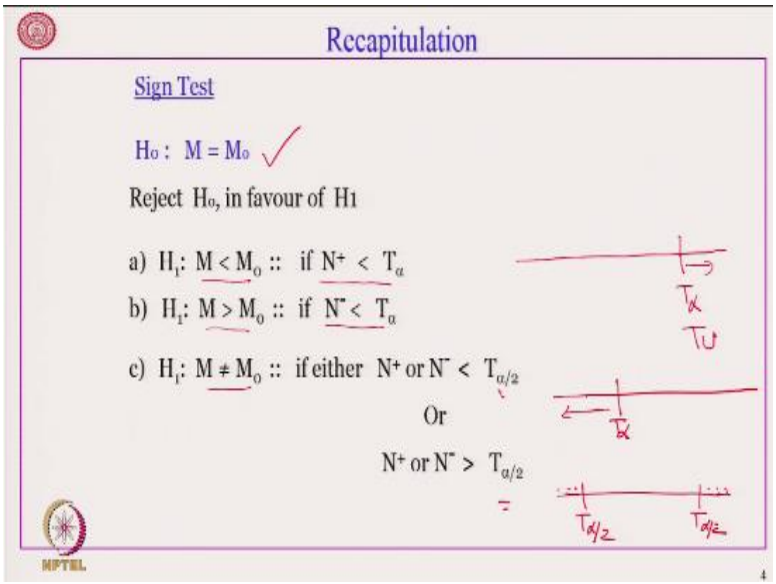
Welcome students to the MOOC series of lectures on Nonparametric Statistical Inference. This is lecture number 3.

(Refer Slide Time: 00:30)



In the last class, we had already discussed two nonparametric tests for centrality in particular we have studied one sample sign test and one sample Wilcoxon Signed Rank test. In this class we shall see extension of the above two as follows. We shall first study paired sample test and also we will look at test for comparing the central location for two different populations that means two sample test.

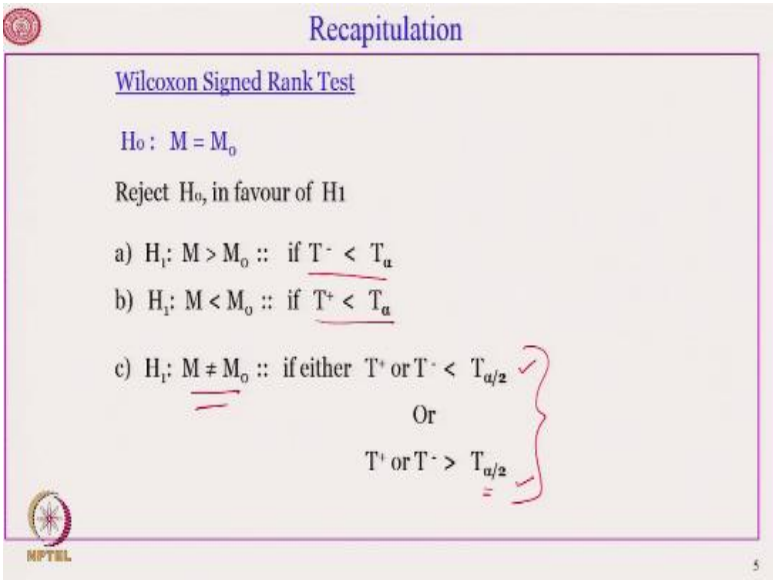
(Refer Slide Time: 01:10)



For Sign test you are testing if $M = M_0$ and there maybe three alternatives $M < M_0$ when the rejection criteria is that $N^+ < T\_\alpha$. If $M > M_0$ that means that $N^- < T_\alpha$ and if $M \neq M_0$ then either of N^+ or N^- is less than the corresponding critical values $T_{\alpha/2}$. Why? Because these are two-sided test.

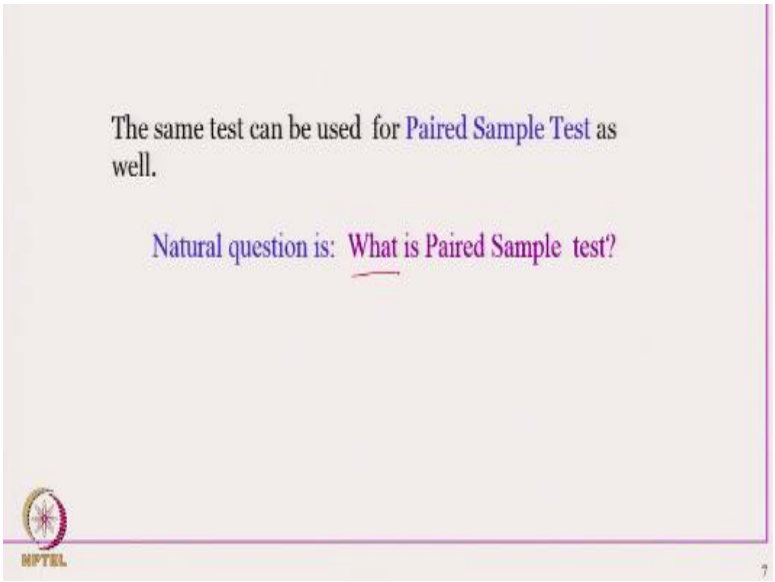
We have given you the explanation before, so if it is a one-sided test of size α then we have to look at if the value is coming out to be beyond the T_α so that we can reject. This is for upper side. If it is lower sided one-tailed then we will look at this and the value should fall here in order to reject the null hypothesis, but when it is two-sided test for equality of two populations this is $T_{\frac{\alpha}{2}}$ and this is also $T_{\frac{\alpha}{2}}$. And we will reject null hypothesis if the obtain statistics falls here or falls here.

(Refer Slide Time: 02:47)



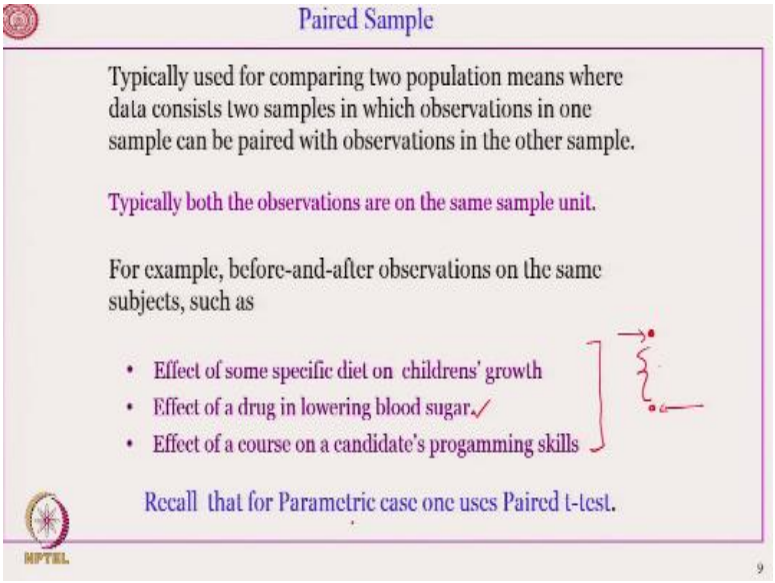
Similarly for Wilcoxon signed rank test we have the rejection criteria if $T^- < T_\alpha$ or $T^+ < T\_\alpha$ or for two-sided case we will look at if any one of them is less than $T_{\frac{\alpha}{2}}$ or greater than $T_{\frac{\alpha}{2}}$ depending upon whether you are doing a lower side or upper tail of the distribution.

(Refer Slide Time: 03:16)



Now the same test can be used for paired sample test as well. Natural question is what is paired sample test?

(Refer Slide Time: 03:30)

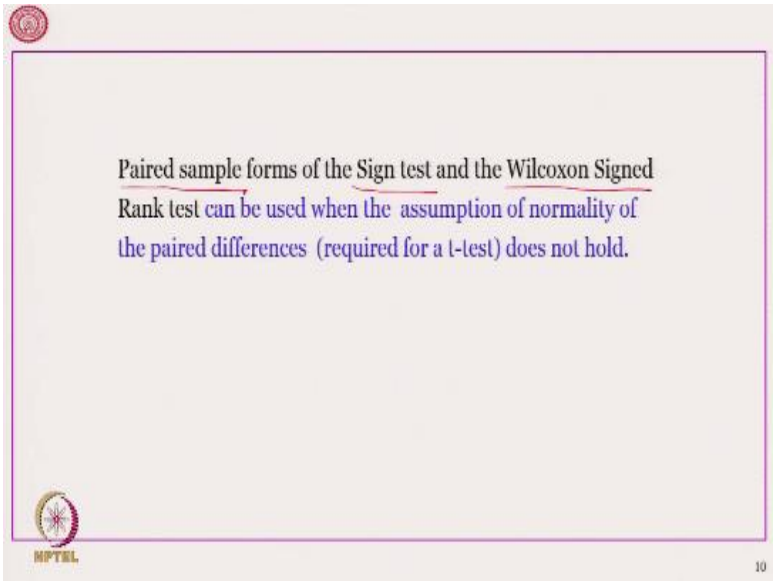


So let me first illustrate that. Typically a paired sample test is used for comparing two population means where data consists of two samples in which observations in one sample can be paired with observations in the other sample, okay. Typically both the observations are on the same sample unit. For example before and end observations on the same subjects, such as effect of some specific diet on children's growth or effect of a drug in lowering blood sugar or effect of a course on a candidate's programming skills.

Why these are called paired sample test? Because in all these cases we look at a subject and we see, say for example the blood sugar case before administration of the drug what is the value and after administration of the drug what is the value. So, this is the difference which may be considered the effect of the drug and then we need to see whether this effect is significant or not.

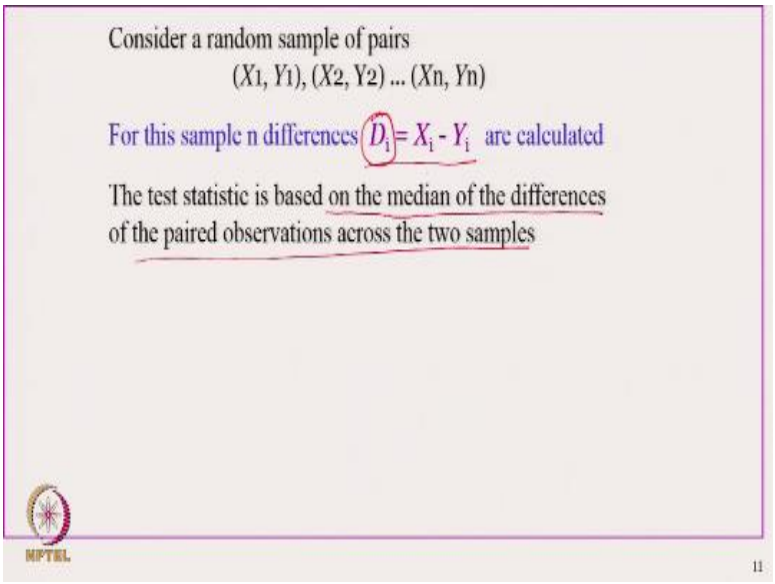
So that is why these are called paired sample test and recall that for parametric case we use paired test for paired t test for such situations.

(Refer Slide Time: 05:05)



But t test assumes normality of the data when we do not have that assumption then we cannot use t test and therefore we have to go for nonparametric testing of hypothesis and in particular we want to use the same Sign test and same Wilcoxon Signed Rank test on the data which is coming from paired samples.

(Refer Slide Time: 05:36)

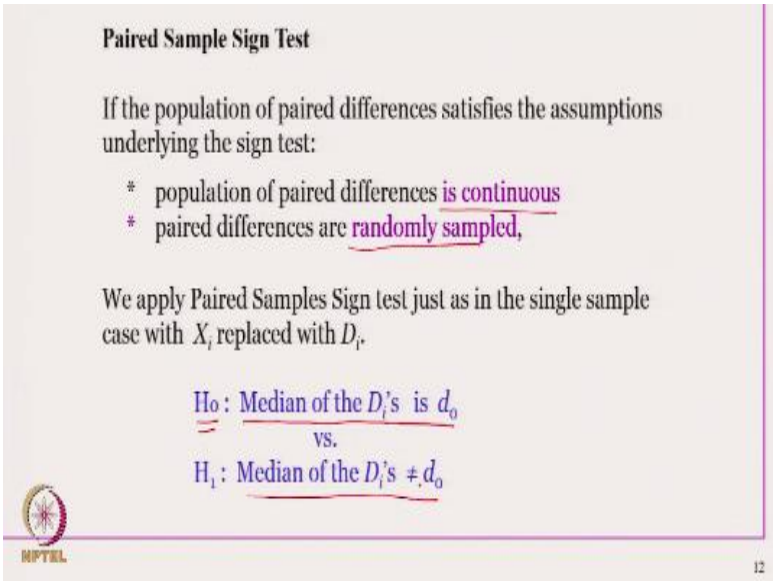


So consider a random sample of pairs $(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)$. For this sample n differences can be calculated $D_i = X_i - Y_i$ that means what it was before and it was after. So that gives you the difference and then the test statistic is based on the median of the difference

of the paired observations. So I hope the concept is clear. Earlier we were looking at sign test or signed rank test for either X or Y.

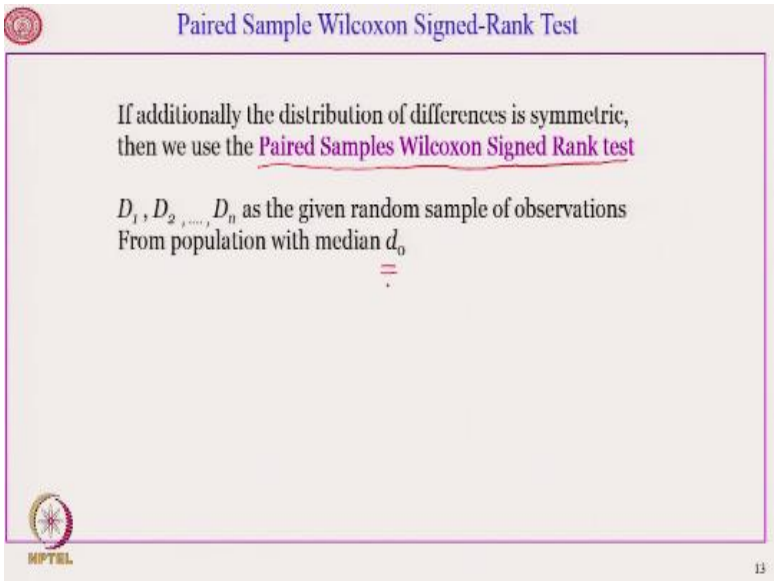
Now since we have paired values we look at the difference and we are going to apply the same technique on the $D_i's$.

(Refer Slide Time: 06:27)



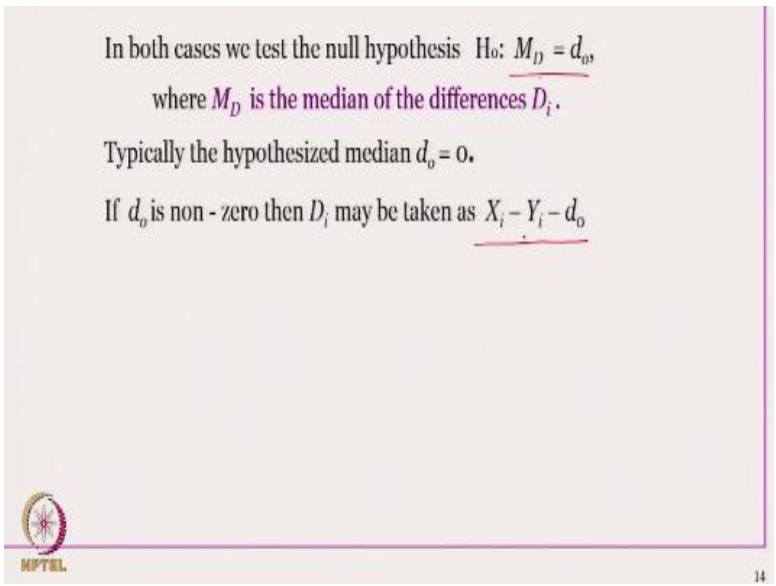
So what we will do in a paired sample Sign test. If the population of paired differences satisfies the assumptions of underlying sign test, right?. What is that? The distribution is continuous and the sample is random. If these two assumptions are satisfied then we are going to use paired sample Sign test where the H_0 or the null hypothesis is that median of the $D_i's$ is d_0, versus median of the $D_i's$ is not equal to d_0. So for example we are looking at only two-sided test.

(Refer Slide Time: 07:09)



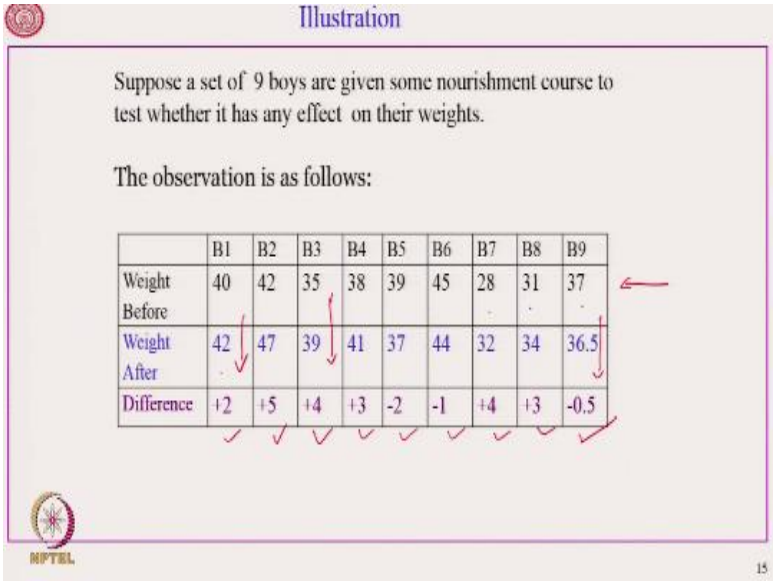
If in addition to that we can also assume that the differences are symmetric than we can use paired sample, Wilcoxon Signed Rank test and then $D_1, D_2, \cdots, D_n$, are the given samples. We may test if they are coming from a population with median is equal to 0 or they are coming from a population with a median is equal to d_0. In the later case we have to further modify the data as follows.

(Refer Slide Time: 07:46)



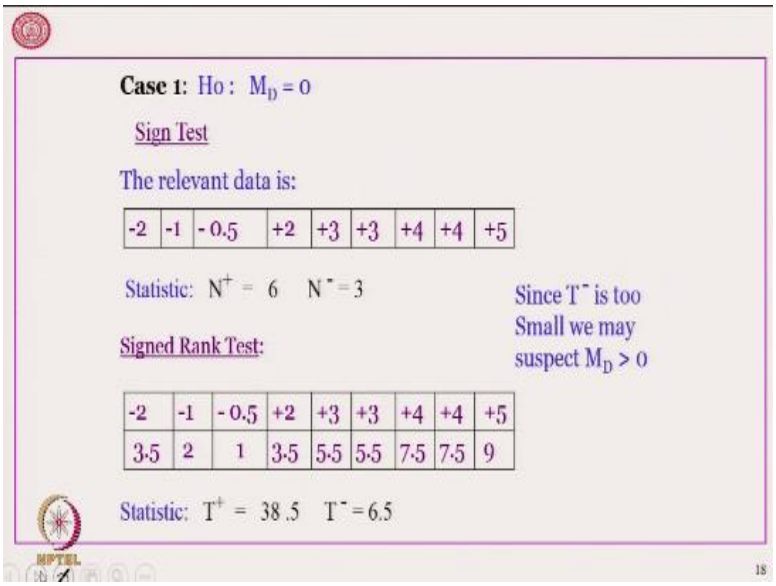
When the H_0 is that median is equal to 0 then there is no problem because that is typically the case, but if $d_0 \neq 0$, then we shall look at $X_i - Y_i - d_0$ and then we will check if there is any, so that this now becomes zero centered and then we can test the null hypothesis.

(Refer Slide Time: 08:14)



So, suppose there are 9 boys are given some nourishment and suppose their weights before the course are these 40, 42, 35, 38, 39, 45, 28, 31 and 37. Now after the course is complete their weights are measured again and we found this is the weight after the course is done. The boy with 40 has now become 42 with weight, the boy with 35 now has a weight gain 39, and similarly the one with 37 now has a weight 36.5.

So we are looking at $Y_i - X_i$ that is the gain after the nourishment course. So we find that the difference is $+2, +5, +4, +3, -2, -1, +4, +3$ and -0.5. (Refer Slide Time: 09:22)



Therefore, our focus of attention is now this vector when we put them in sorted order of magnitude $-2, -1, -0.5$ these are the negative gains and $+2, +3, +3, +4$ etcetera these are the positive gains. Now $H_0: M_D = 0$ that is whether there is no gain or in other words whether

the median is equal to 0 for the difference. Therefore the corresponding statistic is going to be N^+ and N^-.

We see that there are 6 of them which are positive and 3 of them, they are on negative side. Therefore, we know that we have to look at binomial $Bin(9,0.5)$ and we have to see the cumulative probability of N^- and from there we should be able to decide whether to accept this or reject. When we are going for signed rank test then in the sorted order this is the values that we have.

We are now instead of just counting we are giving ranks to them. The smallest one in magnitude is 0.5 therefore it gets the rank 1, the second smallest is magnitude wise 1 therefore it is getting the value 2. Third has, this and this, both of them absolute value 2 so they should get average of 3 and 4 that is why they are getting 3.5 and 3.5. Similarly these two are getting average of 5 and 6 these two are getting average of 7 and 8 and this one is getting the value 9.

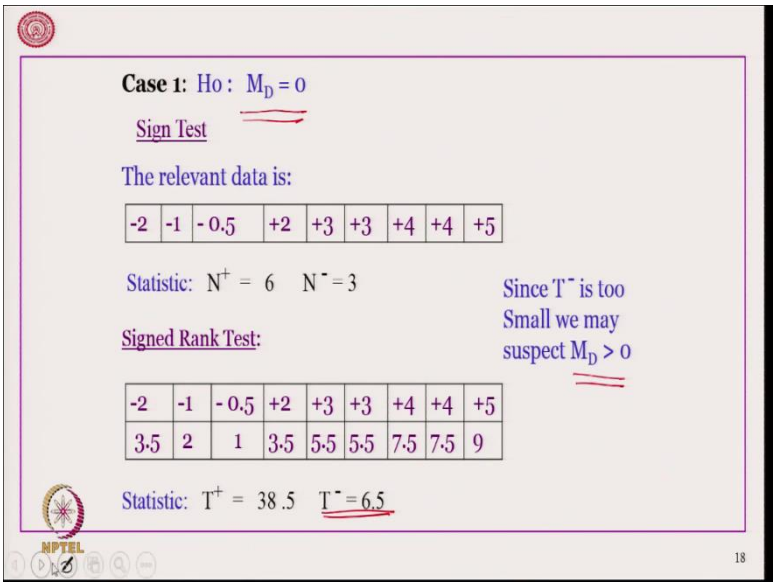
Therefore T^+ is going to be 38.5 and T^- is going to be 6.5. As you can see from here it is 6.5 and since the sum is going to be 45 we can easily get that this is going to be 38.5. Therefore since T^- is too small compared to T^+ we suspect that the median of D can be greater than 0 therefore we need to test that.

(Refer Slide Time: 12:01)

n	Two-Tailed Test		One-Tailed Test	
	α = .05	α = .01	α = .05	α = .01
5	--	--	0	--
6	0	--	2	--
7	2	--	3	0
8	3	0	5	1
9	5	1	8	3
10	8	3	10	5
11	10	5	13	7
12	13	7	17	9
13	17	9	21	12
14	21	12	25	15
15	25	15	30	19
16	29	19	35	23
17	34	23	41	27
18	40	27	47	32
19	46	32	53	37
20	52	37	60	43
21	58	42	67	49

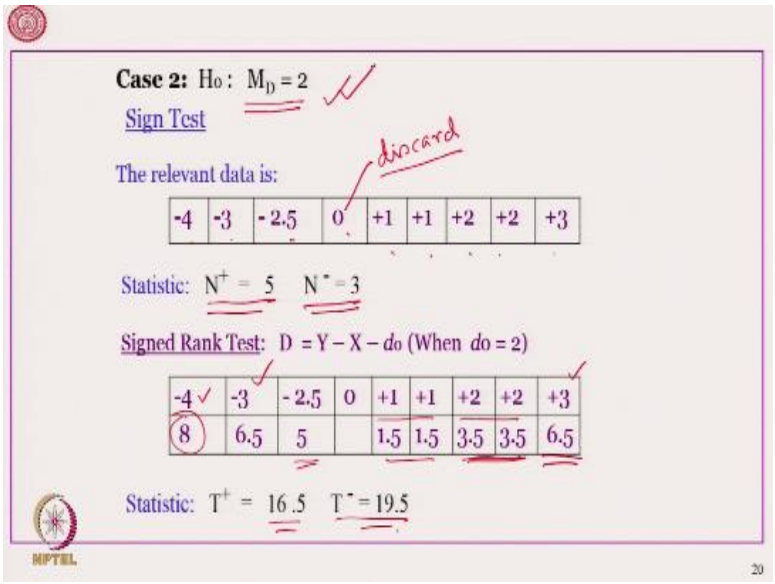
So we go back to our Wilcoxon signed rank test tablet. We have the value 9 and at 5% level we have the critical value to be 8. And since our obtained value is 6.5 which is less than 8 therefore we reject the null hypothesis.

(Refer Slide Time: 12:29)



Case 2, that means that we are not accepting that $M_D = 0$ that means that we are rejecting $M_D = 0$ in favor of $M_D > 0$.

(Refer Slide Time: 12:44)



Now let us look at the second case when we are looking at $M_D = 2$. Therefore now my statistic is going to be we are going to subtract 2 from all of them therefore we are getting $-4, -3, -2.5, 0$ and then 1, 1, 2, 2 and 3.

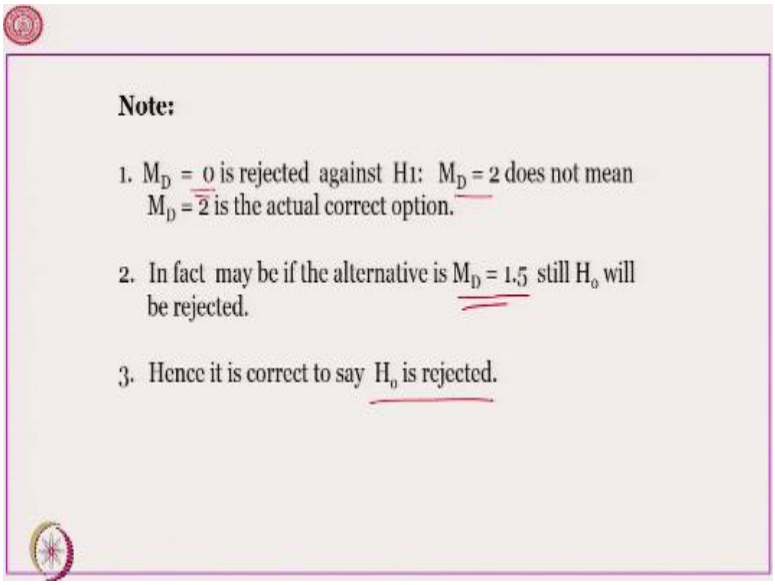
Therefore, when we are applying Sign test we are getting $N^+ = 5$ and $N^- = 3$ because we are discarding this as it is same as the value 0. And since it is zero centered therefore we discard

that value 0. Now we can see that N^+ and N^- are very close therefore Sign test cannot reject $M_D = 2$.

On the contrary suppose we are using signed rank test. Then in this case again we assign the ranks to the deviations and we get these two are getting 1.5 because average of 1 and 2, these two are getting 3.5 average of 3 and 4, this is getting 5 this and this together are getting 6 and 7 average that is 6.5 and this is the maximum one in absolute value therefore getting the rank 8. Therefore if we consider T^+it is $1.5 + 1.5 + 3.5 + 3.5 + 6.5 = 16.5$ thus, $T^+ = 16.5$ and $T^- = 19.5$

Again I am not going into the table to check whether we can accept or reject the null hypothesis, but we can see that T^+ and T^- are very close to each other we can say from here, we cannot reject the null hypothesis $M_D = 2$.

(Refer Slide Time: 14:52)

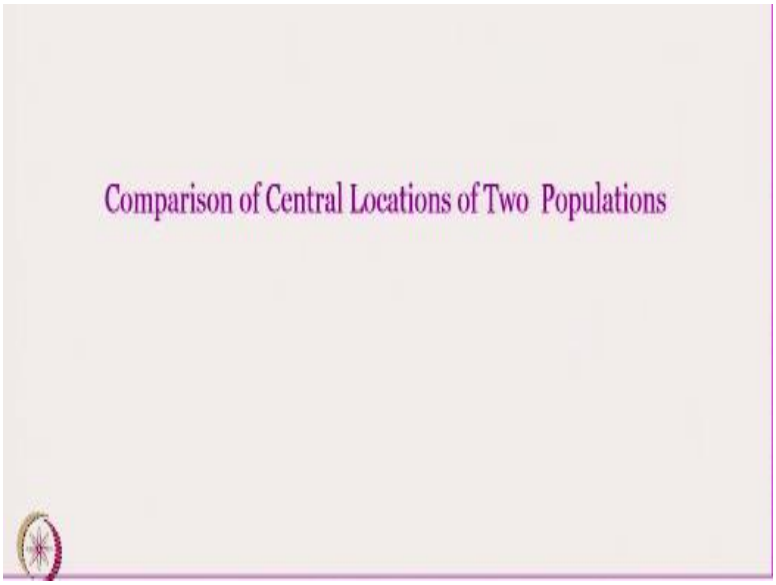


Note that $M_D = 0$ is rejected against $M_D = 2$ does not mean that the median of deviation is actually 2. It is only that, against $M_D = 2$ we are going to reject $M_D = 0$. In fact instead of 2 if we are testing against say $M_D = 1.5$ or say $M_D = 2.25$ we may find that we are still rejecting $M_D = 0$ against the corresponding alternative.

Therefore, we are rejecting a null hypothesis does not mean that the alternative is true. So, instead of saying that accepting the alternative we should say that H_0 is rejected. So this is generally true for testing of hypothesis even for parametric cases it does not mean that rejection

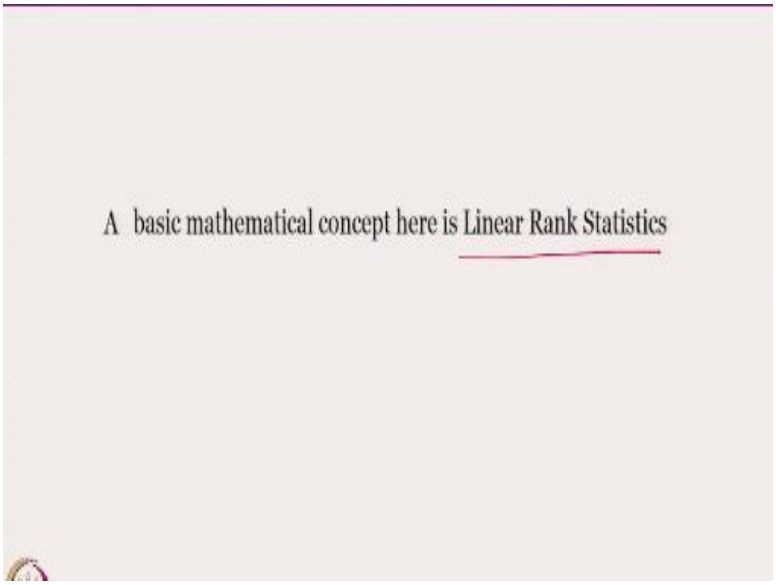
of null hypothesis in against some particular alternative actually means the value of the alternative is the correct one.

(Refer Slide Time: 16:08)



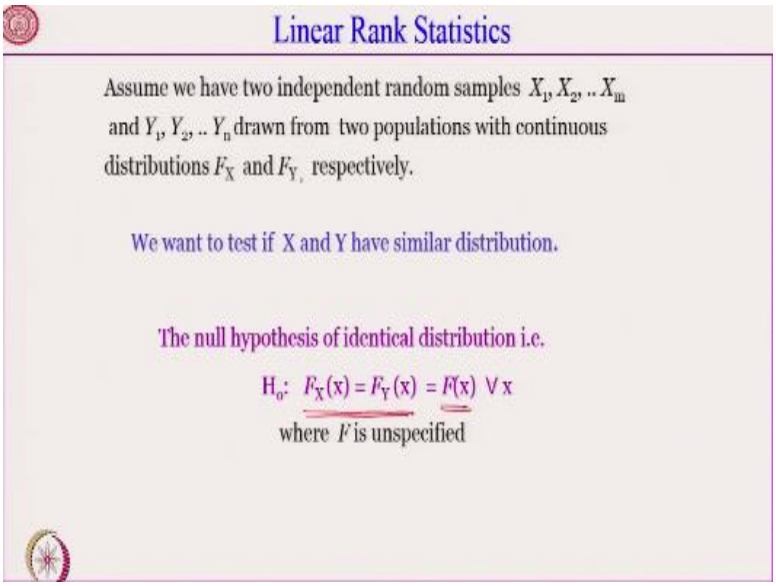
So this is about paired sample test now we are going little bit ahead we want to test the centrality of two different populations. How to do that?

(Refer Slide Time: 16:22)



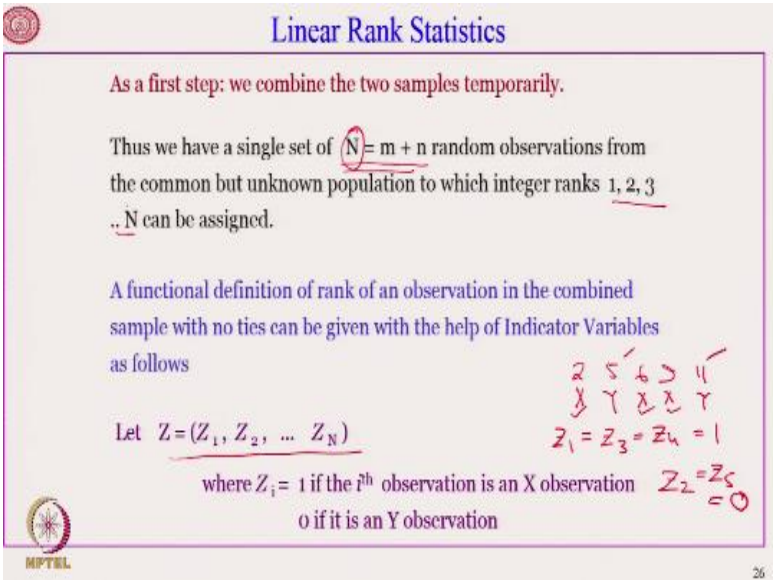
The basic mathematical concept here is that of Linear Rank Statistics. I shall introduce you to linear rank statistics, but the actual mathematical property of linear rank statistic we shall study in lecture 5..

(Refer Slide Time: 16:39)



So what is linear rank statistics? Suppose we have two independent random samples $X_1, X_2, \cdots, X_m$ from X population and $Y_1, Y_2, \cdots, Y_n$ drawn from another population. They have continuous distribution F_X and F_Y respectively. We want to test if X and Y have same or similar distribution. Thus, when we are comparing the distribution function of two different populations we are trying to check if $F_X(x) = F_Y(x)$ is equal to some common distribution function $F(x)$ which is not specified to us. So in effect we are checking if these two samples are coming from the same population.

(Refer Slide Time: 17:34)



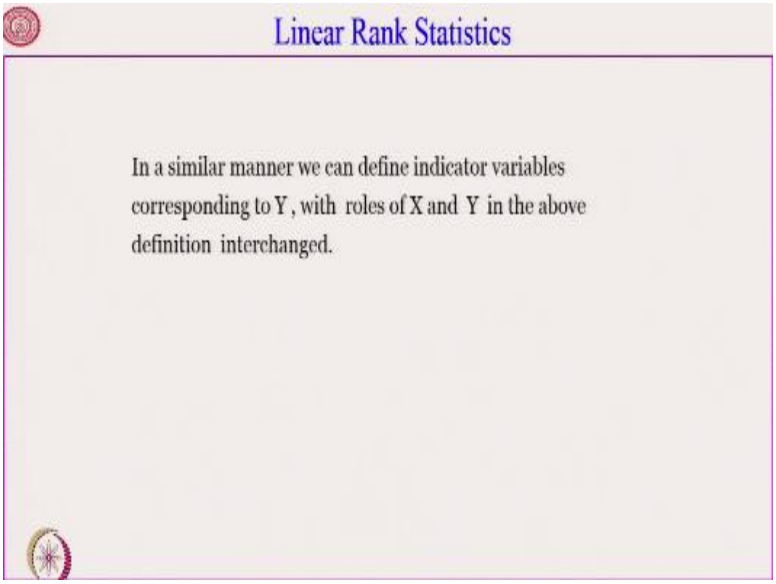
So how we proceed? We proceed as follows. As a first step we combine the two samples temporarily that means we make one array made of both X and Y populations in that array therefore there will be $N = m + n$ many observations and we can assign them ranks 1, 2, 3 up

to N if there is no ties. In case of ties of course we can take appropriate average to give the ranks. So that the total sum of rank remain same which is equal to $1 + 2 + \cdots + N$.

A functional definition of rank of an observation in the combined sample with no ties can be given with the help of an indicator variable. So let $Z = (Z_1, Z_2, \cdots, Z_N)$ is a vector of indicator variables where $Z_i = 1$ if the i^{th} observation is an X observation and $Z_i = 0$ if it is an Y observation.

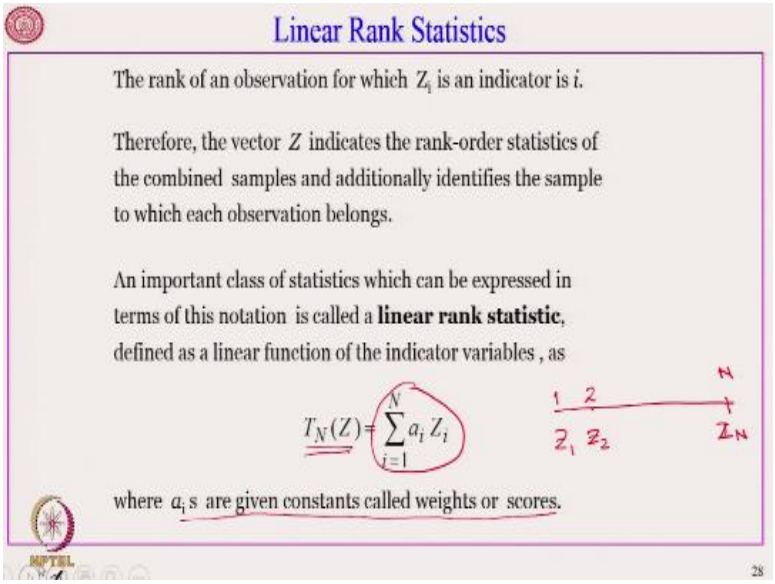
Say for example suppose my observations are 2, 5, 6, 9 and 11 where these are X and these two are Y. Therefore, $Z_1 = Z_3 = Z_4 = 1$ because first, third and fourth observations are from X therefore they are getting the value 1 and $Z_2 = Z_5 = 0$ because these two are from Y population. So, that is how we get an array of size capital N with binary values either 1 or 0 depending upon whether it is from X or from Y.

(Refer Slide Time: 19:53)



Of course, one can do it in the reverse way one can give value 1 for Y observations and the value 0 for X observations, but in that case also the treatment is very similar.

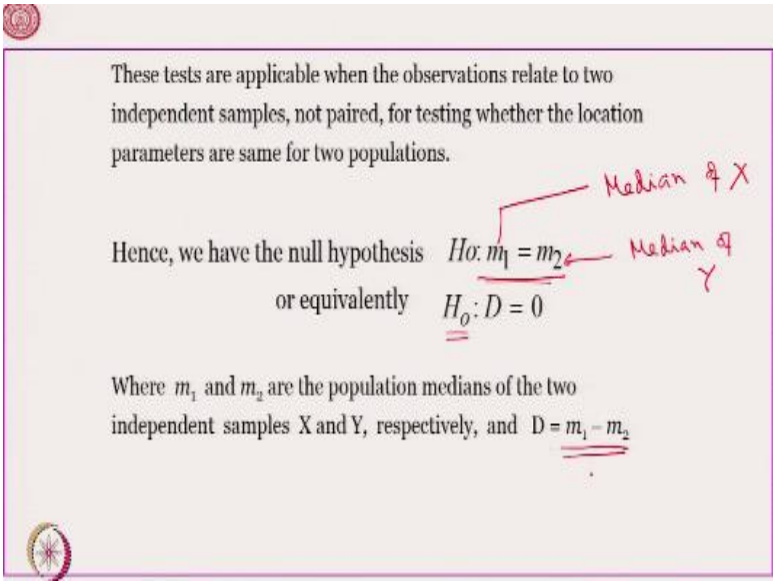
(Refer Slide Time: 20:07)



Now, the rank of an observation for which Z_i is an indicator is i. That means that for Z_1 the rank is going to be 1, the one that is corresponding to Z_2 it will have rank 2 and the one that is corresponding to Z_N that will have rank N in the sorted array. A linear combination of the Z values $T_N(Z) = \sum_{i=1}^{N} a_i Z_i$, is called a linear rank statistic where $a_i's$ are given constants called weights or scores, depending upon the problem we can change the value of ai to understand certain properties.

As I have said earlier that we will look at some such mathematical properties when we shall be in our lecture 5. For the time being we are not going into the mathematical details rather we shall see how we can compare the distribution of two population or two population medians.

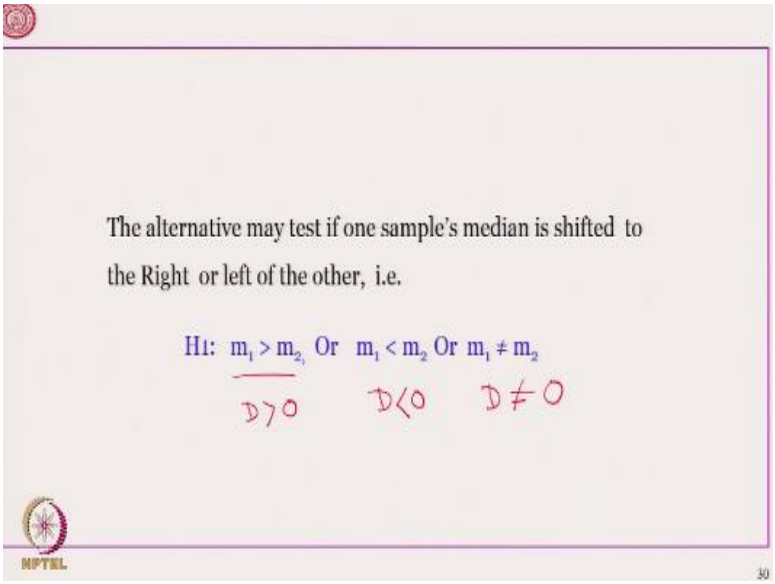
(Refer Slide Time: 21:28)



So what we are testing? We are testing the null hypothesis that $m_1 = m_2$ that means that if m_1 is the median of X population and m_2 is the median of Y population, we are checking whether these two populations have the same median. But notice that having the same median does not mean that these two populations have the same distribution we shall also have to look at the dispersion of these two populations which we shall look at in lecture 4, which is called a scale problem.

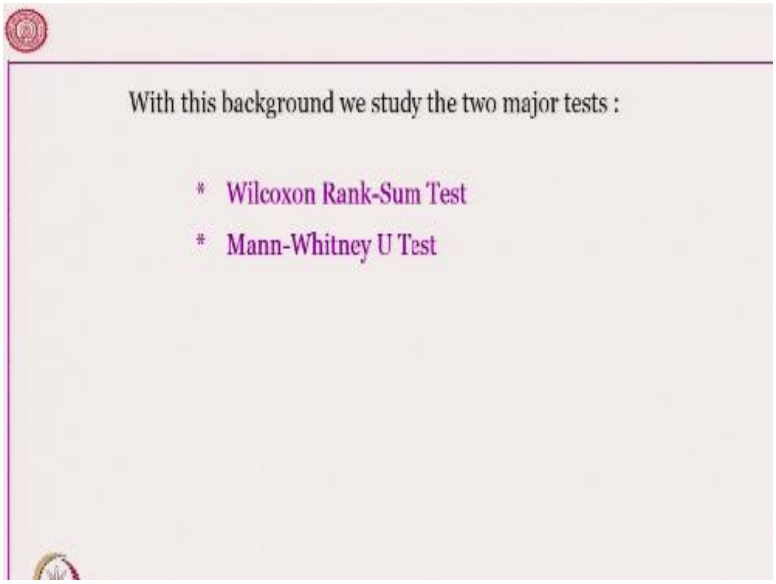
For this lecture we are just looking at the commonality of the central locations and our statistic is therefore appropriately defined, but our null hypothesis is that $H_0: m_1 = m_2$ or $H_0: D = 0$ that means the difference of the two medians m_1 and m_2 is equal to 0.

(Refer Slide Time: 22:47)



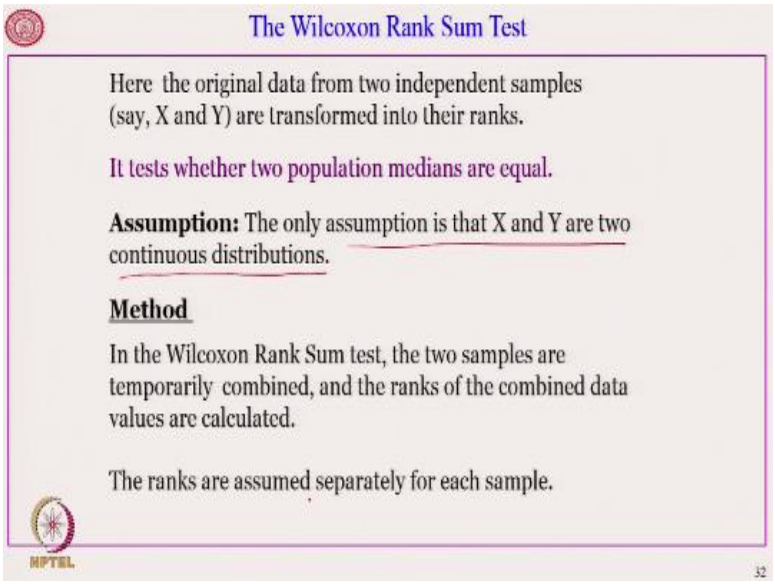
What can be the alternative? The alternative can be as before it can be $m_1 > m_2$ or $D > 0$, $m_1 < m_2$ or $D < 0$, $m_1 \neq m_2$ or $D \neq 0$.

(Refer Slide Time: 23:11)



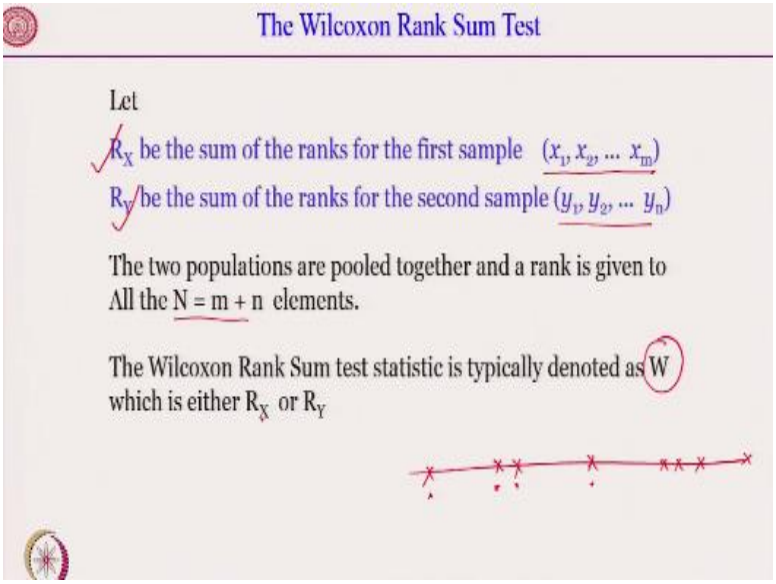
With this background we study two major tests which are called Wilcoxon Rank-Sum Test and Mann-Whitney U test.

(Refer Slide Time: 23:24)



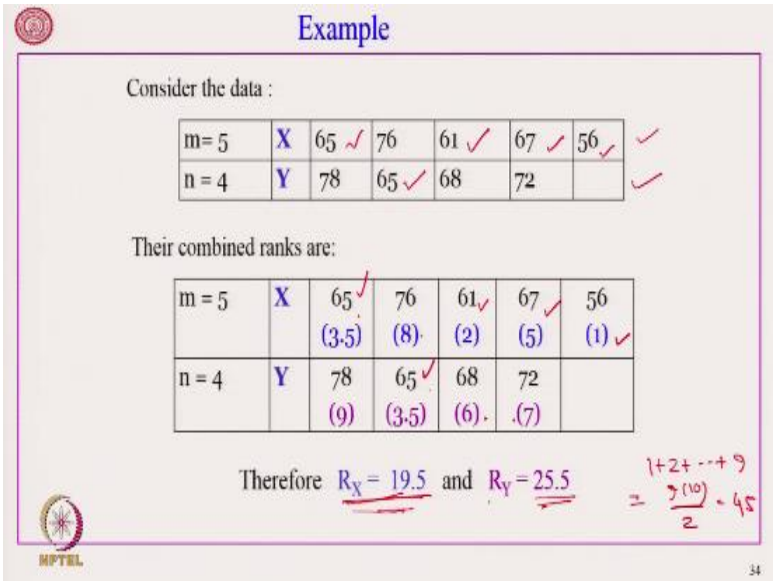
In Wilcoxon Rank Sum Test, the original data X and Y are from two independent samples we want to test whether these two populations have the same median. The only assumptions that is made is that X and Y have continuous distributions this is very important. What is the method? In Wilcoxon Rank Sum Test the two samples are temporarily combined and the ranks of the combined data values are calculated and then we look at the ranks for X and Y separately.

(Refer Slide Time: 24:06)



So let R_X be the sum of ranks for the X samples that means we have m observations $x_1, x_2, \cdots, x_m$. In the pooled array they need not be consecutive. So maybe these are the X observations therefore their ranks are not in continuity 1, 2, 3 like that. Therefore we just consider the ranks of these observations and sum them up that is going to give you R_X. In a similar way the observations coming from the second sample their sum is called R_Y together they have capital N many elements and the rank sum statistic is going to be called W which can be R_X or R_Y.

(Refer Slide Time: 25:04)

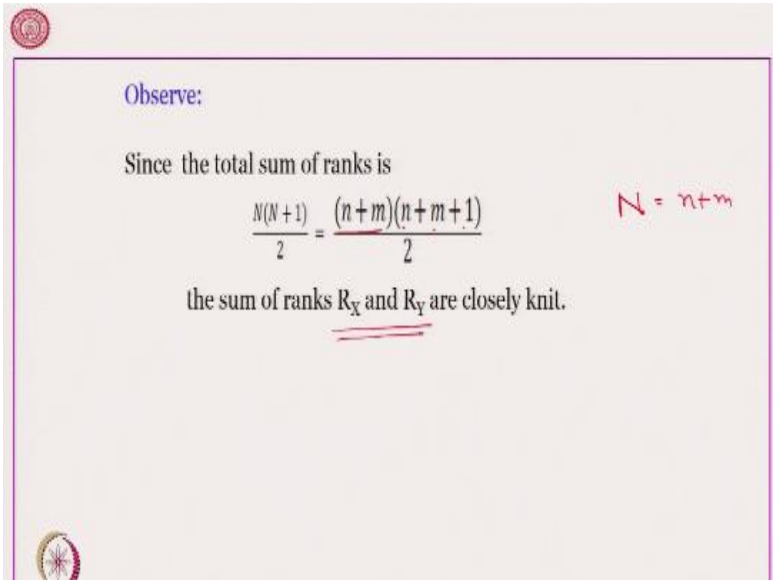


Let me illustrate with an example. Suppose there are 5 X observations and 4 Y observations. Now we sort them together and give them the ranks, the smallest one of them is 56 so it is getting the rank 1. The second smallest is 61 it is getting the rank 2, the third smallest is 65, but

there is another 65 among the Y so both of them are getting the rank 3.5, the fifth one because they are taking care of 3 and 4, the fifth one is 67 so that gets the rank 5 and in a similar way we give ranks 6, 7, 8 and 9.

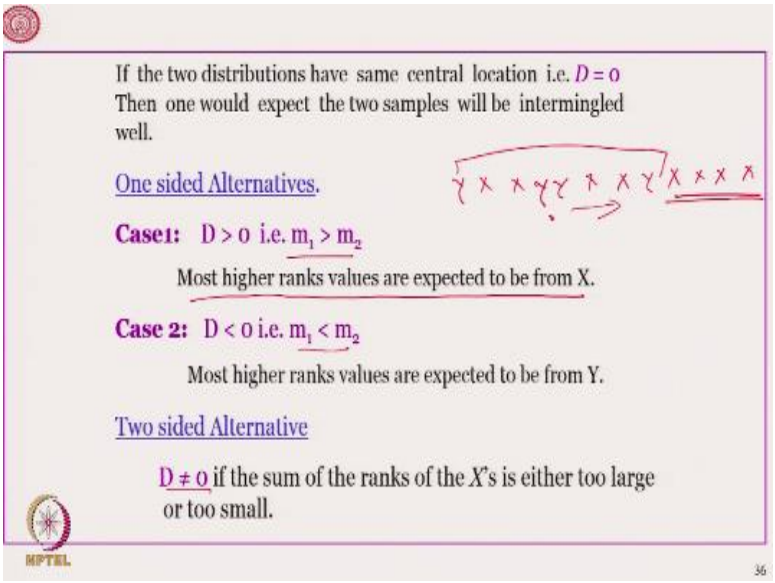
Therefore R_X which are the blue ones is the sum of these ranks which is $5 + 1 + 2 + 8 + 3.5 = 19.5$, thus, $R_X = 19.5$ and R_Y is therefore going to be $7 + 6 + 9 + 3.5 = 25.5$, thus $R_Y = 25.5$. Therefore from this given observation we calculate R_X and R_Y any one of them can be used as a statistic.

(Refer Slide Time: 26:26)



But before that let us observe that the total sum of rank is going to be $\frac{N(N+1)}{2} = \frac{(n+m)(n+m+1)}{2}$this is because $N = n + m$. Therefore R_X and R_Y are very closely related because their sum is equal to constant. If you look at the previous one this sum is equal to 45 and because there are 9 elements the rank is going to be $1 + 2 + \cdots + 9 = \frac{9*10}{2} = 45$. Therefore their sum of the R_X and R_Y is constant for a given sample size.

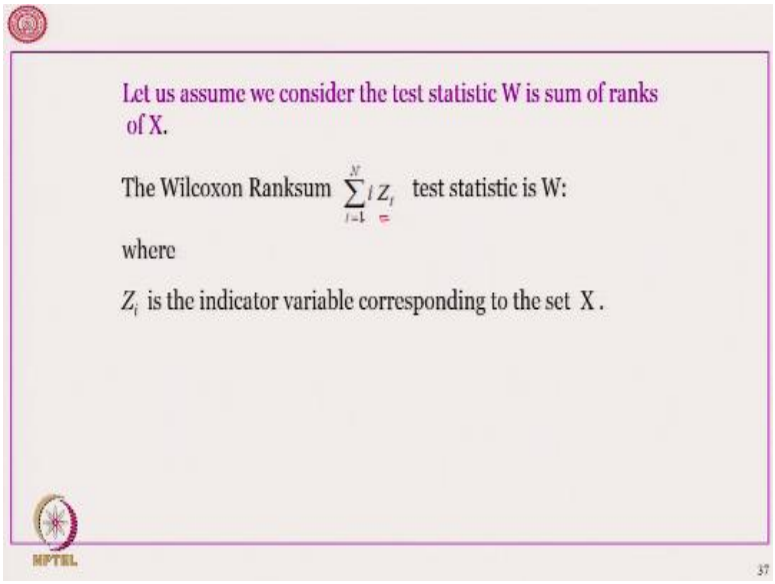
(Refer Slide Time: 27:27)



If the two distributions have the same central location that is $D = 0$ then one would expect that the two samples will be intermingled well. Now what are going to be the alternatives, $D > 0$ that is $m_1 > m_2$ as we have already discussed, $D < 0$ that is $m_1 < m_2$ and $D \neq 0$ that is $m_1 \neq m_2$. Now suppose $m_1 > m_2$. When do we think that? Most higher rank values are expected to be from X.

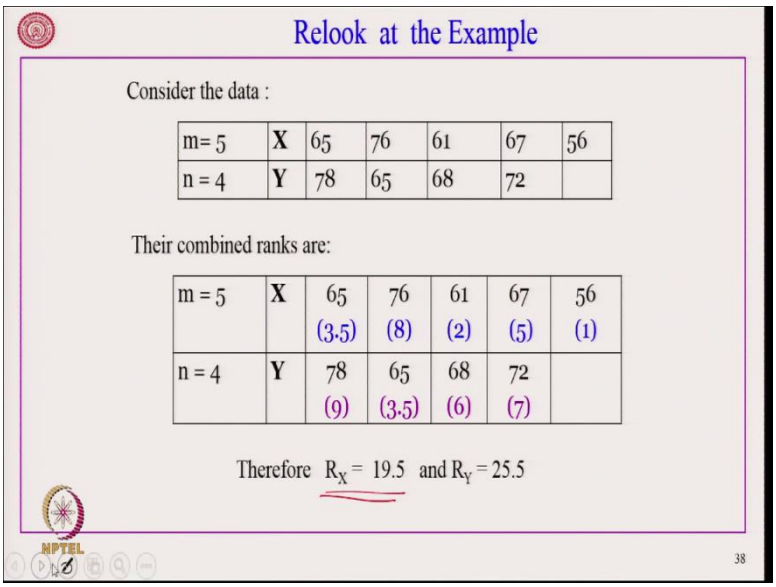
So suppose this is the combined population and m_1 is bigger than m_2 then we would expect that the higher values in the combined population are coming from X. So that we could expect the median is on this side, but if these are the only values that is coming from Y we would expect the median is going to be somewhere here. So that is the intuition that works behind choosing the criteria.

(Refer Slide Time: 28:44)



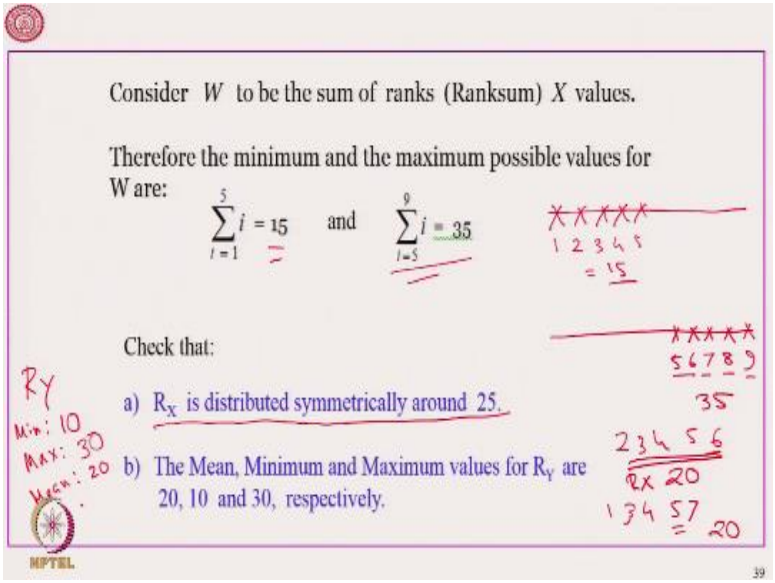
So let us consider the sum of ranks for X that is R_X. Therefore the statistic $W = \sum_{i=1}^{N} iZ_i$, because we have defined Z_i to be 1 when the observation is coming from X and therefore, $\sum_{i=1}^{N} iZ_i$ is going to give us the Wilcoxon rank sum statistic R_X.

(Refer Slide Time: 29:10)



This we have already calculated. So corresponding to this population our statistic is going to be $R_X = 19.5$.

(Refer Slide Time: 29:22)



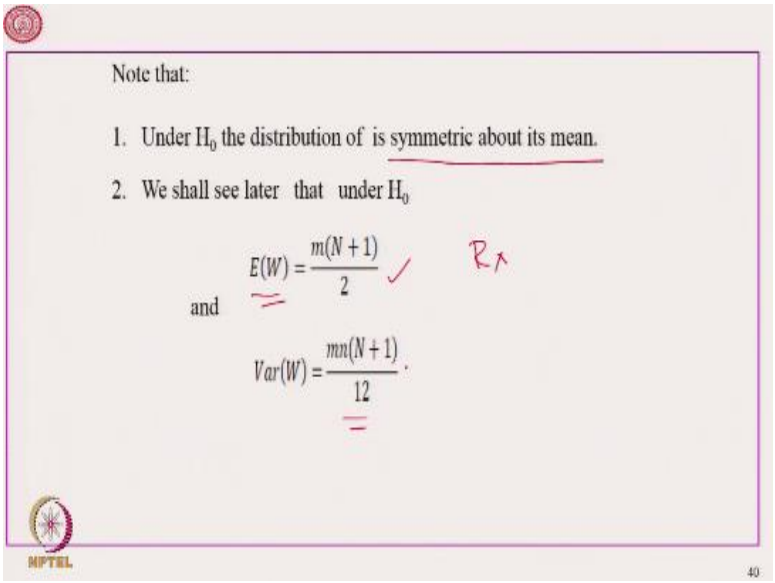
Now let us observe a few interesting things that when there are 9 observations 5 of them are from X and 4 of them are from Y then the minimum value of R_X is equal to 15 because if there are 9 observations the minimum value that sum of ranks will get is, the first 5 of them are from X and therefore their ranks are going to be $1 + 2 + 3 + 4 + 5 = 15$.

What is the maximum possible value? The maximum possible value is going to be when the highest 5 values are taken by X therefore these are going to be 5, 6, 7, 8 and 9 and therefore their sum is $5 + 6 + 7 + 8 + 9 = 35$. Now this is a distribution because R_X can take the value 15 only in this way, but R_X can take the value 20 in different ways because suppose our X takes the value second, fourth, fifth, sixth and third.

So that will give us the value of $R_X = 20$, but we can also get it in the form say 1, 3, 4, 5, 7 that will also give us the value 20. Therefore as we look at the value of the statistic R_X we shall find that it can take the values between 15 to 35 with different probabilities. I want you to check that R_X is distributed symmetrically around the value 25.

In a similar way one can check that if we consider R_Y, then minimum value is equal to 10 when Y observations are the first, second, third and fourth in the combined sorted array the maximum value is going to be 30 because when Y values are 6, 7, 8 and 9 then the sum of ranks is going to be 30 and you can check that the mean is equal to 20 and it is symmetric around that value. So this I leave it as an exercise.

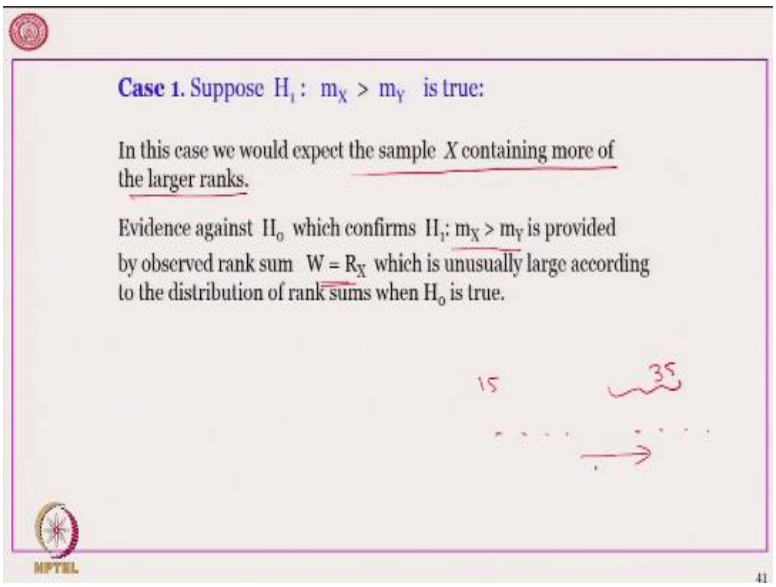
(Refer Slide Time: 32:14)



Let us go forward under H_0 the distribution is symmetric about the mean we shall see later that expected value of W is equal to $\frac{m(N+1)}{2}$ if we consider R_X to be the statistic and variance of W is going to be $\frac{mn(N+1)}{12}$. As I said we shall prove in lecture 5.

.

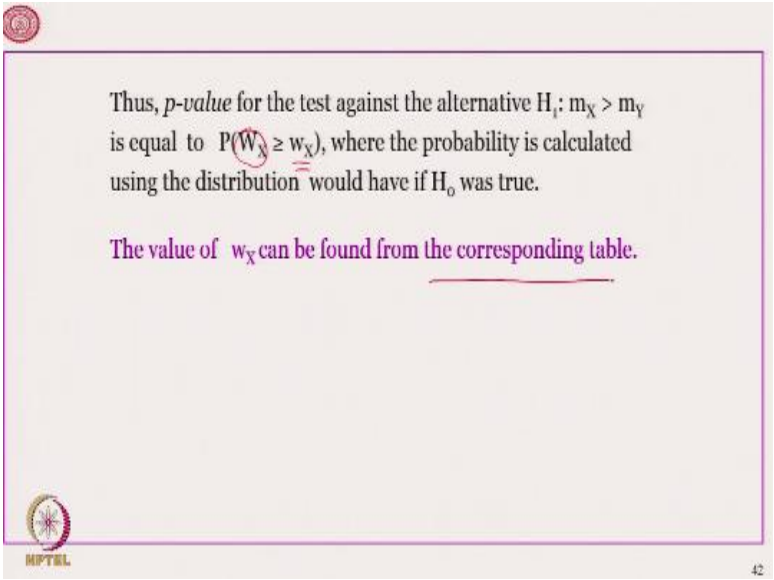
(Refer Slide Time: 32:40)



Now suppose $H_1: m_X > m_Y$ which is true. In that case what is going to happen you would expect that this sample X contains more of the larger ranks. Therefore evidence against H_0 which confirms that $m_X > m_Y$ is provided by the rank sum W is equal to R_X, which is unusually large according to the distribution of rank sums when H_0 is true. As I said with respect to X the smallest value is 15 but the highest value is 35.

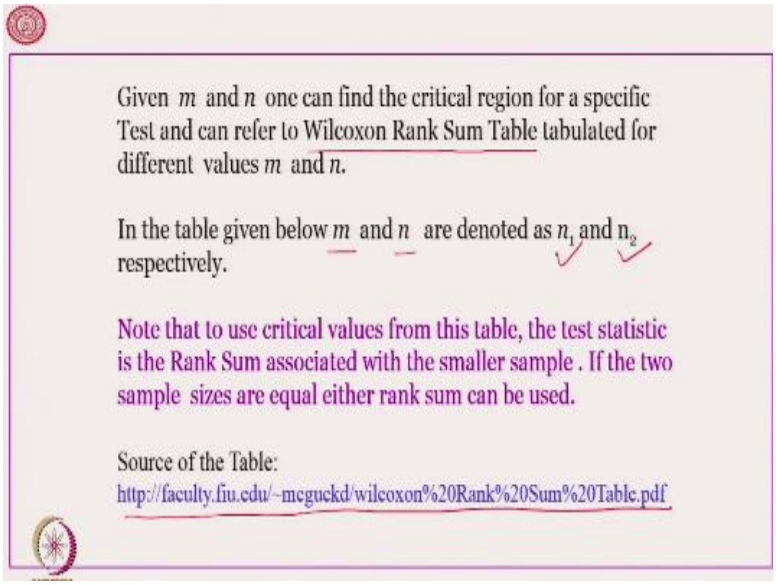
So, if the observation we find that the sum of ranks of X is coming out to be very close to 35 that means that all the highest rank things are coming from X, lower rank elements are coming from Y then we would expect that the median of X is actually bigger than the median of Y.

(Refer Slide Time: 33:50)



Therefore, in that case we shall reject the null hypothesis if the statistic obtained from the population is greater than equal to some critical value and where from we get the critical value? We get the critical value from some table.

(Refer Slide Time: 34:10)

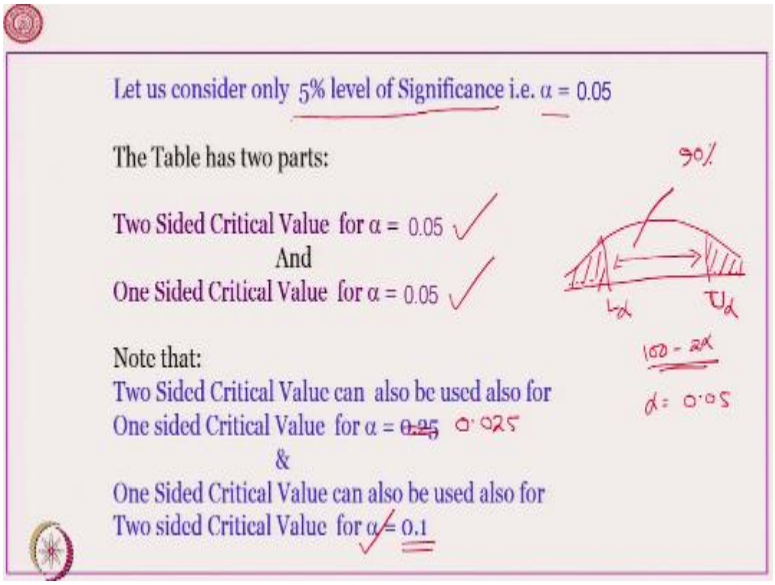


Given m and n one can find the critical region for a specific test and can refer to Wilcoxon Rank Sum Table tabulated for different values of m and n. I shall give you a sample of that table. Note that in that table m and n are denoted by n_1 and n_2 respectively. So these are the

sample sizes n_1 and n_2. Note that to use critical values from this table the test statistic is the rank sum associated with the smaller sample.

If two sample sizes are equal than either rank sum can be used. So, the table that I am showing to you has been taken from this source.

(Refer Slide Time: 35:08)

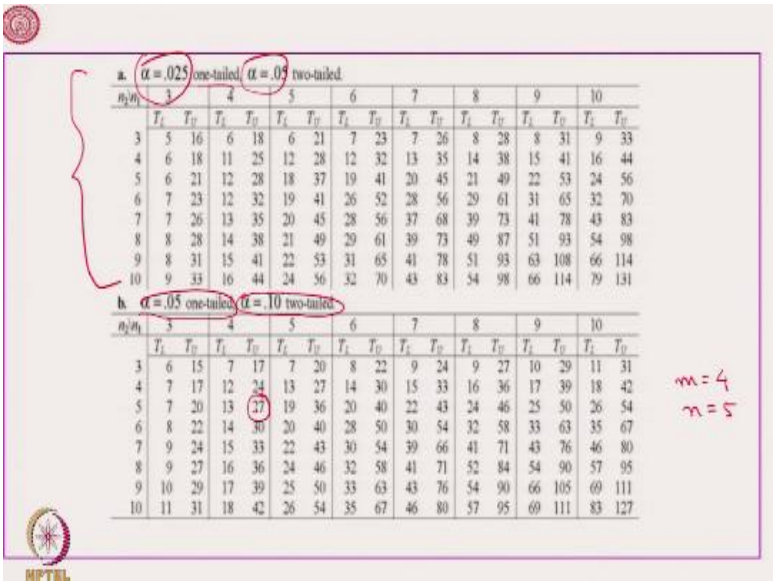


So before we see the table let me just tell you something. So consider that you are testing at 5 percent level of significance that means $\alpha = 0.05$. Now this can be one-sided, this can be two-sided. Hence, accordingly the table has two parts. Two-sided critical value for alpha is equal to 0.05 and one-sided critical value for alpha is equal to 0.05. Note that the two-sided critical value can also be used for one-sided critical value for alpha is equal to 0.025.

And one-sided critical value can also be used for two-sided critical value for alpha is equal to 0.1. So basically what I am saying suppose this is the distribution of the statistic it is the one-sided critical value for alpha. So this is the upper side, let me call it U_α and this is the lower side let me call it L_α. Therefore, we are rejecting it for a one-sided test if the statistic value is on this side or the statistic value is on this side that is for one-sided test.

Therefore, if I am going to test it for two-sided then with equality of tails, if I consider this and this then the total length or the confidence of this interval is equal to $100 - 2\alpha$ that is if $\alpha = 0.05$ then this is going to be 90 percent that is why we are saying that one-sided critical value can also be used for two-sided critical value for alpha is equal to 0.1.

(Refer Slide Time: 37:32)



a. $\alpha = .025$ one-tailed, $\alpha = .05$ two-tailed.

n_2/n_1	3		4		5		6		7		8		9		10	
	T_L	T_U	T_L	T_U	T_L	T_U	T_L	T_U	T_L	T_U	T_L	T_U	T_L	T_U	T_L	T_U
3	5	16	6	18	6	21	7	23	7	26	8	28	8	31	9	33
4	6	18	11	25	12	28	12	32	13	35	14	38	15	41	16	44
5	6	21	12	28	18	37	19	41	20	45	21	49	22	53	24	56
6	7	23	12	32	19	41	26	52	28	56	29	61	31	65	32	70
7	7	26	13	35	20	45	28	56	37	68	39	73	41	78	43	83
8	8	28	14	38	21	49	29	61	39	73	49	87	51	93	54	98
9	8	31	15	41	22	53	31	65	41	78	51	93	63	108	66	114
10	9	33	16	44	24	56	32	70	43	83	54	98	66	114	79	131

b. $\alpha = .05$ one-tailed, $\alpha = .10$ two-tailed.

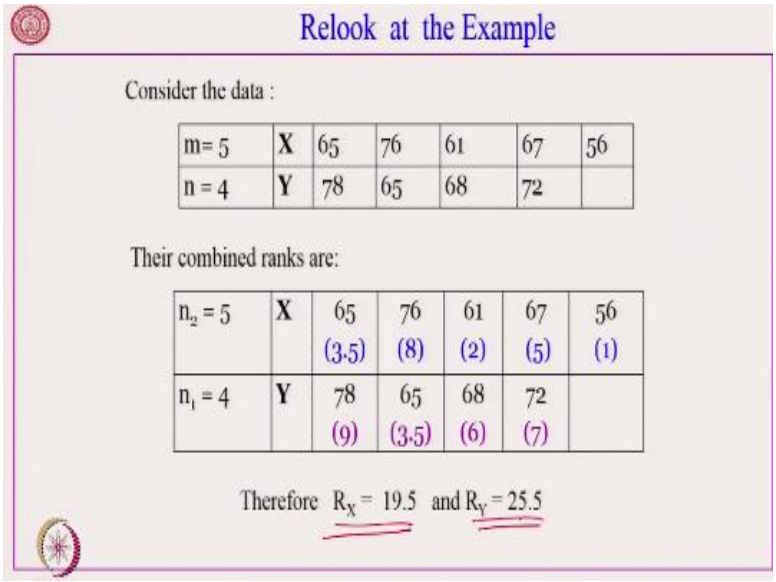
n_2/n_1	3		4		5		6		7		8		9		10	
	T_L	T_U	T_L	T_U	T_L	T_U	T_L	T_U	T_L	T_U	T_L	T_U	T_L	T_U	T_L	T_U
3	6	15	7	17	7	20	8	22	9	24	9	27	10	29	11	31
4	7	17	12	24	13	27	14	30	15	33	16	36	17	39	18	42
5	7	20	13	27	19	36	20	40	22	43	24	46	25	50	26	54
6	8	22	14	30	20	40	28	50	30	54	32	58	33	63	35	67
7	9	24	15	33	22	43	30	54	39	66	41	71	43	76	46	80
8	9	27	16	36	24	46	32	58	41	71	52	84	54	90	57	95
9	10	29	17	39	25	50	33	63	43	76	54	90	66	105	69	111
10	11	31	18	42	26	54	35	67	46	80	57	95	69	111	83	127

$m = 4$
$n = 5$

So let us now have a look at the table it is given that the upper part this part is for alpha is equal to 0.25 for one-tailed or alpha is equal to 0.5 for two-tailed. Similarly this side alpha is equal to 0.05 for one-tailed or alpha is 0.1 for two-tailed. Therefore if I am testing for $m = 4$ and $n = 5$ then we are looking at 5% level this value which is 27.

Therefore on the upper side if the obtained value is bigger than 27, then we are going to reject that null hypothesis otherwise we are going to accept the null hypothesis.

(Refer Slide Time: 38:36)



Relook at the Example

Consider the data :

m = 5	X	65	76	61	67	56
n = 4	Y	78	65	68	72	

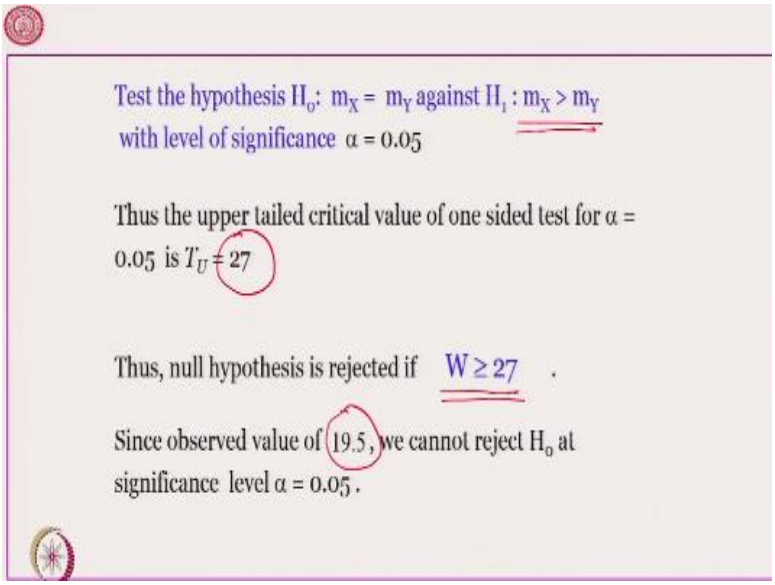
Their combined ranks are:

$n_2 = 5$	X	65 (3.5)	76 (8)	61 (2)	67 (5)	56 (1)
$n_1 = 4$	Y	78 (9)	65 (3.5)	68 (6)	72 (7)	

Therefore $R_X = 19.5$ and $R_Y = 25.5$

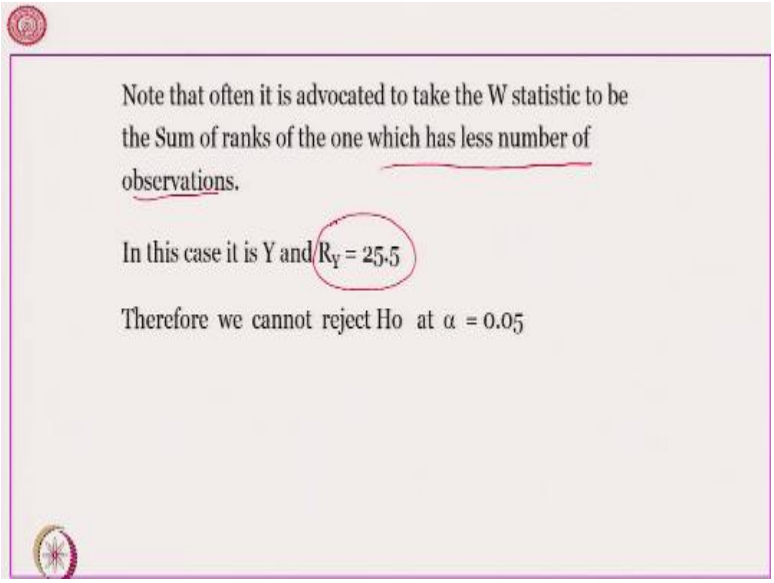
So let us look at this $m = 5$, $n = 4$ we have already computed this and we have found that $R_X = 19.5$ and $R_Y = 25.5$.

(Refer Slide Time: 38:52)



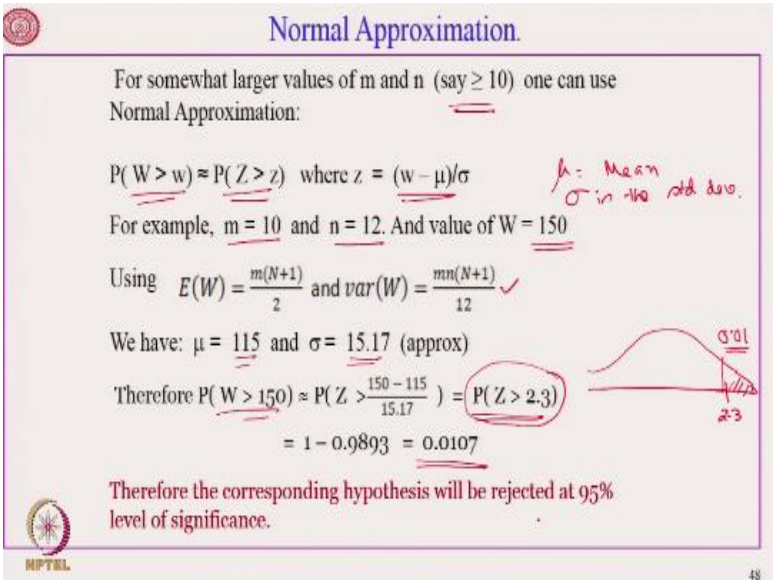
So at 5 percent level of significance when the alternative is that the median of X is greater than median of Y then the critical value that we obtained from the table just now we have discussed is 27. Therefore the null hypothesis is going to be rejected if the obtained value of the statistic W is greater than equal to 27. However we have obtained the value is 19.5, therefore we cannot reject the null hypothesis at this significance level α is equal to 0.05.

(Refer Slide Time: 39:36)



Sometimes it is advocated that you take the W statistic to be the sum of ranks of the one, out of X and Y which has less number of observations. In our case it was Y because Y has only four 4 observations whereas X has 5 and we have seen that $R_Y = 25.5$. Therefore using Y also we cannot reject the null hypothesis at 0.05.

(Refer Slide Time: 40:10)

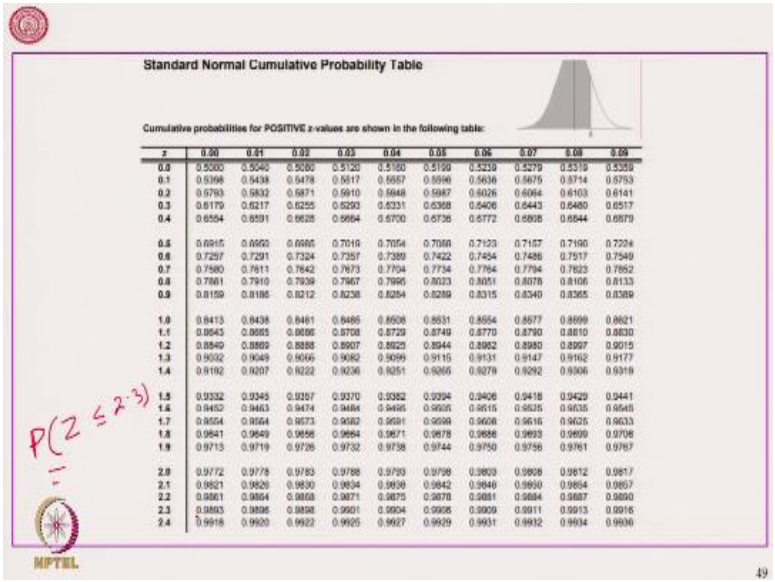


Now let us look at the normal approximation, if m and n are somewhat large say greater than equal to 10 then one can use normal approximation because it is difficult to compute these tables for all different combinations of m and n. So after some point we can look at from large sample theory and go for the normal approximation so $P(W > w) \approx P(Z > z)$ $where\ z = \frac{w-\mu}{\sigma}$ where μ is equal to the mean and σ is the standard deviation.

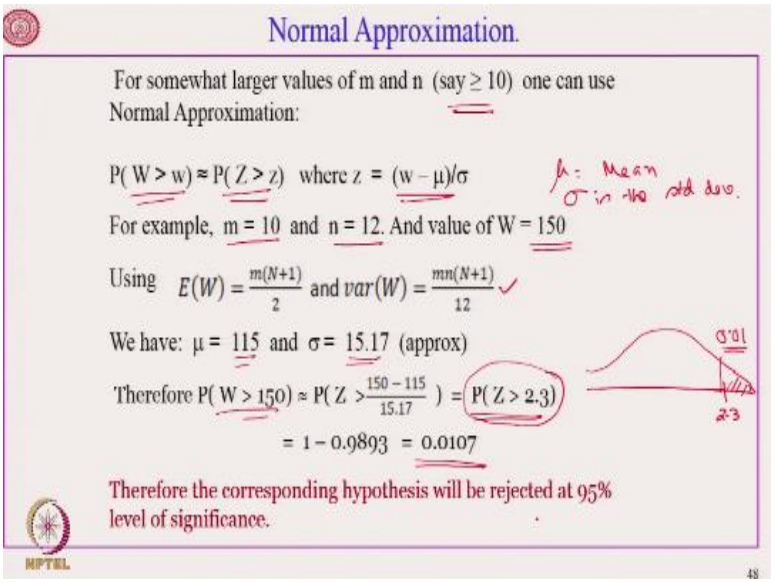
As I have given you the values for mean and the variance or the standard deviation, but those values we will prove again in lecture number 5. So for example when m is 10 and n is 12 and suppose the sum of ranks of X is 150. Question is whether we are going to accept the null hypothesis or reject that one. Since expected value of W is equal to $\frac{m(N+1)}{2}$ and variance of W is equal $\frac{mn(N+1)}{12}$.

Therefore $\mu = 115$ and $\sigma = 15.17$. Therefore probability W greater than 150 after making the normalization is coming out to be probability that a standard normal random variable Z is greater than equal to 2.3.

(Refer Slide Time: 42:12)



(Refer Slide Time: 42:35)



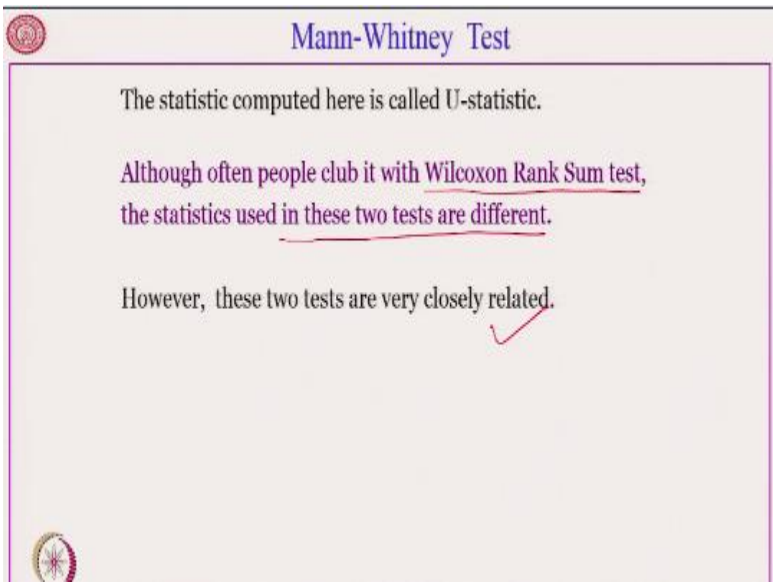
Now let us look at the normal table we can see that probability a normal distribution is less than equal to 2.3, that value is given as 0.9893. Therefore, probability that normal variable is greater than 2.3, that value is coming out to be 0.0107. Therefore, if this is the normal distribution and this is 2.3 then this area is only 0.01 that means that the probability a standard normal variable will take a value greater than 2.3 that probability is only 1 percent.

Now suppose you are testing at 5 percent level of significance therefore at 5 percent level of significance this is going to be rejected.

(Refer Slide Time: 43:25)

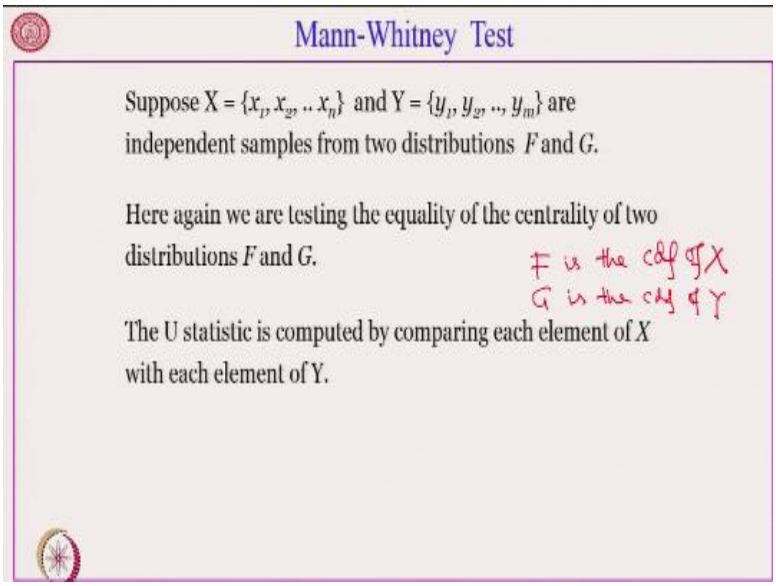
Let us now discuss the other important test which is called Mann-Whitney Test.

(Refer Slide Time: 43:34)



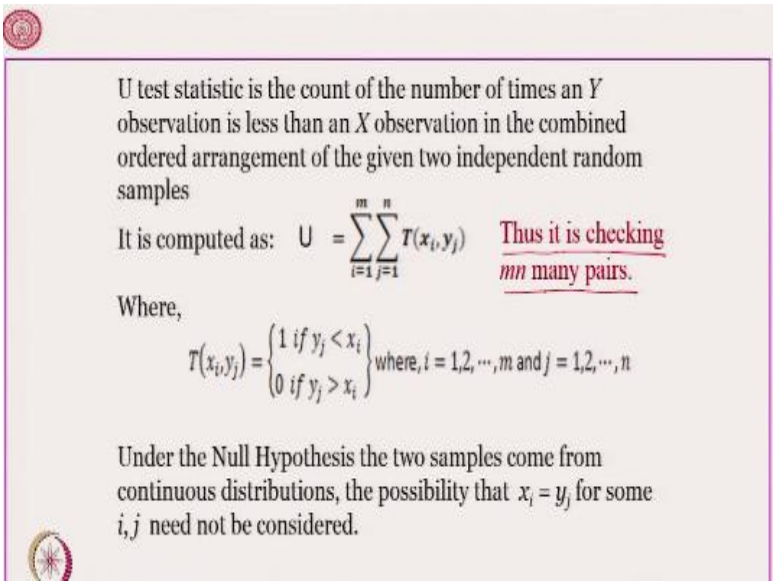
The statistic that is computed here is called the U-statistic or Mann Whitney U-statistic. Often people club it with Wilcoxon Rank Sum test and in some literature you may find people are calling it Wilcoxon Mann Whitney Test, but we have to remember that these two test rank sum test and Mann-Whitney U-Test are slightly different although they are very closely related. So we shall explore the test and then we shall see what is the relation between Mann Whitney Test and Wilcoxon Rank Sum test.

(Refer Slide Time: 44:21)



So, suppose $x_1, x_2, \cdots, x_m$ and $y_1, y_2, \cdots, y_n$ are independent samples from two distributions F and G. So here again we are testing the equality of the centrality of the two distributions F and G where F is the cdf of X and G is the cdf of Y. The U-statistic is computed by comparing each element of X with each element of Y.

(Refer Slide Time: 45:06)

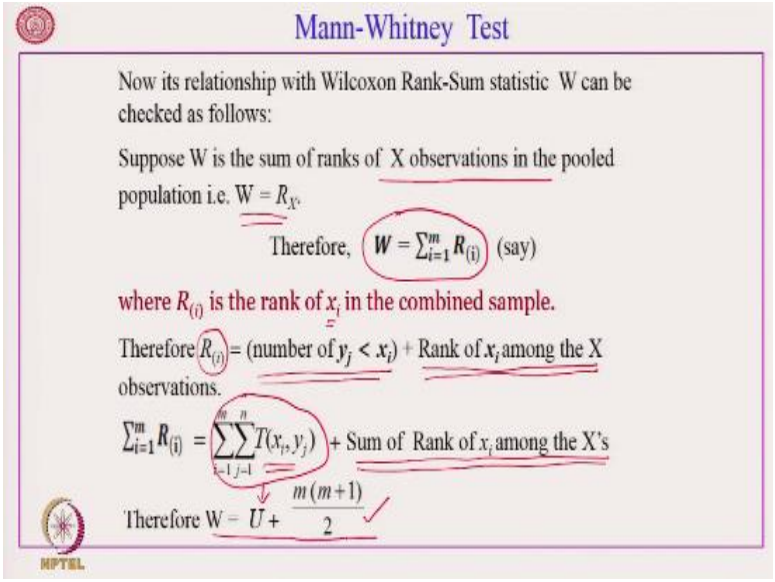


So let me explain basically what we are doing, it is the count of the number of times an Y observation is less than an X observation in the combined ordered arrangement of the given two independent samples. Therefore, it is computed as $U = \sum_{i=1}^{m}\sum_{j=1}^{n} T(x_i, y_j)$ where $T(x_i, y_j) = \begin{cases} 1 \ if \ y_j < x_i \\ 0 \ if \ y_j > x_i \end{cases}$ where, $i = 1,2,\cdots,m$ is the number of X samples and $j = 1,2,\cdots,n$, is the number of Y samples.

Now you may question what happens if $y_j = x_i$. So typically we discard such things because under the null hypothesis the two samples come from continuous distributions. Therefore the possibility that $x_i = y_j$ for some i, j need not be considered, but in reality we may get some such data, but in reality we may get some such data.

Therefore we discard both of them from the two populations in our subsequent computation and accordingly the values of m and n need to be adjusted. Therefore, if we look at it we are checking mn many pairs because each X observation will be compared to all the Y observations and vice-versa.

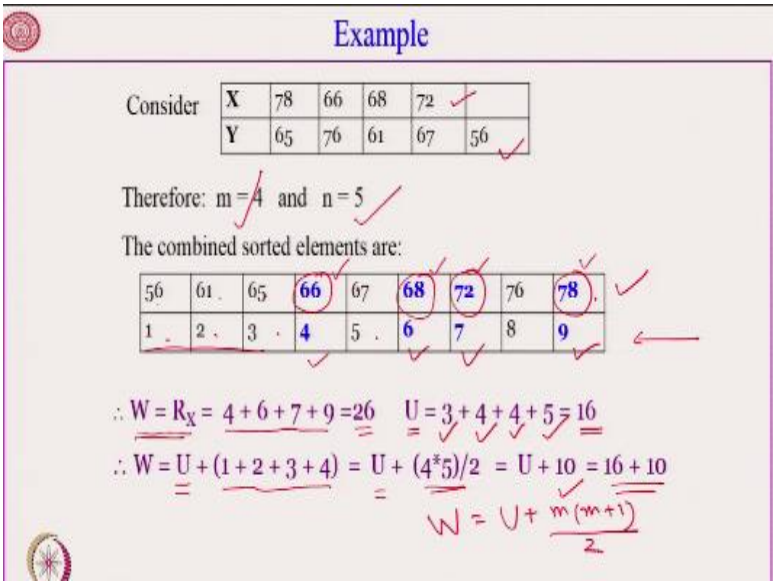
(Refer Slide Time: 46:50)



Now question comes what is the relation between Wilcoxon Rank-Sum statistic W and Mann Whitney U Statistic? Suppose W is the sum of ranks of X observations in the pooled population therefore W is equal to R_X which is the sum of ranks of the X observations. So we can write $W = \sum_{i=1}^{m} R_{(i)}$ this $R_{(i)}$ denotes the rank of the i^{th} X observation in the combined sample.

Therefore, what is $R_{(i)}$ that is the rank of the i^{th} X sample, this is going to $\left(\text{number of } y_j < \text{x}_\text{i}\right) + \text{Rank of } x_i \text{ among the X observations.}$ Therefore $\sum_{i=1}^{m} R_{(i)} = \sum_{i=1}^{m}\sum_{j=1}^{n} T(x_i, y_j) + \text{sum of rank of } x_i \text{ among the X observations}$ Now among the X's we have m observations therefore the sum of ranks is going to be $\frac{m(m+1)}{2}$ and therefore this element which is nothing but the U-Statistic is coming here and from here we can find that the relationship is that $W = U + \frac{m(m+1)}{2}$.

(Refer Slide Time: 48:40)



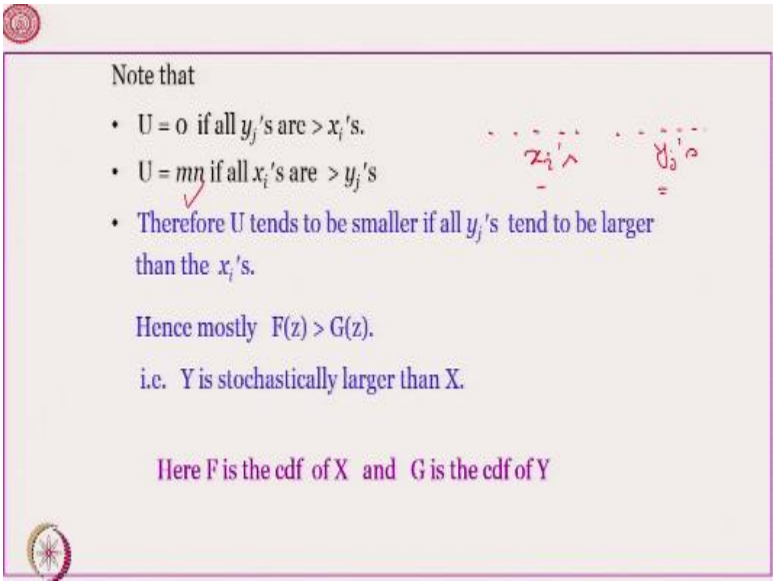
So let me give you an example here consider X is equal to 78, 66, 68 and 72 that is there are 4 observations therefore m is equal to 4. Y is equal to 65, 76, 61, 67 and 56 therefore n is 5. Now in the combined population when we arrange them in sorted order it looks like this of which the blue ones are X observations and their ranks in the combined population are 4, 6, 7 and 9.

Therefore the Wilcoxon rank sum statistic W is equal to $4 + 6 + 7 + 9$ which is equal to 26. What is U-Statistic? We are checking how many Y's are less than each X. Therefore, we are comparing 1, 2, 3, 4, 5 all 5 of Y's with 66, 68, 72 and 78 that is effectively we are comparing 20 pairs of observation, but from these sorted array we can easily count that there are 3 Y's which are less than 66 thus I get a 3.

There are 4 Y's, 1, 2, 3 and 4 which are less than 68 same 4 are also less than 72 and all 5 Y's are less than 78 that is why the U-Statistic is equal to $3 + 4 + 4 + 5 = 16$. Therefore we can see that $W = U + (1 + 2 + 3 + 4)$ because the sum of ranks of X elements within themselves are going to be 1, 2, 3, 4 that is $U + \frac{4*5}{2}$ therefore 16 plus 10.

Hence we verify that relationship between W and U that $W = U + \frac{m(m+1)}{2}$

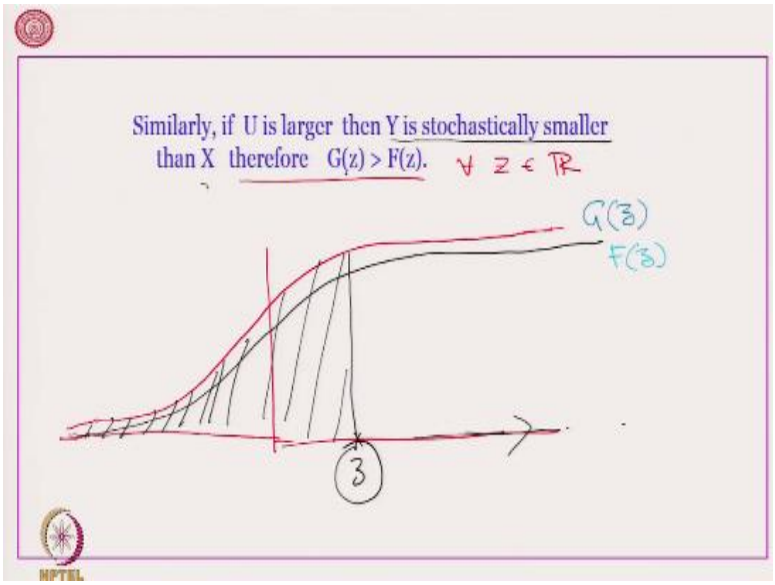
(Refer Slide Time: 51:16)



Now note that $U = 0$ if all y_j 's are $>$ all x_i 's,that is pretty obvious because if $x_i's$ are this and $y_j's$ are this then all x_i is less than y_j there is not a single instance when an Y observation is less than an X observation therefore the value of the summation is going to be 0. On the other hand if all the $x_i's$ are greater than all the $y_j's$ therefore for all the mn many pairs will get a score 1 and thus we shall get the value mn.

Therefore U tends to be smaller if all $y_j's$ tend to be larger than the $x_i's$. Hence, mostly $F(z)$ is going to be greater than $G(z)$ that is Y is stochastically larger than X. Here F is the cdf of X and G is the cdf of Y.

(Refer Slide Time: 52:33)



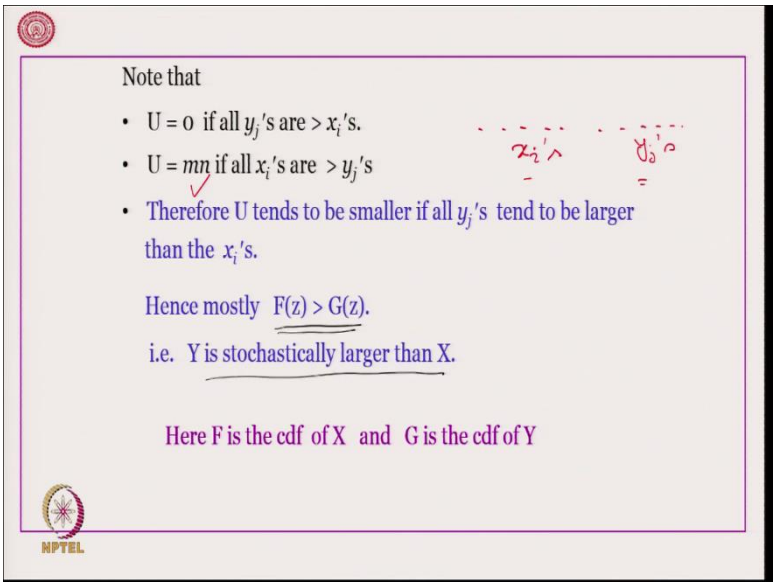
Similarly, if U is larger than Y is stochastically smaller than X and therefore $G(z)$ is greater than $F(z)$for all z belonging to $\mathbb{R}$. So let me explain this with a diagram. Suppose $G(z)$ is the

distribution of Y and suppose it has a shape like this we know that each distribution function at minus infinity is going closer to 0 and at plus infinity it is going towards 1 so let us call it $G(z)$

Now G(z) is greater than F(z) for all z therefore how will F(z) look like. It will look like something like this although eventually it is also going to touch 1, but most of the values for the real number z we can see that G(z) is greater than F(z), so let us call it F(z). Therefore, what we can say? We can say that given any particular value of z say this one, the G(z) is greater than F(z), therefore the proportion of Y values is bigger than the proportion of X values which are less than equal to z.

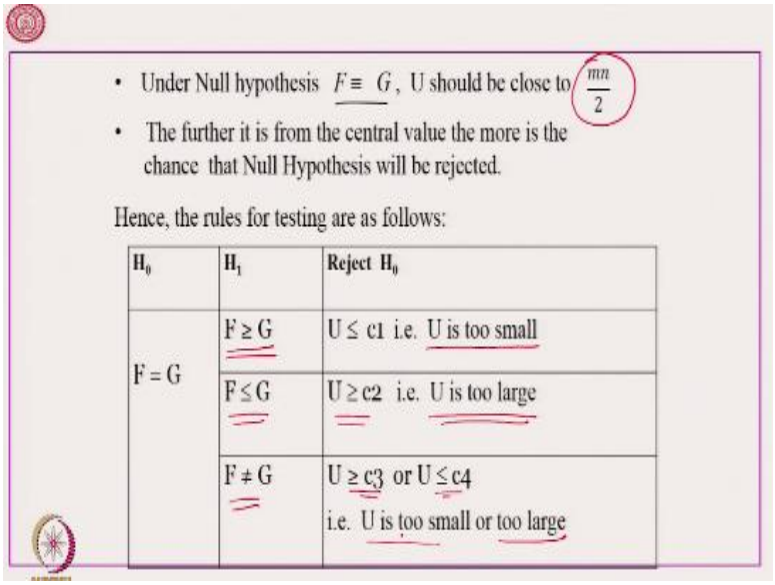
Therefore, most of the X values are going to happen on the other side of z that is on this side we shall get more X values than Y values that is why we say that Y is stochastically smaller than X if G(z) is greater than F(z).

(Refer Slide Time: 55:14)



Similarly, as we have seen in the previous slide that F(z) is greater than G(z) implies Y is stochastically larger than X.

(Refer Slide Time: 55:26)

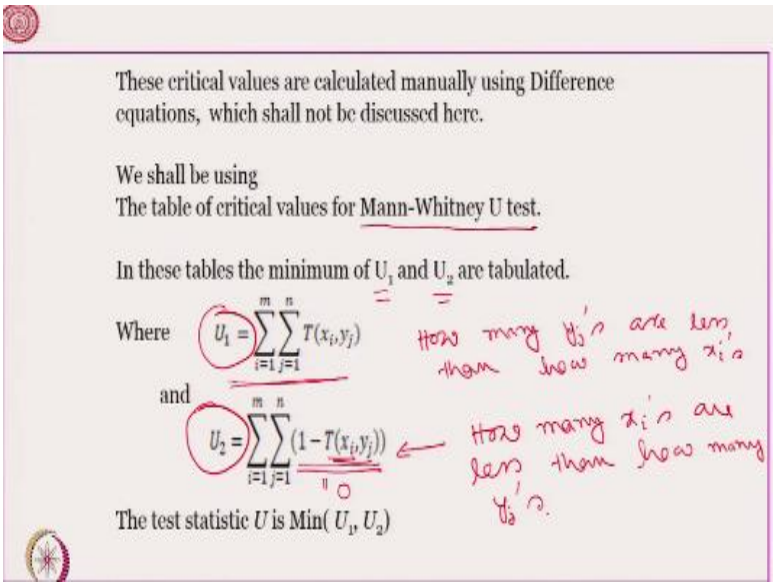


Therefore, the test that we want to do whether they have the similar distribution that is $F = G$ then U should be close to $mn/2$ because that is the average value because the range of U is from 0 to mn and therefore what we can see that the average value will come to be $mn/2$. The further it is from the central value the more is the chance that null hypothesis is going to be rejected.

Therefore, depending upon the alternative the rejection criteria will change. For example if alternative is $F \geq G$ then that means that Y has more spread than X that means that Y is stochastically larger than X then we shall reject the null hypothesis if U is too small. Similarly, if the alternative is $F \leq G$ then we shall reject the hypothesis null hypothesis if U is too large that is U is greater than equal to some critical value.

On the other hand if we have a two-sided alternative then we should check whether U is too small or too large that is it is greater than some threshold or less than some other critical value and based on that we will be rejecting the null hypothesis.

(Refer Slide Time: 57:10)

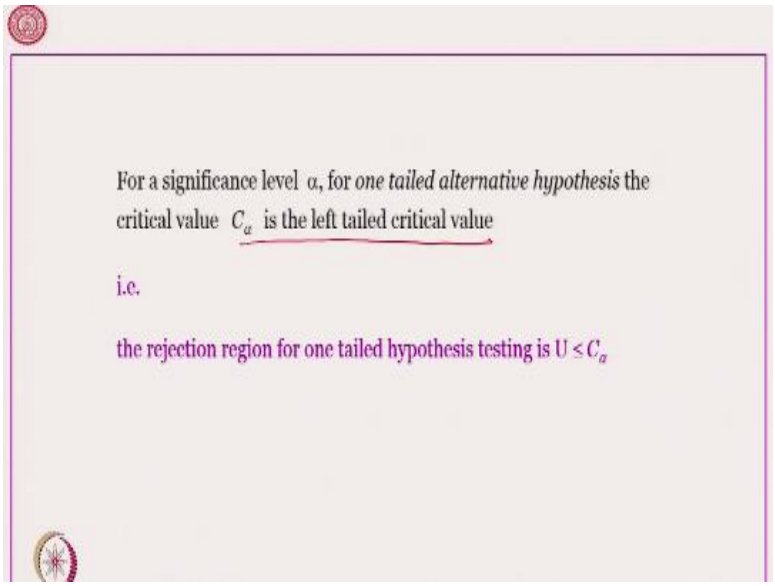


These critical values are calculated manually using difference equation which we are not going to discuss here, but let me tell you a practical way of using this. The table of critical values is available which is called Mann-Whitney U Test, but for practicality we actually compute 2 different statistics let them be U_1 and U_2 where U_1 is the one that we have already discussed $\sum_{i=1}^{m}\sum_{j=1}^{n} T(x_i, y_j)$ where $T(x_i, y_j) = 1$ if $y_j < x_i$.

Therefore here we are checking how many $y_j's$ are less than how many $x_i's$. So that gives us the statistic U_1. Now let us consider $U_2 = \sum_{i=1}^{m}\sum_{j=1}^{n}(1 - T(x_i, y_j))$. Note that if $T(x_i, y_j) = 1$ then this value is equal to 0 and if $T(x_i, y_j) = 0$ then this value is coming out to be 1. Therefore, effectively this counts how many $x_i's$ are less than how many $y_j's$.

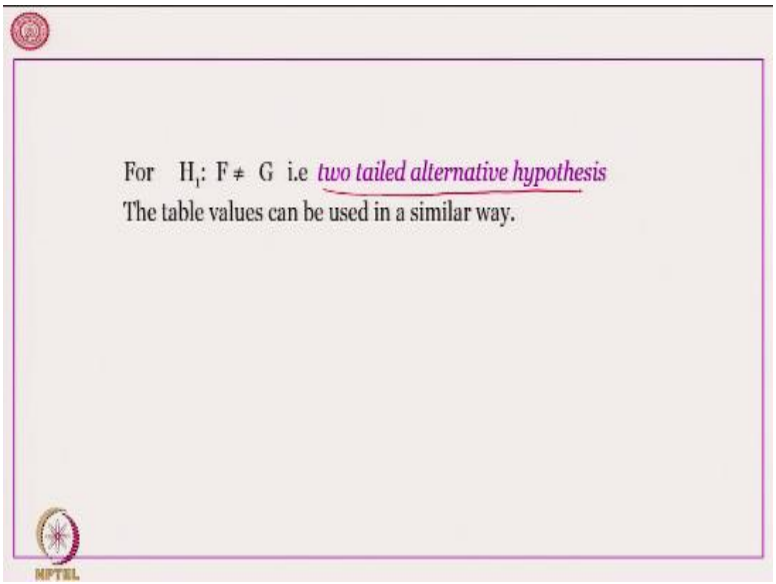
Therefore, when we calculate U_1 and U_2 we take the minimum of them and we call that the U Statistic.

(Refer Slide Time: 59:15)



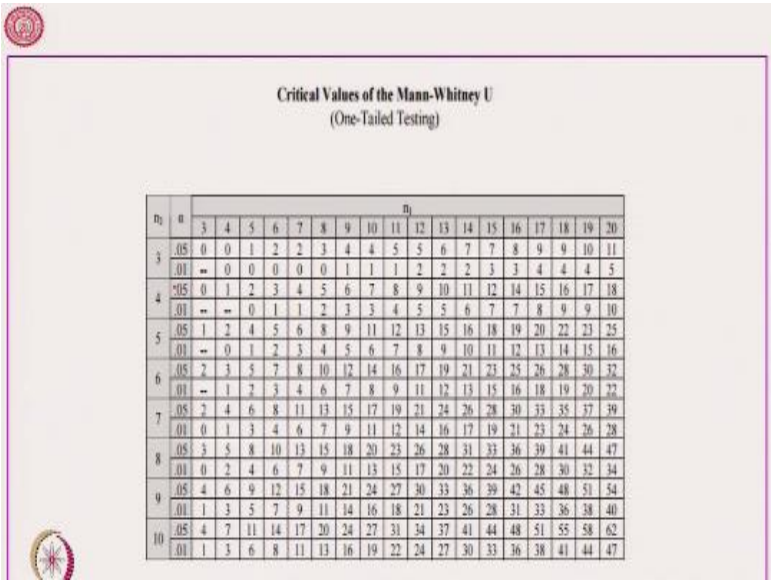
So, for a significance level α for one tailed alternative hypothesis the critical value C_α is the left tailed critical value. We will check if it is smaller than some critical value C_α or not and that is what we want to test.

(Refer Slide Time: 59:34)



For $H_1: F \neq G$ then we shall check in a two tailed alternative hypothesis in a very similar way.

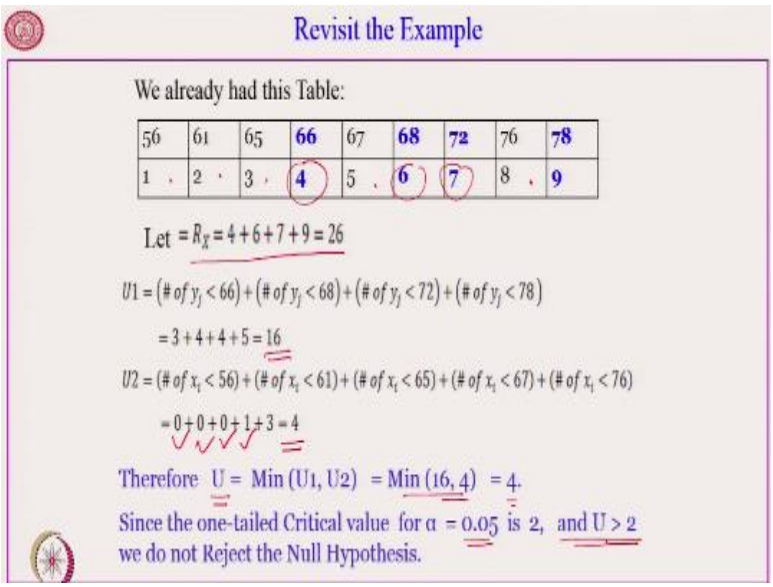
(Refer Slide Time: 59:44)



n_2	α	n_1 = 3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
3	.05	0	0	1	2	2	3	4	4	5	5	6	7	7	8	9	9	10	11
	.01	--	0	0	0	0	0	1	1	1	2	2	2	3	3	4	4	4	5
4	.05	0	1	2	3	4	5	6	7	8	9	10	11	12	14	15	16	17	18
	.01	--	--	0	1	1	2	3	3	4	5	5	6	7	7	8	9	9	10
5	.05	1	2	4	5	6	8	9	11	12	13	15	16	18	19	20	22	23	25
	.01	--	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
6	.05	2	3	5	7	8	10	12	14	16	17	19	21	23	25	26	28	30	32
	.01	--	1	2	3	4	6	7	8	9	11	12	13	15	16	18	19	20	22
7	.05	2	4	6	8	11	13	15	17	19	21	24	26	28	30	33	35	37	39
	.01	0	1	3	4	6	7	9	11	12	14	16	17	19	21	23	24	26	28
8	.05	3	5	8	10	13	15	18	20	23	26	28	31	33	36	39	41	44	47
	.01	0	2	4	6	7	9	11	13	15	17	20	22	24	26	28	30	32	34
9	.05	4	6	9	12	15	18	21	24	27	30	33	36	39	42	45	48	51	54
	.01	1	3	5	7	9	11	14	16	18	21	23	26	28	31	33	36	38	40
10	.05	4	7	11	14	17	20	24	27	31	34	37	41	44	48	51	55	58	62
	.01	1	3	6	8	11	13	16	19	22	24	27	30	33	36	38	41	44	47

So let us look at the Mann-Whitney U-Statistic for two-tailed the values are given for different values of m and n, m and n which are given n_1 and n_2 respectively. In our case similarly there is one-tailed testing so that the value for different m and n is given for one-sided test and the alpha are taken to be 0.05 and 0.01.

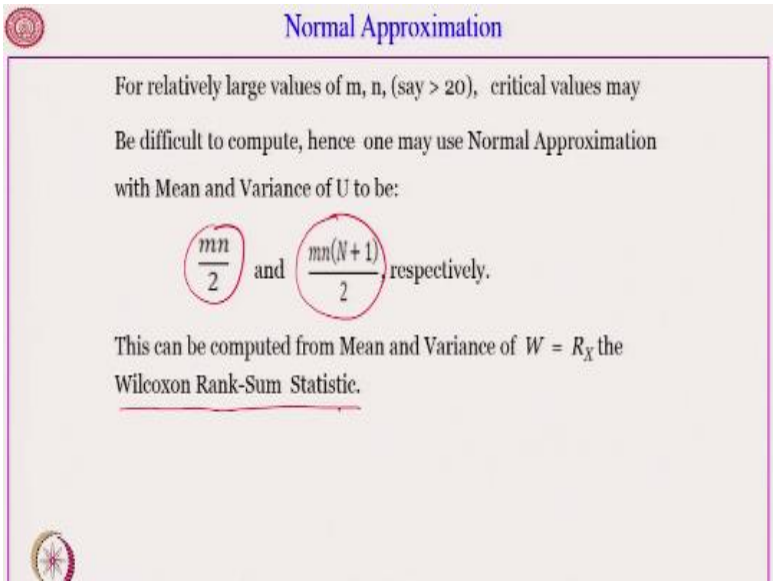
(Refer Slide Time: 1:00:20)



So let us revisit the example. We have this is the sorted value we have already calculated $R_X = 26$, U_1 is coming out to be how many Y's are less than X this also we have computed to be 16 I am not repeating that, but in a similar way U_2 is going to be this is going to be 0 for this because no X is smaller than this Y, 0 for this, 0 for this, 0 for this therefore we got 3 0s.

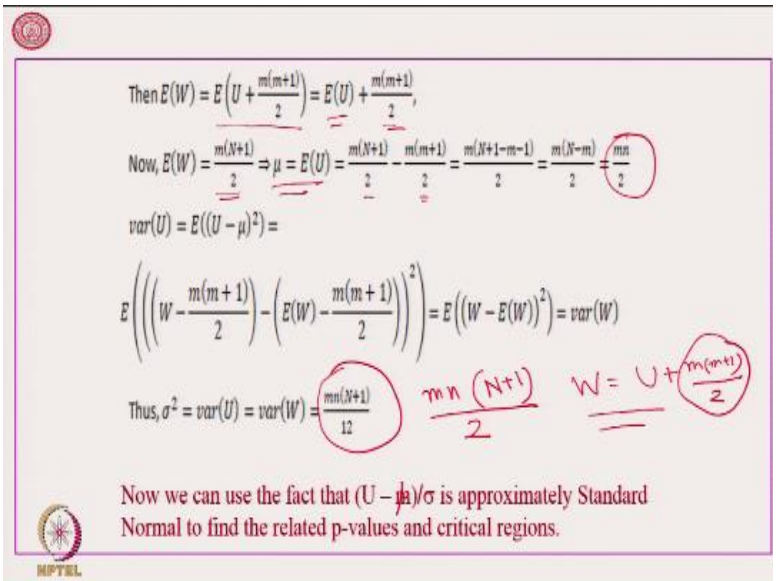
It is going to be 1 for 67 because there is an X value smaller than this and this is going to be 3 for 76 as there are 3 X value smaller than this therefore the value of the statistic is 4. Hence, Mann-Whitney U-Statistic is equal to minimum of 16 and 4 is equal to 4. So suppose we are now looking at one-tailed critical value for alpha is equal to 0.5. If we go back to the table for one-tailed, for 0 and 5 the value at 5 percent level is given out to be 2. Therefore since U is greater than 2 we do not reject the null hypothesis.

(Refer Slide Time: 1:01:49)



For a relatively large values of m,n critical values maybe difficult to compute hence one may use normal approximation with mean and variance is equal to $mn/2$ and $\frac{mn(N+1)}{2}$ respectively. This can be computed from mean and variance of W is equal to R_X which is the Wilcoxon rank sum statistic.

(Refer Slide Time: 1:02:20)



How we will compute? $E(W) = E\left(U + \frac{m(m+1)}{2}\right) = E(U) + \frac{m(m+1)}{2}$,. Now we have already stated that although we did not prove it yet that $E(W) = \frac{m(N+1)}{2}$ implies $E(U) = \frac{m(N+1)}{2} - \frac{m(m+1)}{2} = \frac{mn}{2}$

Variance, I have calculated it, but actually one does not need to calculate explicitly because we know that the relationship between W and U are $W = U + \frac{m(m+1)}{2}$. Therefore because it is a constant we know that variance in W and variance of U are going to be the same. Therefore, we can see that same value of variance will come here which is $\frac{mn(N+1)}{12}$.

Therefore, we can use the fact that $(U - \mu)/\sigma$ is approximately standard normal to find the related p values and critical regions. Okay friends I stop here today. So in this class we have studied two important nonparametric tests namely Wilcoxon rank sum test and Mann-Whitney Test. These are used for testing the equality of the central location of two different populations.

But as I said at the very beginning that when we compare two different distributions we not only look at the centrality, we also look at the dispersion. Therefore, it is not enough to check the medians of the two samples we should also check whether their dispersions are also same nature or not, or whether we can accept that they have the same level of dispersion. The corresponding problem is called the two sample scale problem.

In the next class we shall start with this. Okay friends I stop here today. Thank you.