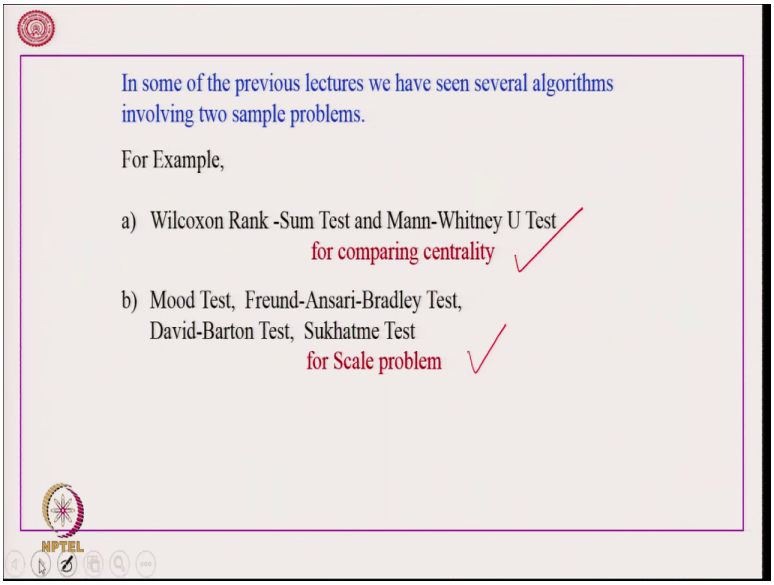
Nonparametric Statistical Inference
Professor. Niladri Chatterjee
Department of Mathematics
Indian Institute of Technology, Delhi
Lecture No. 08

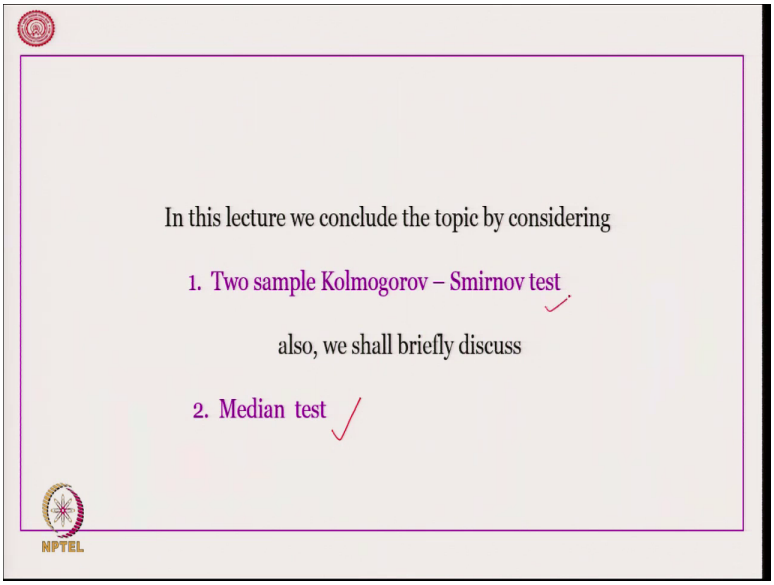
Welcome students to MOOCs series of lectures on non-parametric statistical inference. This is lecture number 8.

(Refer Slide Time: 0:31)



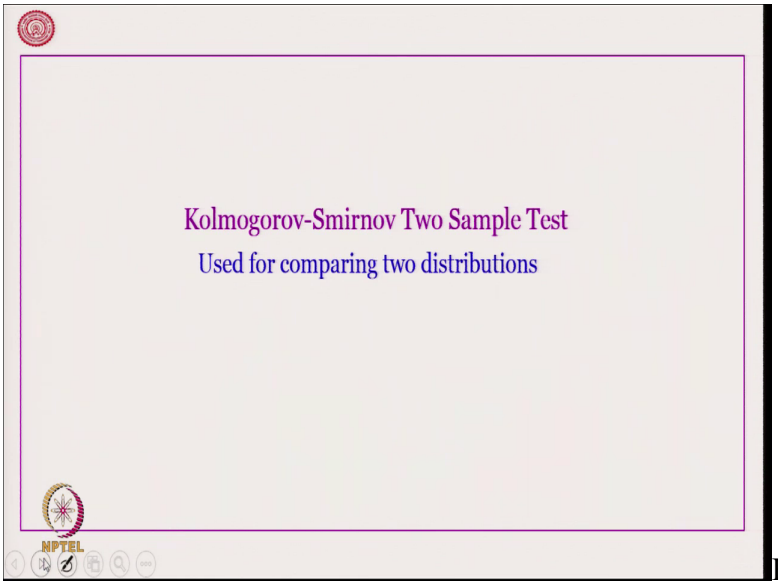
In some of our earlier lectures we have seen several tests for two samples for example, we have studied Wilcoxon rank sum test and Mann Whitney U test for comparing the centrality or central locations of two populations. We have also studied for scale problem mood test, Freund Ansari Bradley test, David Barton test and Sukhatme test etc. In this lecture we shall conclude the topic by studying some more tests.

(Refer Slide Time: 1:17)



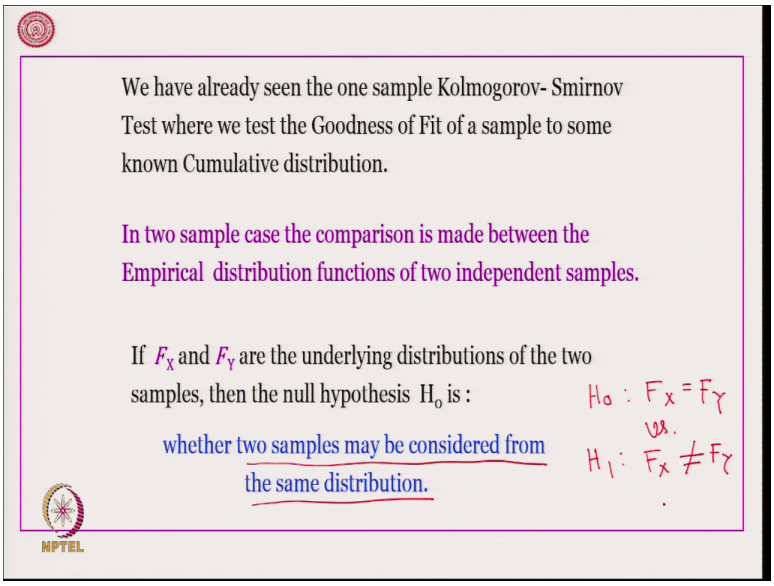
In particular the one that we study extensively is two sample Kolmogorov Smirnov test. This is typically used for comparing the distribution of two populations and also we shall briefly discuss the median test for the same purpose but definitely the two sample Kolmogorov Smirnov test is much powerful than the median test.

(Refer Slide Time: 1:52)



Kolmo have already mentioned that it is used for comparing two distributions.

(Refer Slide Time: 2:03)

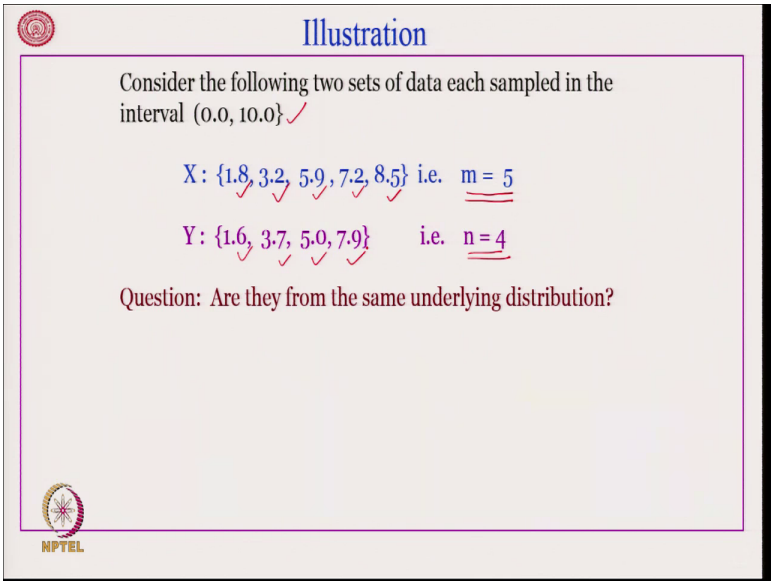


We have already seen one sample Kolmogorov Smirnov test where we are comparing whether the sample that we have taken from a population, that population is actually having some known distribution that is what we have written known cumulative distribution. However, in two sample test we are not comparing a sample with a known distribution rather we are checking whether the two samples are coming from the same distribution or having the same empirical distribution.

So, we have to understand that we do not know the actual distribution of any of the two samples but just we want to check if they are having same distribution.

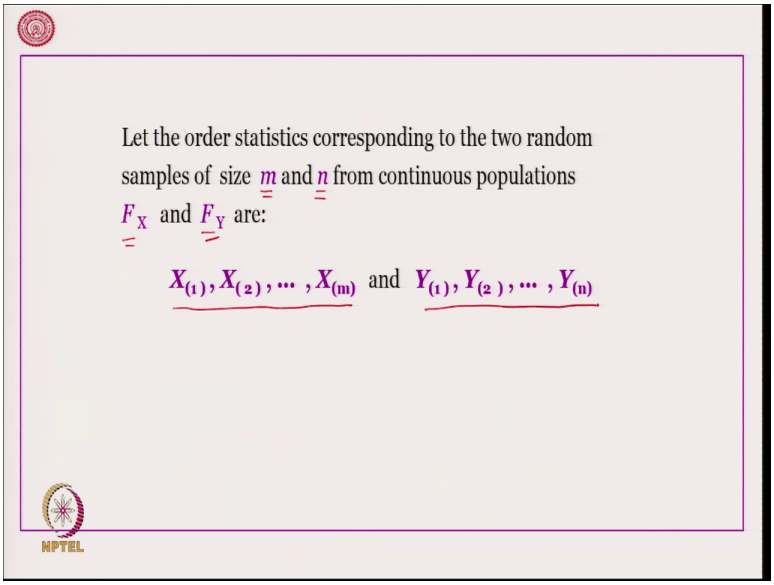
So, let Fx and Fy be the underlying distributions of the two samples then the null hypothesis Ho is that whether the two samples may be considered from the same distribution. Therefore, the null hypothesis Ho we can write it as Fx is equal to Fy against the two sided alternative that is whether they are different from each other. So, this is what we want to test.

(Refer Slide Time: 3:49)



So, let me first give you an example suppose we have taken samples from the interval 0 to 10. Sample 1 that is X we have selected 5 points therefore we are saying m is equal to 5 and sample Y which has 4 points which we are saying n is equal to 4 and the values are as I have written 1.8, 3.2, 5.9, 7.2 and 8.5 for X and 1.6, 3.7, 5 and 7.9 for Y. Can we say that they are from the same underlying distribution?

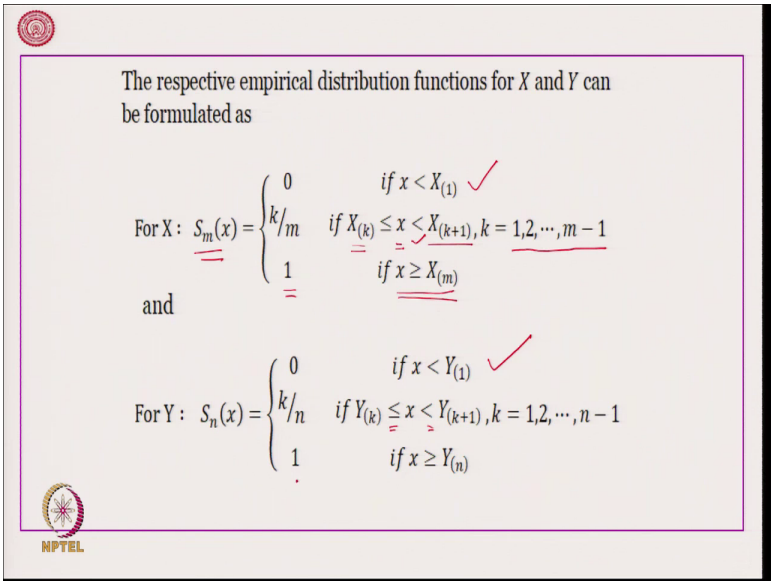
(Refer Slide Time: 4:36)



So, in general the problem can be stated that we have two samples of size m and size n and they are respectively coming from distribution Fx and Fy. We want to check if they have the same

underlying distribution so we have taken the order statistic x1 x2 xm and y1 y2 yn. We consider the order statistics because that helps us in computing the sample cumulative distribution.

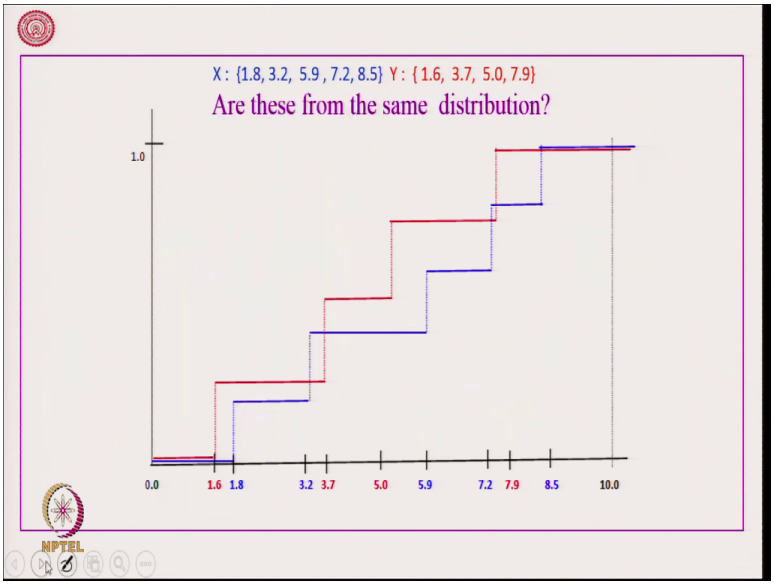
(Refer Slide Time: 5:17)



Therefore the respective empirical distribution functions for X and Y can be formulated as for X it is Sm X which is 0 if x is less than X1 which is k by m if the kth order statistic less than equal to X and strictly less than Xk plus one for value of k is equal to 1, 2, 3 up to m minus one and for any x greater than the mth order statistic or the largest of the X the value of the empirical cumulative distribution function is one.

In a similar way, for Y we compute the empirical distribution as zero if the value of X is less than the first order statistic it is k by n if the value of X is between the kth order statistic and the k+1th order statistic. Note that here it is less than equal to but here it is strictly less than the same was here as well and if the value of X is greater than the nth order statistic then the value is 1.

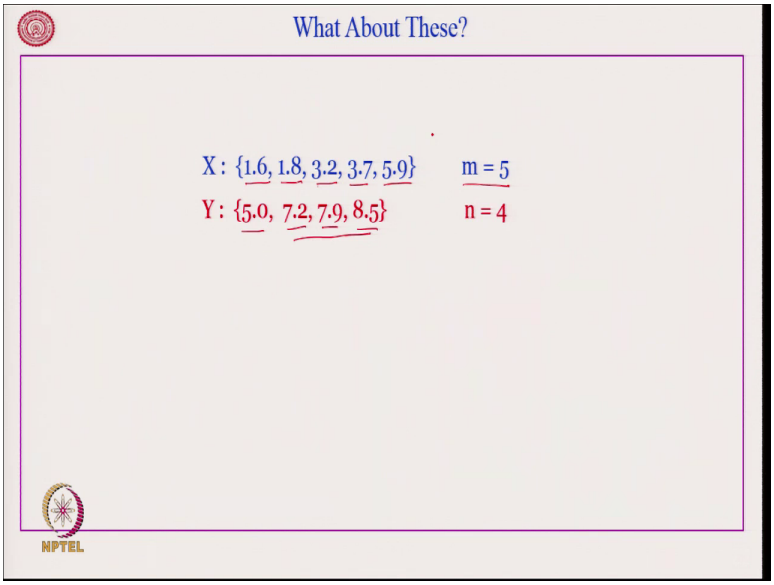
(Refer Slide Time: 6:38)



So, let us plot them. Here you can see for X we have the blue colored cumulative distribution function based on the sample. If we assume all of them are equally likely then each one of them is giving or causing a jump of 0.2 and like that. After the fifth order statistic it attains the value 1. So, the blue one is for X. Similarly, for Y there are 4 sample values therefore at each step it is being raised by point 25 and after the nth order or the fourth order statistic it attains the value 1.

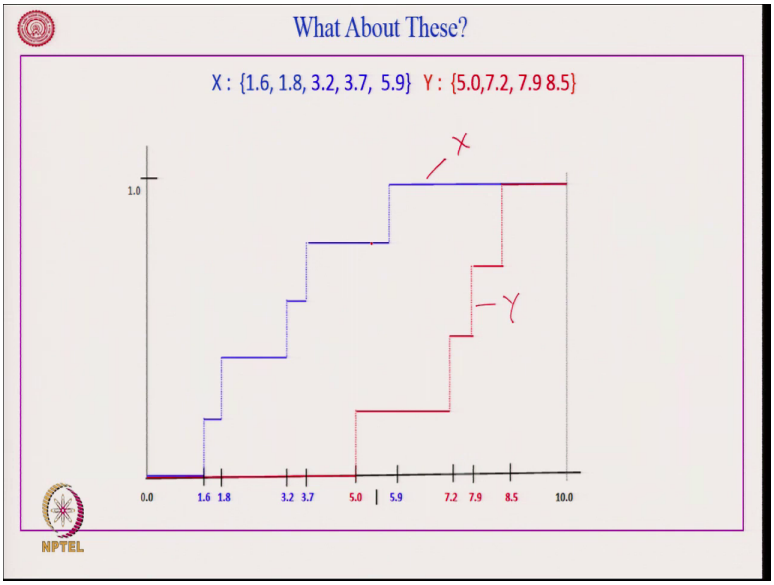
Now in front of us we have the two empirical distribution functions. Can we say that they are coming from the same distribution? So how to do that, basically we shall see if the two sample distributions are really close to each other. Now whether they are close or not that will depend upon certain threshold and that is the fundamentals of Kolmogorov Smirnov test. We are coming to that but let us consider another two samples.

(Refer Slide Time: 8:07)



Say for X, as before we have taken 5 samples. The values are 1.6, 1.8, 3.2, 3.7 and 5.9 and for Y the values are 5.0, 7.2, 7.9, 8.5. What does it mean? If we observe carefully we can see that most of the Y values are larger than the X values. Therefore, we feel that Y is stochastically larger than X. But how do we justify?

(Refer Slide Time: 8:46)

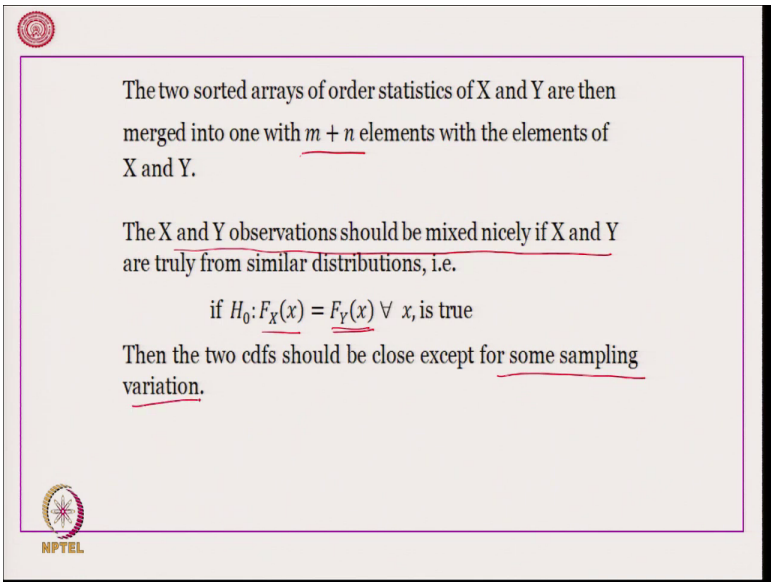


So, we again we compute their respective empirical distribution function and if we compare the blue ones is for X and the red one is for Y and when we compare the distribution functions we can see that they are not at all close to each other.

(Refer Slide Time: 9:07)

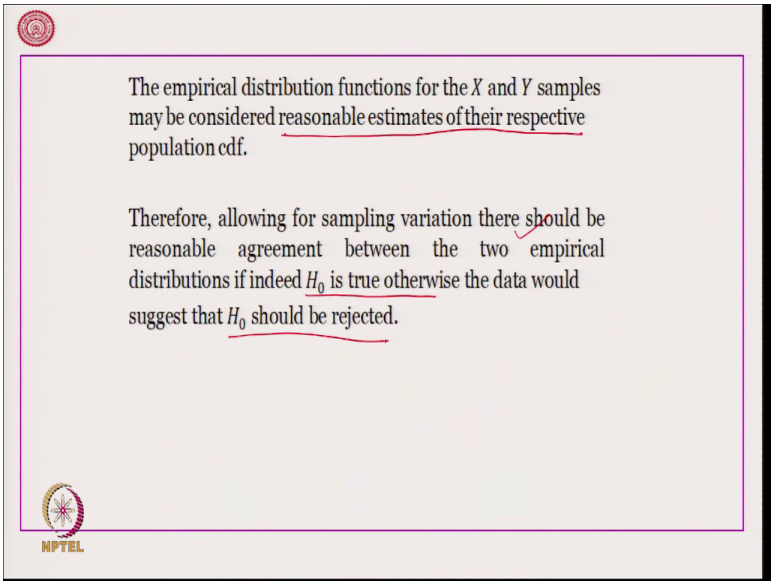
Therefore, we need to now design the test for testing Ho.

(Refer Slide Time: 9:13)



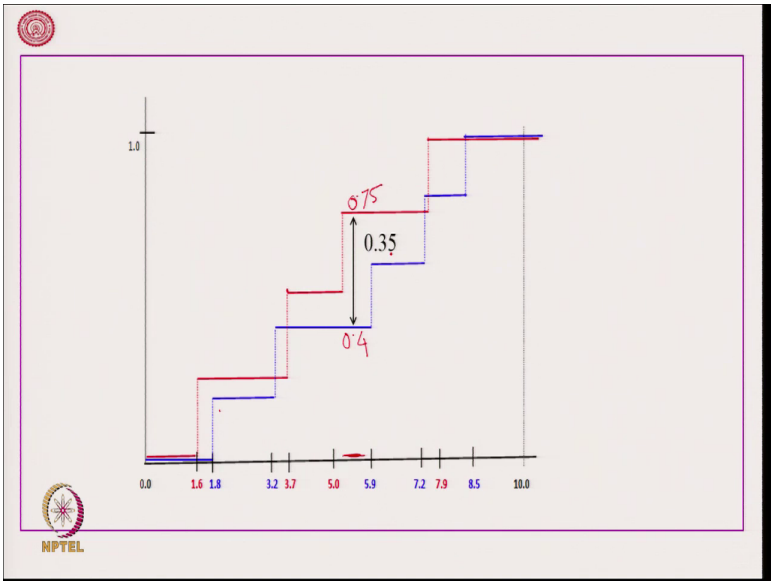
So, it goes as follows, the two sorted arrays of order statistics of X and Y are then merged into one with m plus n elements combined X and Y elements. If they are coming from the same distribution, then the X and Y observation should be mixed nicely that is for all X roughly Fx(x) and $F_Y(x)$ should be very close except that there may be some sampling variation.

(Refer Slide Time: 9:50)



And the empirical distribution functions for X and Y samples we can consider them to be the reasonable estimates of their respective population cumulative distribution function. Therefore allowing some sampling variation there should be reasonable agreement between the two empirical distributions. If indeed the null hypothesis Ho is true and if that is not satisfied then we are going to say that Ho should be rejected.

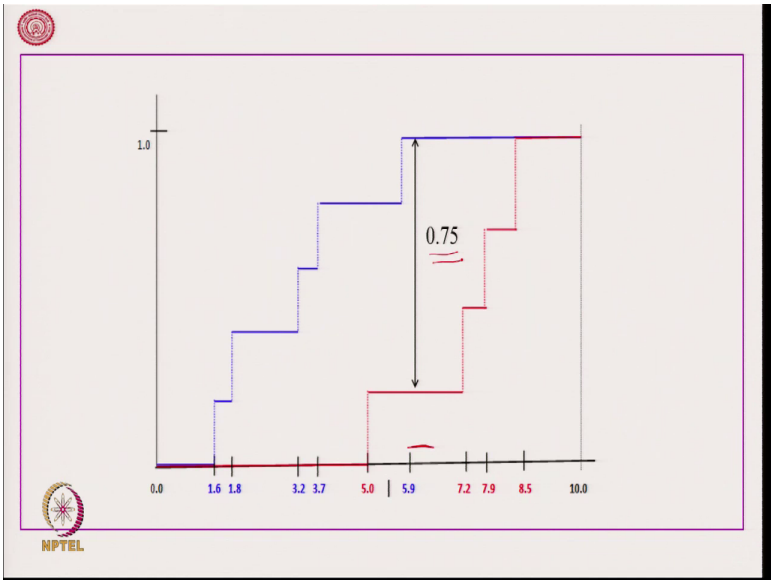
(Refer Slide Time: 10:26)



So, let us go back to the two diagrams that we have seen. For the first one if we compare the two empirical distribution functions we can see that the maximum difference coming out to be for X

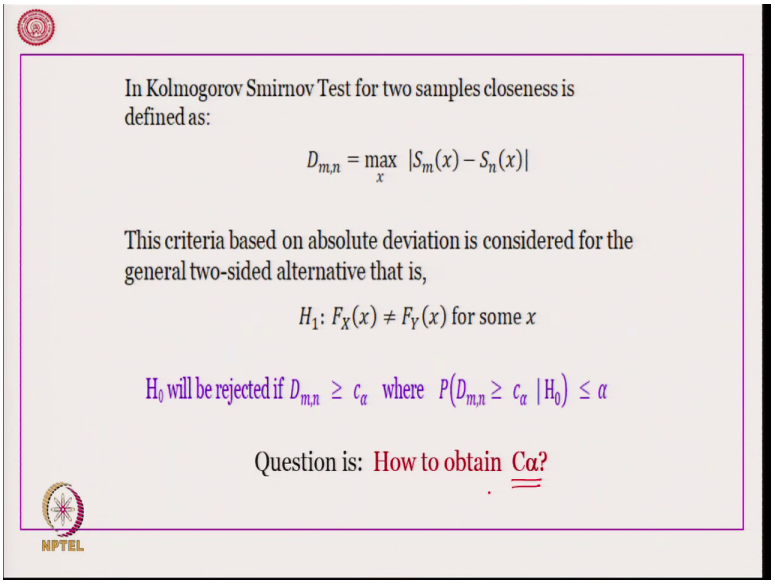
around this region and the difference in their F values is 0.35. How do we get that? Because at this point it is 0.2 + 0.2 therefore 0.4 and at this point it is 0.25 + 0.25 + 0.25. Therefore, 0.75. Therefore, the difference is coming out to be 0.35.

(Refer Slide Time: 11:19)



On the other hand with respect to the second diagram in a similar way we can see that around this region the maximum difference that is coming and that magnitude is 0.75.

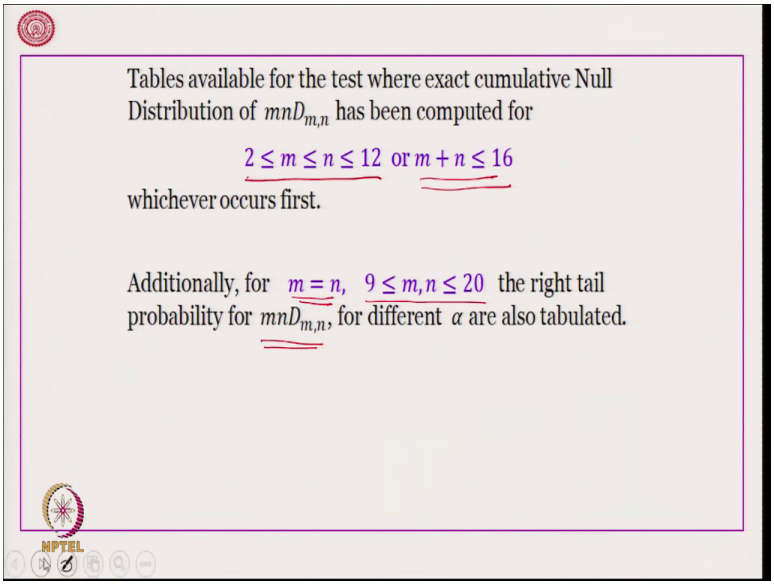
(Refer Slide Time: 11:37)



The Kolmogorov Smirnov test for two samples closeness is defined as the maximum over X the modulus of Sm(x) – Sn(x). So, what is Sm(x)? It is the proportion of X observations ≤ x and similarly it is the proportions of Y observations less than equal to X. So, we look at the maximum difference between them and if the deviation is greater than some threshold then we shall reject the null hypothesis in favor of the alternative that they are not the same.

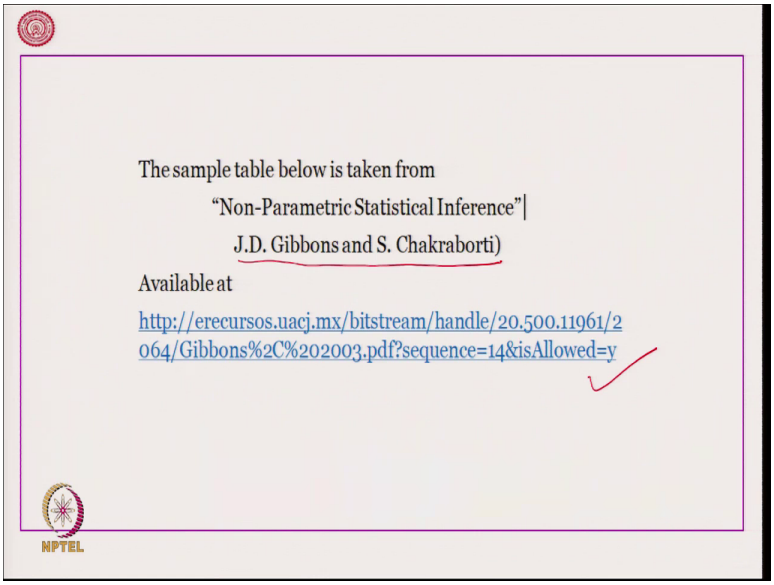
That is what we have written that Ho will be rejected if Dm,n which is the maximum of the difference maximum taken over all X's is greater than some critical value Cα for level of significance α where probability Dm,n ≥ Cα under the null hypothesis is ≤ α. The question is how to obtain these critical values Cα? And that is what Kolmogorov Smirnov two sample test is all about.

(Refer Slide Time: 13:32)



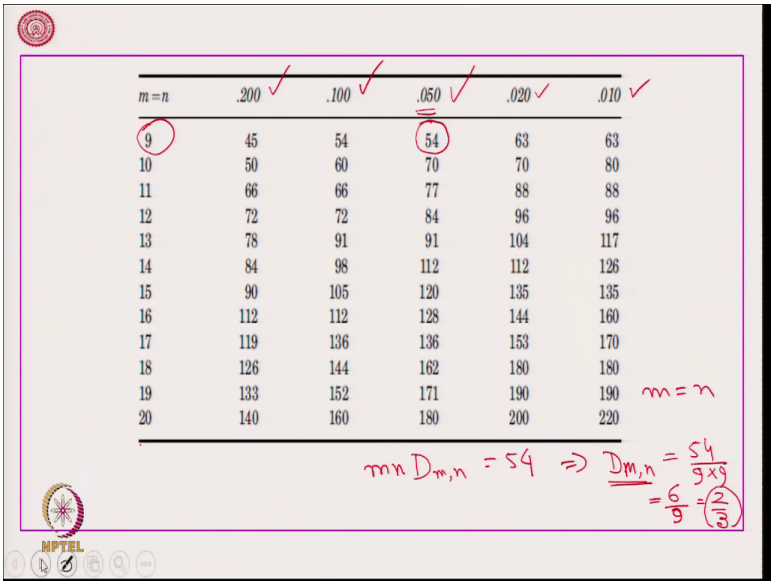
So, tables are available for this test for m and n between 2 to 12 such that their sum is less than equal to 16. Additionally, for situations when m is equal to n that is the sample size is same for both X and Y then for all the values between 9 to 20, the right tail probability for m, n multiplied by Dm,n is tabulated for some different values of α.

(Refer Slide Time: 14:17)



So, let us first look at the table which we have taken from this URL and it is from the book by J.D. Gibbons and S Chakraborti on non-parametric statistical inference. I have been referring to this book many times. So, if you want you can look at it, it is available over the internet and you can look at this table.

(Refer Slide Time: 14:42)



Say for example, when m is equal to n then for 9 to 20, the value the critical values is given for α is equal to 0.2, point 1, 0.05, 0.02 and 0.01. So, what does it mean? So, let us consider m is equal

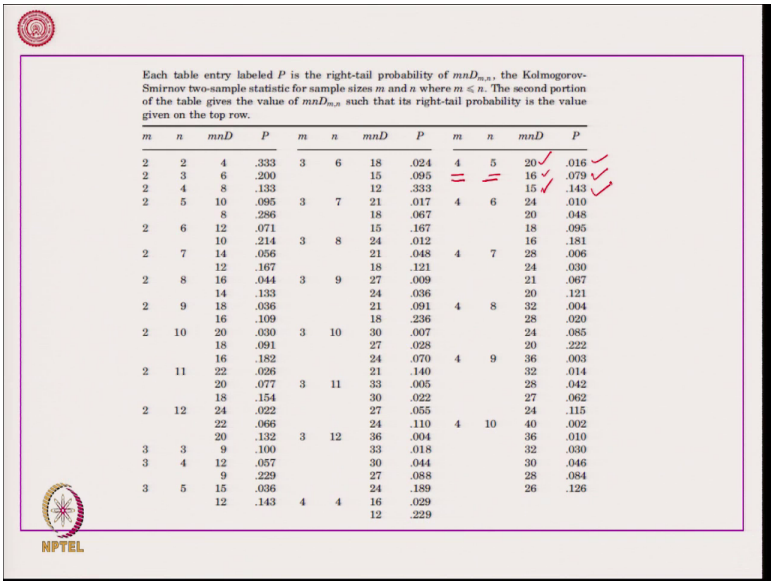
to n is equal to 9 and let us look at 5% level of significance. So, the value given is 54. Therefore m, n into Dm,n is equal to 54, since m and n are both are 9 that means that D m,n or $D_{9,9}$ is equal to 54 upon 9 x 9 = 6 / 9 = 2 /3. What does it mean?

It means that if the sample size are 9 from both X and Y populations then the maximum difference in their empirical distribution that value will be ≥ 2 / 3 or say 0.67 has a probability is point 05 that is 5%.

So, let me repeat if two samples are taken of size 9 and 9 then the probability that the maximum difference between their empirical distribution functions will be greater than equal to point 67 that probability is point 05.

In a similar way for each m, n you can calculate the critical value for Dm,n where m is equal to n.

(Refer Slide Time: 16:58)

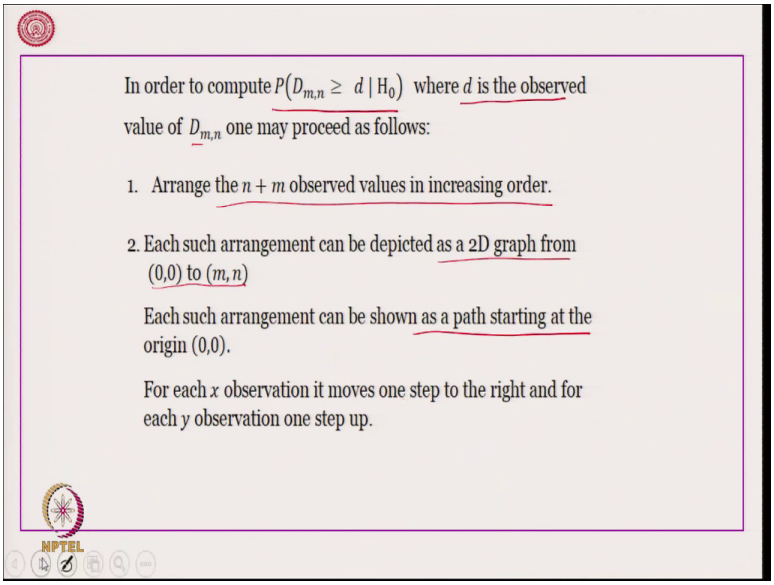


Each table entry labeled P is the right-tail probability of $mnD_{m,n}$, the Kolmogorov-Smirnov two-sample statistic for sample sizes m and n where $m \leq n$. The second portion of the table gives the value of $mnD_{m,n}$ such that its right-tail probability is the value given on the top row.

m	n	mnD	P	m	n	mnD	P	m	n	mnD	P
2	2	4	.333	3	6	18	.024	4	5	20	.016
2	3	6	.200			15	.095			16	.079
2	4	8	.133			12	.333			15	.143
2	5	10	.095	3	7	21	.017	4	6	24	.010
		8	.286			18	.067			20	.048
2	6	12	.071			15	.167			18	.095
		10	.214	3	8	24	.012			16	.181
2	7	14	.056			21	.048	4	7	28	.006
		12	.167			18	.121			24	.030
2	8	16	.044	3	9	27	.009			21	.067
		14	.133			24	.036			20	.121
2	9	18	.036			21	.091	4	8	32	.004
		16	.109			18	.236			28	.020
2	10	20	.030	3	10	30	.007			24	.085
		18	.091			27	.028			20	.222
		16	.182			24	.070	4	9	36	.003
2	11	22	.026			21	.140			32	.014
		20	.077	3	11	33	.005			28	.042
		18	.154			30	.022			27	.062
2	12	24	.022			27	.055			24	.115
		22	.066			24	.110	4	10	40	.002
		20	.132	3	12	36	.004			36	.010
3	3	9	.100			33	.018			32	.030
3	4	12	.057			30	.044			30	.046
		9	.229			27	.088			28	.084
3	5	15	.036			24	.189			26	.126
		12	.143	4	4	16	.029				
						12	.229				

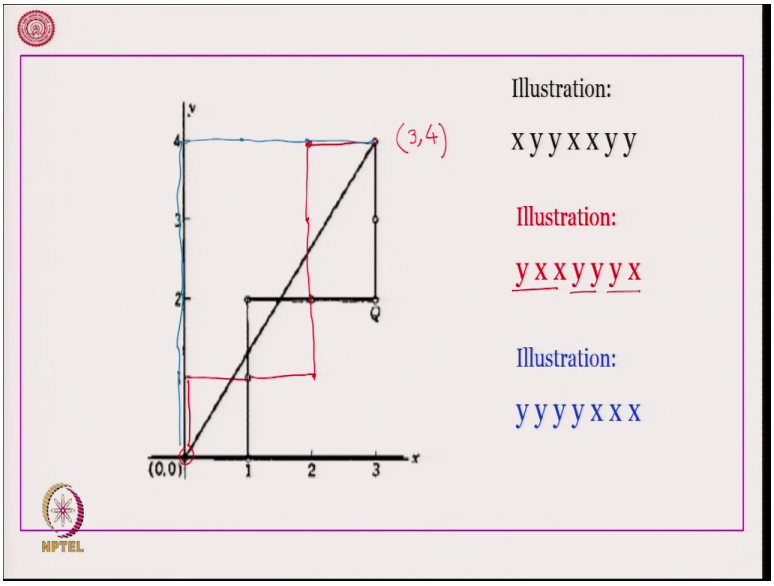
If m and n are not same then the values are tabulated, I think you cannot read it because it is too small. So, for example when m is equal to 4 and n is equal to 5 then the m,n D is given out for 20, 16, 15 etcetra and corresponding right tail probabilities are tabulated here.

(Refer Slide Time: 17:21)



So, how do we do? In order to compute probability Dm,n ≥ D under Ho where D is the observed value of Dm,n we proceed as follows exactly what we have done arrange the m + n observed values in increasing order. For each such arrangement can be depicted as a 2D graph from 0, 0 to m, n. Thus each such arrangement can be shown as a path starting from 0, 0 and going up to m, n. How do we show this path? So, for each of X observation we move to the right and for each Y observation we move two up like that after m plus n many movements we reach from 0, 0 to m, n. So, let me illustrate that.

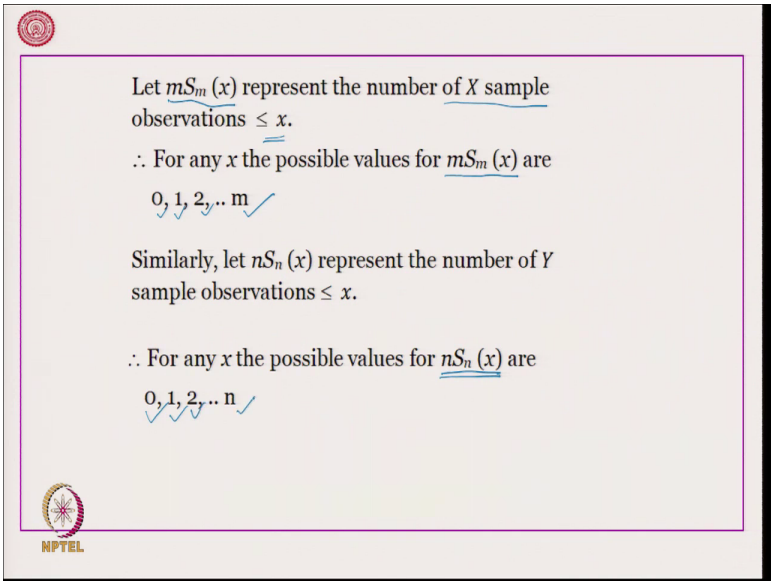
(Refer Slide Time: 18:21)



Consider m = 3 and n = 4. Therefore corresponding to each x observation we are going from current position to the right. Therefore, first we got an X, we come here then we got Y then another Y then X then X and then Y and Y. Let me illustrate again with some other example so consider y x x y y y x so what will be the path? The path is going to be like this from 0, 0 with y we shall go to 1 then 2 x's so we come here, from there we come here, then 3 y's 1 2 3 and then 1 x we reach m, n which is (3 , 4).

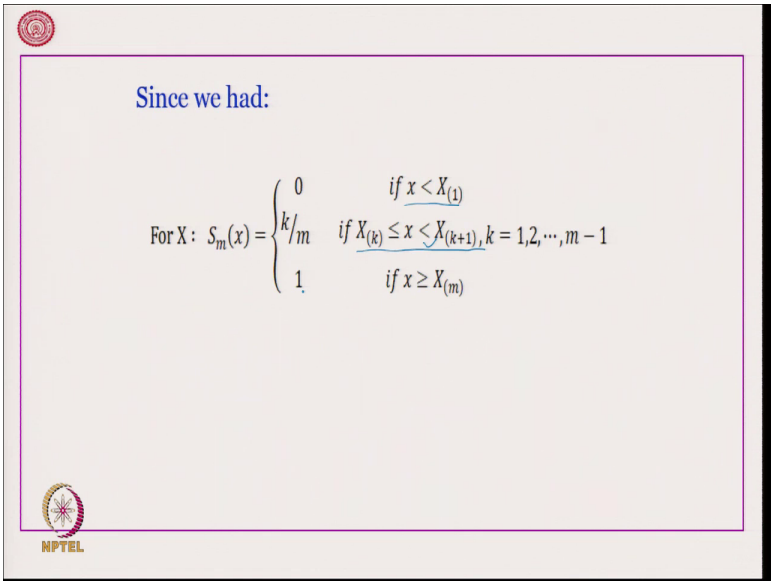
Let me give you one more illustration for y y y y x x x we shall go like this 1 2 3 4 x x x. Thus, you can understand that depending upon the arrangement of the ordered sample we shall get different paths. Each arrangement will give rise to a different path. What is the advantage? The advantage is that with the help of these paths will be trying to compute the probabilities of the maximum difference D. How to do that? That is the mathematics that is what I am going to explain now.

(Refer Slide Time: 20:31)



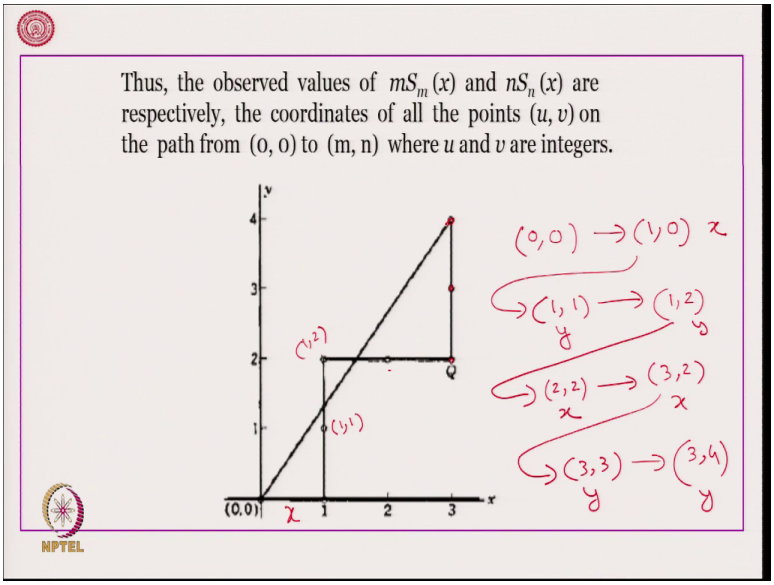
So, let mSm(x) represent the number of X sample observations ≤ x and therefore mSm(x) will take the values 0, 1, 2 up to m. In a similar way nSn(x) which is with respect to the Y observations will take the values 0, 1, 2 up to n. Why it is happening?

(Refer Slide Time: 21:03)



This is happening because of this definition that we have already seen that Sm(x) is 0 if X is less than the first order statistic it is k/m if x is between kth and k+1th order statistic when in this one is a strict inequality and it is 1 if x is ≥ xm.

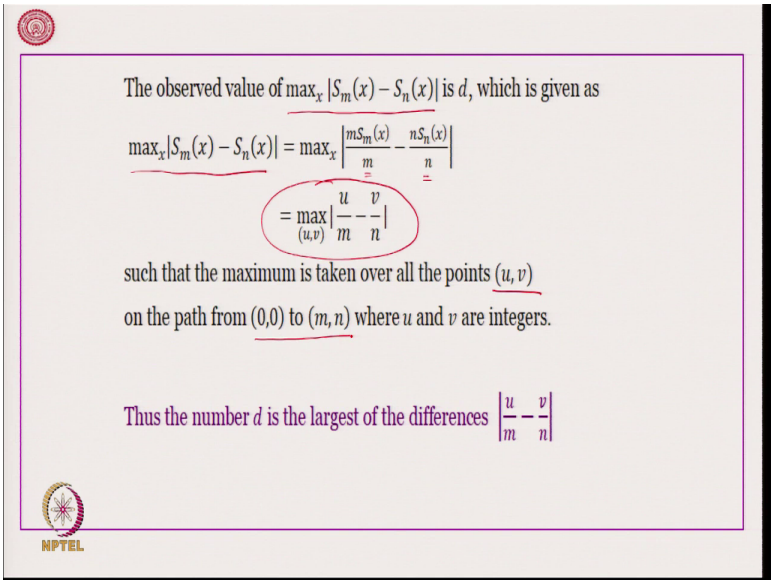
(Refer Slide Time: 21:31)



Thus, the observed values of mSm(x) and nSn(x) are effectively coordinates of all the points u v on the path from 0, 0 to m, n. Therefore, so for this example from 0, 0 we have gone to 1, 0 therefore, the first observation was an x. From 1, 0 we go to 1, 1 because that observation was an

y. From 1, 1 we go to 1, 2 because we got an y then from 1, 2 we go to 2, 2 because we got an x. From 2, 2 we go to 3, 2 because we got another x and from 3, 2 we go to 3, 3 because we got an y and from there we come to 3, 4 because of another y. So, that is the description of the path corresponding to this diagram.

(Refer Slide Time: 23:09)



Therefore, the observed value of maximum over x Sm(x) – Sn(x) is D which is given by

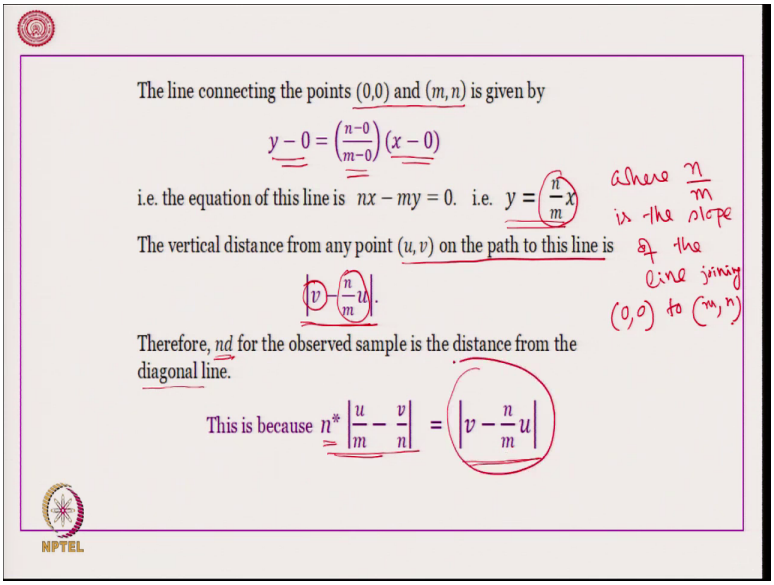
$$\max_x \left| \frac{mS_m(x)}{m} - \frac{nS_n(x)}{n} \right|$$

We have multiplied numerator and denominator. Now mSm X is U and nSn X is V, therefore it can be written as

$$\max_{(u,v)} \left| \frac{u}{m} - \frac{v}{n} \right|$$

So, this maximum has to be taken for all the points u v on the path from 0, 0 to m n. Thus the number D is the largest of the difference u / m - V / n. Why it is or how it is calculated? So, let us look at it from coordinate geometry.

(Refer Slide Time: 24:21)



The line joining the point (0, 0) to (m, n) can be written as

$$y-0=\left(\frac{n-0}{m-0}\right)(x-0)$$

Or in other words we can write it as

$$y=\frac{n}{m}x$$

where n by m is the slope of the line joining origin or (0, 0) to (m, n). Therefore, the vertical distance from any point u v on the path to this line where u and v are integers because either we are going to towards right one step if it is an x or going up one step if it is an y.

Therefore, the vertical distance of the point u v on the path from this diagonal line is going to be

$$\left|v-\frac{n}{m}u\right|.$$

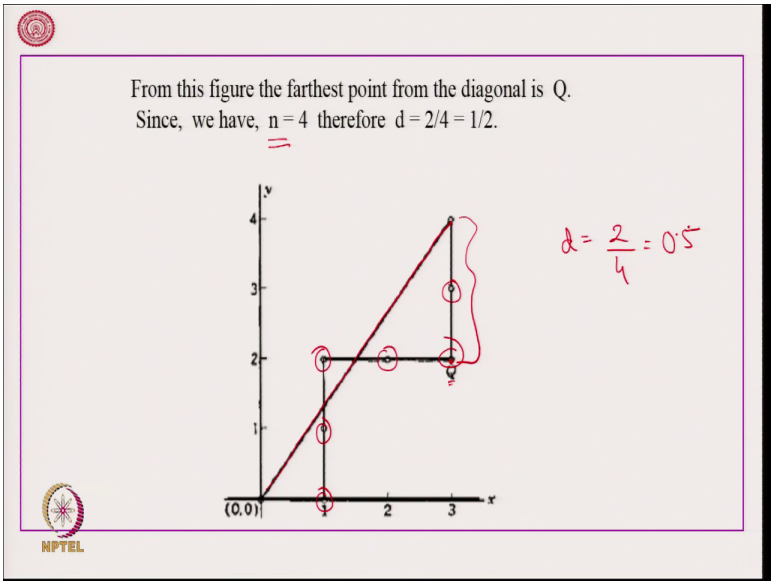
why? Because given u v on the point for the point u the corresponding value of y is equal to v but corresponding point on the line is going to be because the slope is n by m X, it is n by m U. Therefore, this is going to be the absolute difference. Now this suggests that n times D for the

observed sample is the distance of the point from the diagonal line since n times u by m minus v by n is equal to v minus n by m U in absolute terms. So, here we have seen

$$\left|\frac{u}{m} - \frac{v}{n}\right|$$

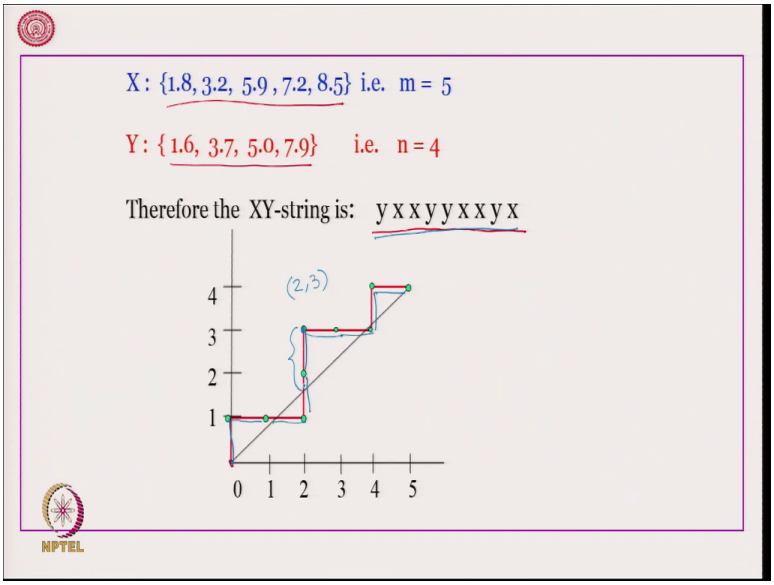
is the difference. So, if I multiply by that by n we get the distance from the diagonal line.

(Refer Slide Time: 26:43)



Say for example, if we look at it these are the points on the path and this is the diagonal line that we are talking about therefore, maximum distance is coming at this point which is Q and that distance is here we get two units but since n = 4 this will lead us the maximum difference with respect to the cumulative distribution is going to be D = 2 /4 is equal to 0.5.

(Refer Slide Time: 27:25)

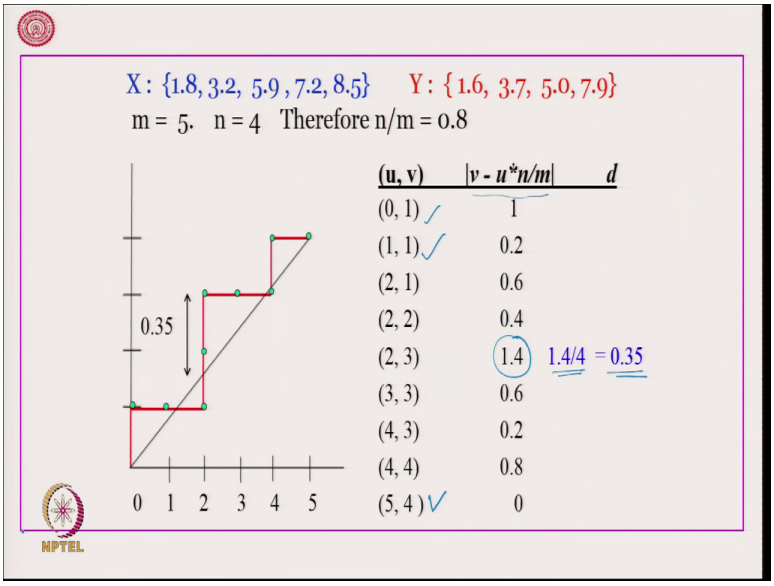


So, let me illustrate with another example. Suppose these are my x's these are my y's that we consider. Therefore, the movement is going to be in the sorted array y x x y y x x y x that is from 0,0 to 0,1 then 1, 1 then 2,1 then 2,2, then 2,3 then 3,3 then 4,3 then 4,4 and then 5,4. So, that is

the path that this sample allows us to traverse. Therefore what is going to be the maximum distance? If we look at it, it is coming out to be this.

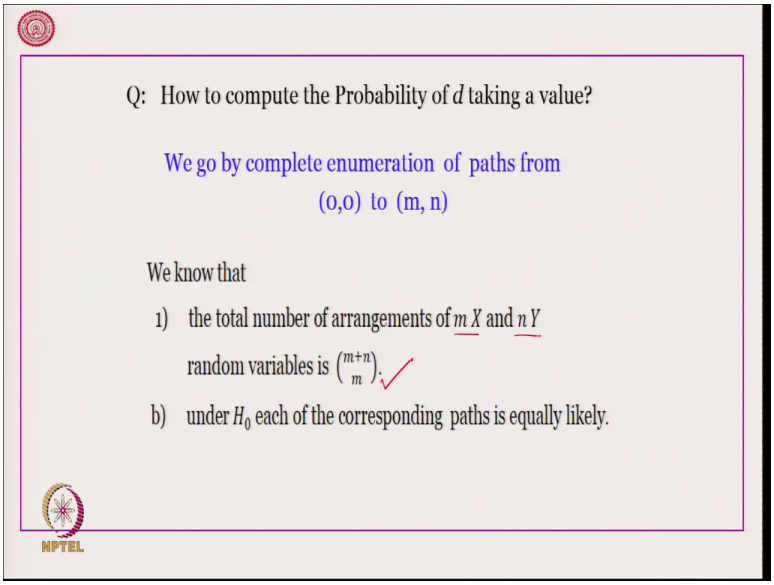
This is the maximum distance for this u v which is nothing but 2, 3 and the distance we can compute mathematically very easily. But let us go into this diagram. Here we find that this maximum distribution is 0.35.

(Refer Slide Time: 28:47)



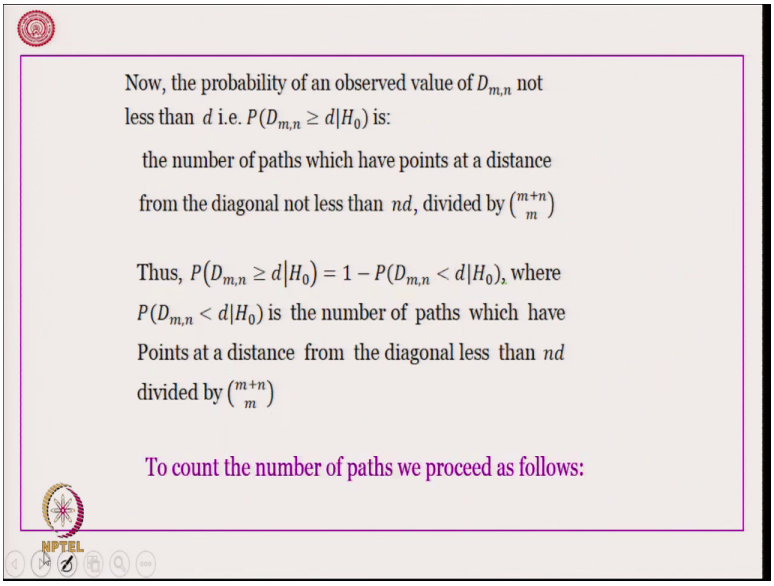
So. what we did corresponding to all the points that we have already marked 0,1, 1,1 up to 5,4 we have computed (v - u * n/m) and the maximum of them is coming out to be 1.4. Therefore, 1.4 divided by n which is 4, we get point 35. And if you remember in this graph we have already seen that the value was point 35. Thus, we transform the problem into the problem of obtaining the maximum difference in the cumulative functions into the problem of computing a path and from there we are looking at the maximum vertical distance from the diagonal.

(Refer Slide Time: 29:44)



So, the question is how to compute the probability of D taking a particular value? Here comes the path based interpretation very handy and in fact we shall compute the probability through complete enumeration of paths from 0, 0 to m, n. We know that the total number of arrangements of m X and n Y is equal to $^{m+n}C_m$. Now under Ho each of the corresponding path is equally likely.

(Refer Slide Time: 30:28)

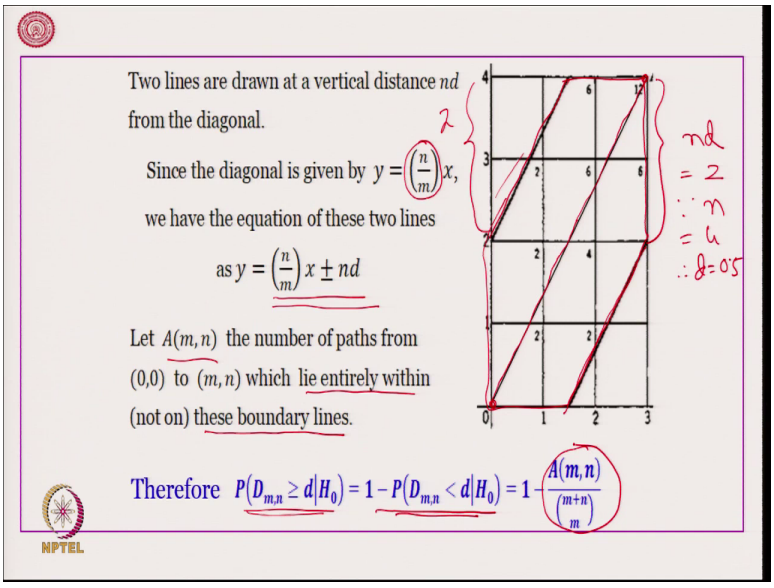


Therefore, the probability of an observed value of Dm,n not less than D that is probability D m, n greater than equal to D under Ho is the number of paths which have points at a distance from the

diagonal not less than n d divided by m plus n c m. Now we write it in a slightly different way what I do is the following. Probability D m, n greater than equal to D is equal to 1 minus probability D m. n less than D under Ho where D m,n less than D is the number of paths which have points at a distance from the diagonal less than nd.

And of course that will be divided by $^{m+n}C_m$. Please focus on this part it is less than and as we can see here it is strictly less than. Therefore, the points which are at a distance exactly nd will not to be taken under consideration for this calculation. Therefore, we now need to count the number of paths and we proceed as follows.

(Refer Slide Time: 32:01)



Two lines are drawn at a vertical distance n d from the diagonal. So, this is the diagonal and we are looking at a vertical distance here we have drawn two lines, one is this and the other in the this one and this vertical distance is equal to n d which we can see from this graph is equal to 2 since n is equal to 4 therefore, D is equal to 0.5 and you can see that here also the distance is 2. Therefore, equation of these two lines are

$$\left(\frac{n}{m}\right)x \pm nd$$

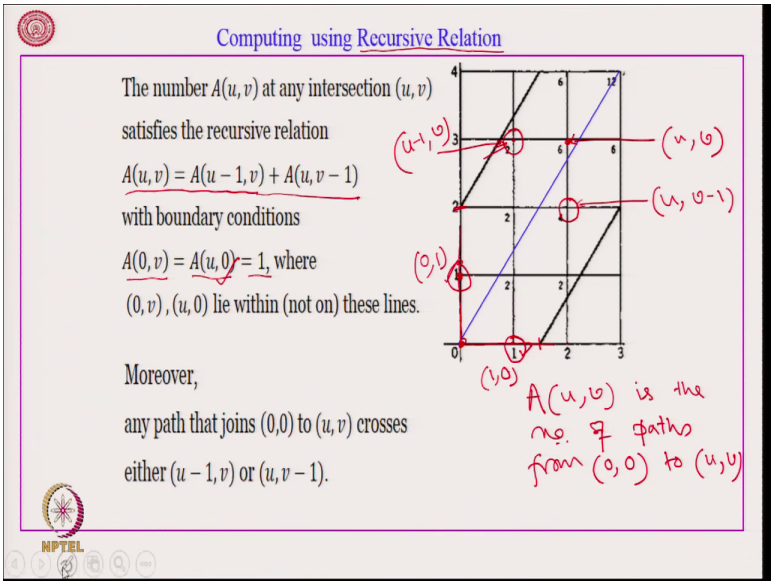
Because we know that the slope of the diagonal is n by m. Therefore, in counting the number of paths from 0, 0 to m, n we shall not consider any of the points which lie on this line or in other

words we shall count only those paths which are entirely in this region. Let A(m,n) be the number of paths from 0, 0 to m, n which lie entirely within this region that is within this boundary lines but not on this boundary lines.

And therefore, probability D m, n greater than equal to D will be one minus probability D m, n less than D and that is going to be

$$1 - \frac{A(m,n)}{\binom{m+n}{m}}$$

(Refer Slide Time: 34:10)



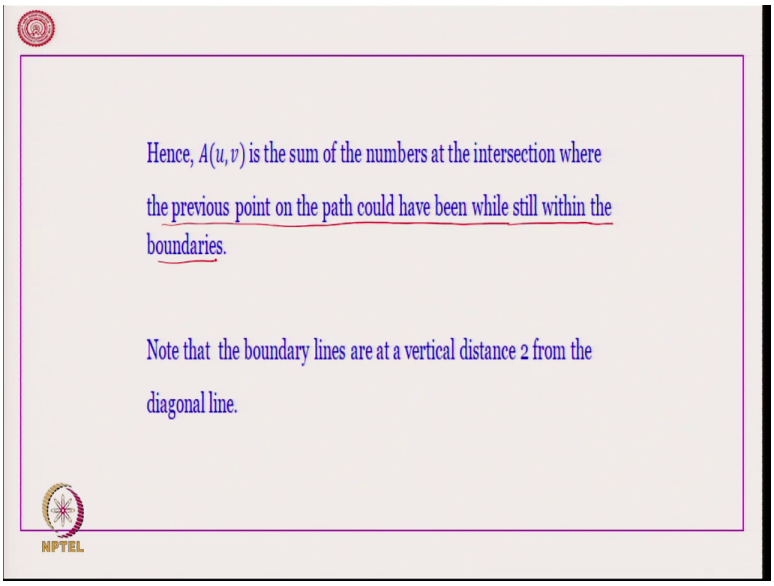
Now as I said we shall go for complete enumeration but we can make use of a recursive relation as follows. Let A uv is the number of paths from 0, 0 to u, v. Let us take an arbitrary point so let us consider this to be u, v. Now how many paths are there lying internally in this region from 0, 0 to u, v? Let us note that the last transition to come to u, v will be either from this point or from this point and this is nothing but (u, v - 1) The same value of x but height is less by 1 and this one is (u - 1 , v).

Therefore, we can say that A(u, v) is equal to A (u-1, v) + A(u, v -1) and what are the boundary conditions? The boundary conditions are A(0, v) = A(u, 0) = 1, A(0, v) means x value is 0 and y

value is v. So, so long we can go along the line each one of them we can access from 0, 0 but there is only one path that is this line. And similarly, when we move along the x axis there is only one path therefore A(u, 0) is also 1. Now in this case since the line comes here, the only two points that we are considering is this one which is (1, 0) and this one which is (0 ,1)..

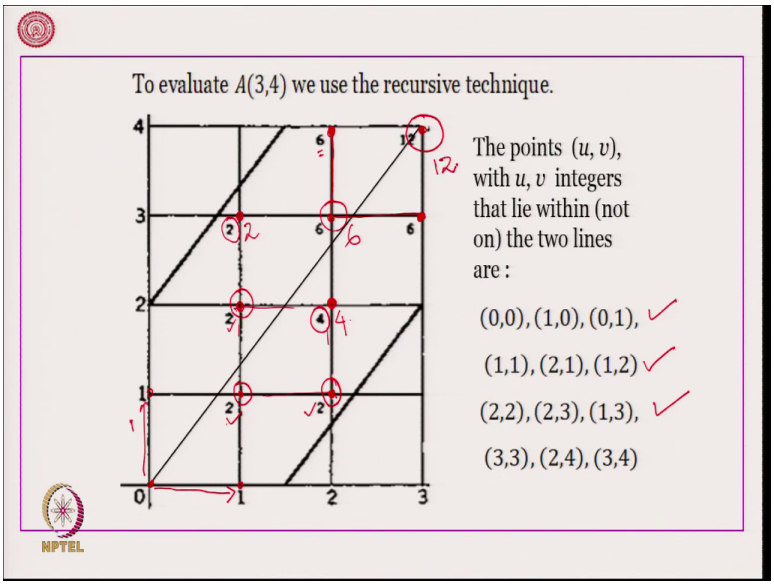
And the number of ways of reaching this point is 1 and reaching this point is also 1. Now with this boundary condition, we now start doing the recursion.

(Refer Slide Time: 37:07)



Therefore, A uv is the sum of the numbers at the intersection where the previous point on the path could have been while still within the boundaries. Therefore, effectively as we have discussed there are only two points from where we can reach a particular u v and as I said that we have to add the number of ways of reaching to those points in order to get the count of A(u, v).

(Refer Slide Time: 37:40)

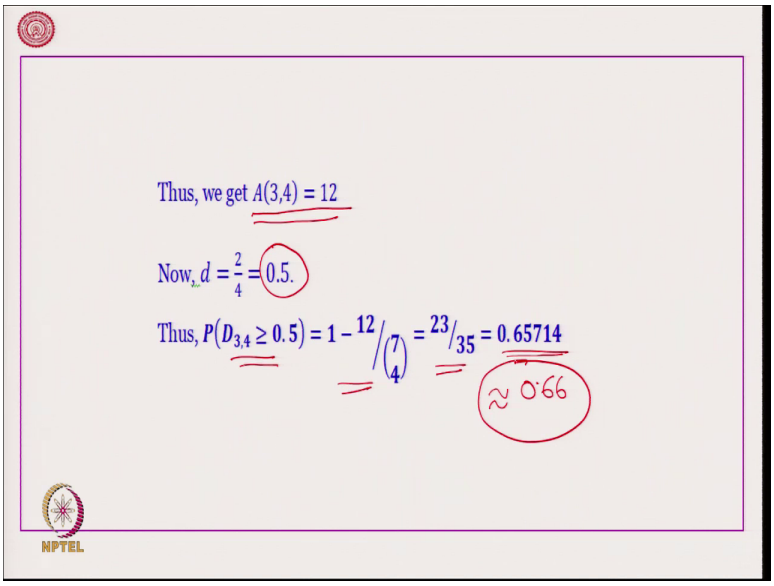


So, let us see which are the points that are completely within the region 0, 0, 1, 0, 0,1, 1, 1, 2, 1 and 1, 2, 2, 2, 2, 3 and 1, 3 and 3, 3, 2, 4 and finally 3, 4. So, we have to count the number of paths which are going through only these points and to reach 3, 4. So, here is only one and here is only one therefore, number of ways of reaching this point is 2. Now, from here we can go into this point in one step therefore number of paths to reach to this point is 2 and in a similar way number of paths to reach this point is also 2.

Therefore, we can come to this point by any of these two paths or any of these two paths and that gives us the count to be 4. Now we can reach to this point only in one way from this point. Therefore, number of paths to this point is still 2 and here we have already calculated this number to be 4. Therefore, total number of pass to this point is 6.

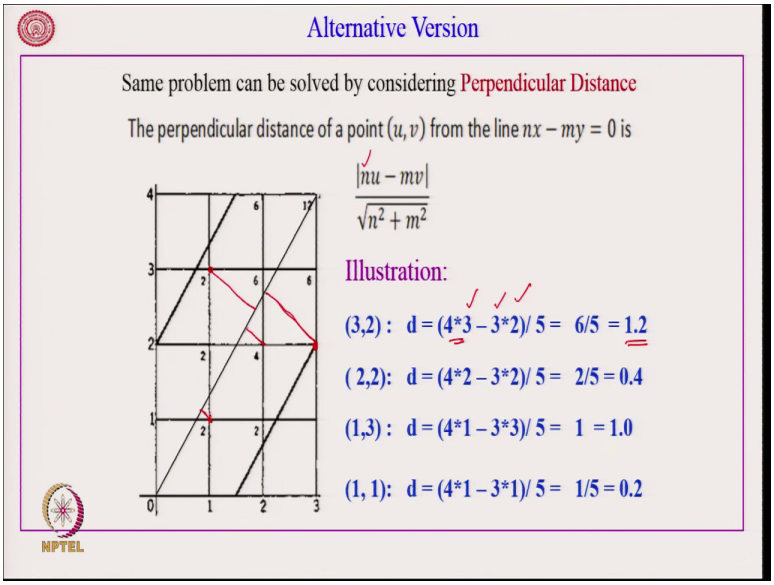
And since from here we can come to this point in only one way therefore the total number of paths to this point from the origin is 6 similar logic number of paths to come to this point is also 6 and therefore the total number of paths coming to this point being strictly within this region is 12.

(Refer Slide Time: 39:55)



Once we have computed that we find that A 34 is equal to 12 and note that here D was 0.5 as we have explained. Therefore, probability $D_{3,4} \geq$ to 0.5 1 - 12 / $^{7}C_{4}$ = to 23 / 35 is = 0.65714 which is very close to 0.66. Therefore, the probability that the maximum distance is going to be > 0.5 is 0.66. And in this way for different values of D we can compute their probabilities.

(Refer Slide Time: 40:52)



Now we present an alternative way of calculating that. In the previous one we have taken vertical distance. If you remember we were talking about this distance. Now in the alternative version we consider perpendicular distance that means, from any point (u, v) we draw a perpendicular to the diagonal and we look at its length that is the perpendicular distance from a point (u, v) to the diagonal line.

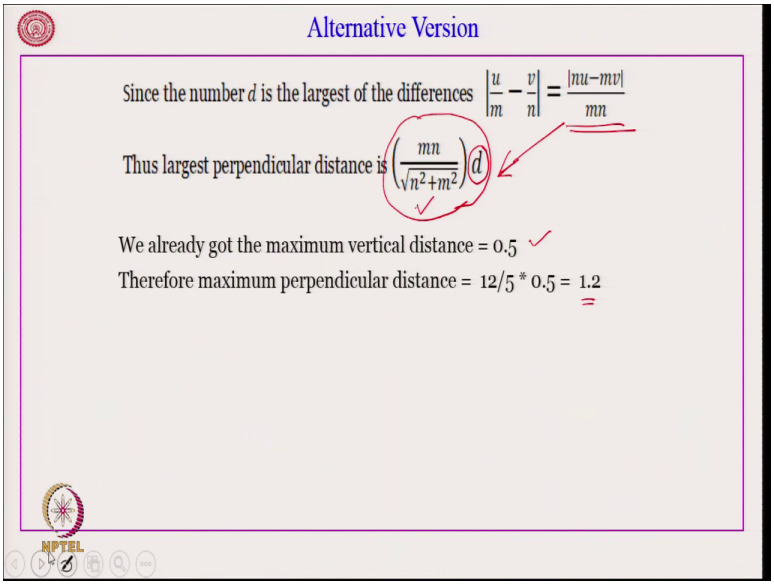
And we know for coordinate geometry that the perpendicular distance of a point (u, v) from the line nx - my = 0 is absolute value of

$$\frac{|nu - mv|}{\sqrt{n^2 + m^2}}$$

Say illustration, from the point (3, 2) the distance is going to be 4 star 3 because n = 4 and u = 3 - m = 3 and v = 2. Therefore this is coming out to be 6 /5 is equal to 1.2.

So, if this is the perpendicular distance, then that length is 1.2. In a similar way from 2, 2, this is the point 2, 2 and its distance from the diagonal line is 0.4 from 1, 3 which is this point this distance is going to be 1 and similarly from 1, 1 which is this point the perpendicular distance is going to be 0.2.

(Refer Slide Time: 42:59)

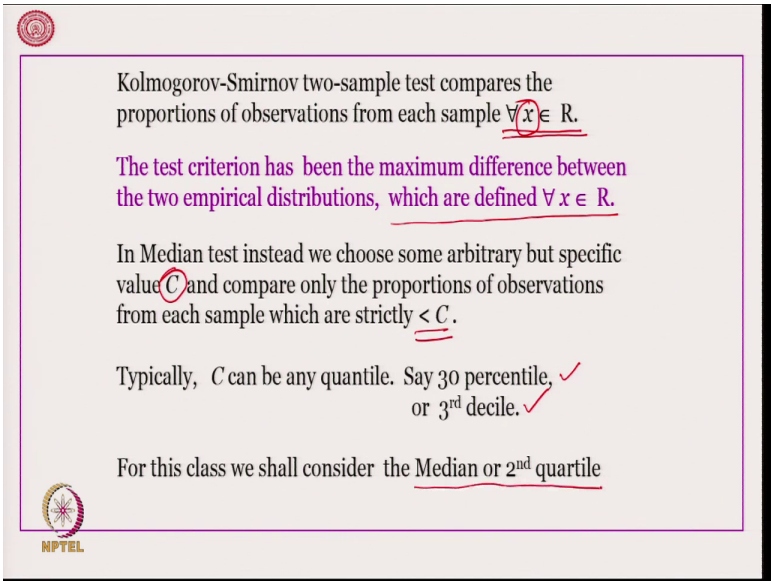


Since the number D is the largest of the difference in u minus m v upon m n that we have seen, we can now very easily correlate the perpendicular distance with this by writing that the largest perpendicular distance is going to be m n upon root over n square plus m square multiplied by d. Therefore, perpendicular distance is also coming out to be a multiple of D where the multiplier is m n upon root over n square plus m square.

We have already seen that the maximum vertical distance was 0.5 therefore, the maximum perpendicular distance is also coming out to be from here 1.2 and we have already obtained that 1.2 is coming from this point. Therefore, we can understand that that is going to be the point which is at a maximum distance D from the diagonal. Ok friends we now discuss another test for the same problem which is called the median test.

I shall give you an idea about the test we will not be dealing with it in great detail because it is less powerful than the Kolmogorov Smirnov two sample test. However, for the sake of completeness let us see what is the idea behind the Median test.

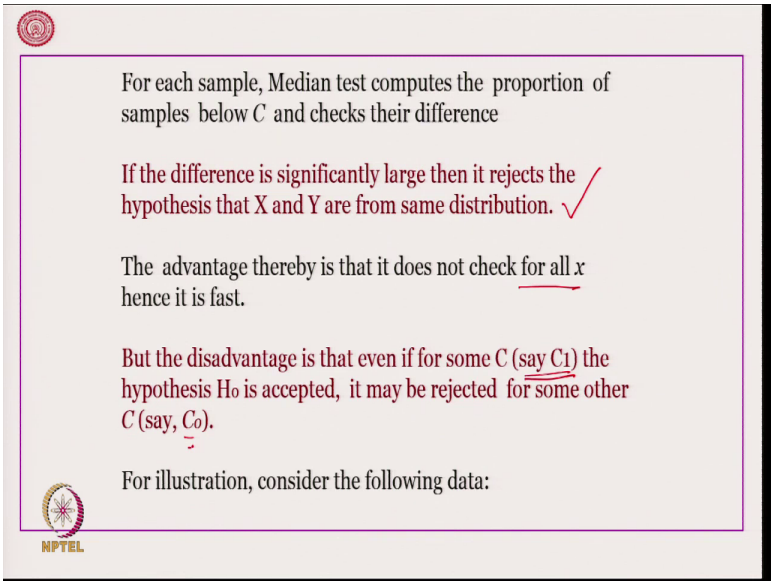
(Refer Slide Time: 44:41)



Now Kolmogorov Smirnov two sample test compares the proportions of observations from each sample for all x belonging to R. Hope you remember that we are going through the entire range and we are trying to check the difference for all x the difference between the cumulative relative frequency for x sample and y sample for a particular value x. And the test criterion has been that the maximum difference between the two empirical distributions which are defined $\forall$ x $\in$ R.

However, in median test instead of choosing the entire x $\in$ R, we choose some specific value C and compare only the proportions of observations from each sample which are $<$ C. Typically C can be any quantile, say for example it can be the thirtieth percentile or we can say that that is the third decile like that it can be any quintile. However for this class we shall consider only median or the second quartile.

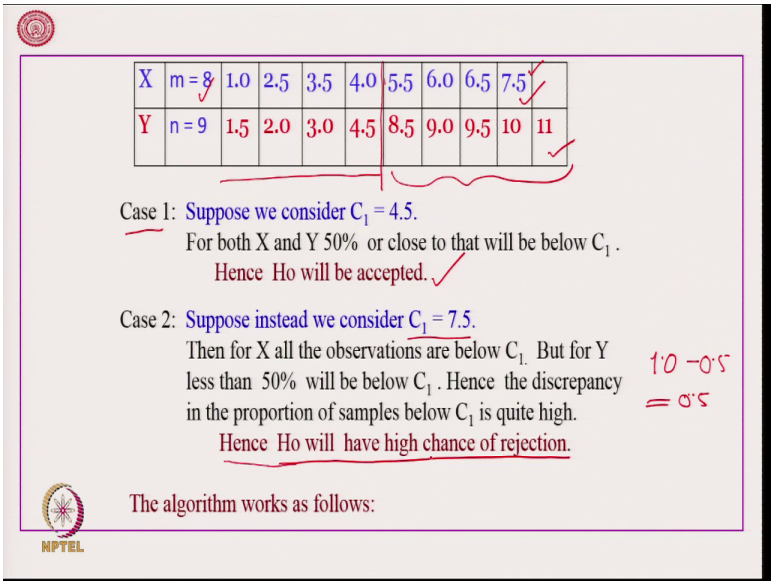
(Refer Slide Time: 46:15)



Now for each sample median test computes the proportion of samples below C and checks their difference. If the difference is significantly large, then it rejects the hypothesis that X and Y are from the same distribution.

The advantage there by is that it does not check for all x, hence it is fast. But the disadvantage is that even if for some C say let me call it C1 the hypothesis Ho is accepted, it may be rejected for some other C, say Co,. Let me illustrate it with the following data.

(Refer Slide Time: 47:05)

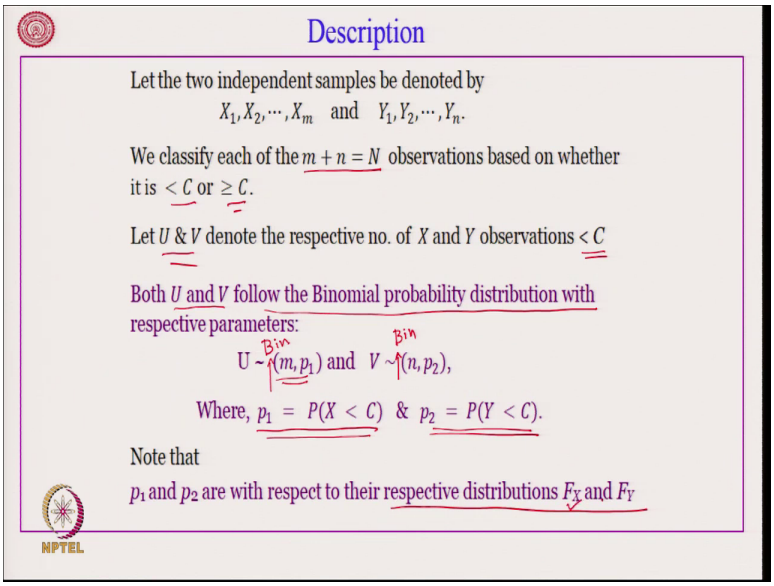


Consider 8 observations from X that is m is equal to 8 and these are the values 1.0, 2.5, 3.5 up to 7.5. And for Y there are 9 observations which are given there: 1.5, 2.0 up to 11. We want to check if they are coming from the same distribution. Now consider case 1where we have chosen C1 to be say 4.5. So, 4.5 which comes somewhere here we can see that 4 observations from X and 4 observations from Y are below the 4.5 or $\leq$ 4.5.

Therefore, there is hardly any difference in that proportions and therefore Ho is going to be accepted. Now consider instead that a value of C1 which is 7.5 then what is going to happen? We can see that all the X observations are below 7.5 and therefore, 100 percent of the observations are below 7.5. What about Y? With respect to y we can see that just 50 percent of them are below 7.5 because the remaining 5 observations are all bigger than equal to 7.50.

Therefore, that difference is going to be 1 minus 0.5 that is equal to 0.5 and therefore what I feel that Ho will have a high chance of rejection. Let us now see how the algorithm works.

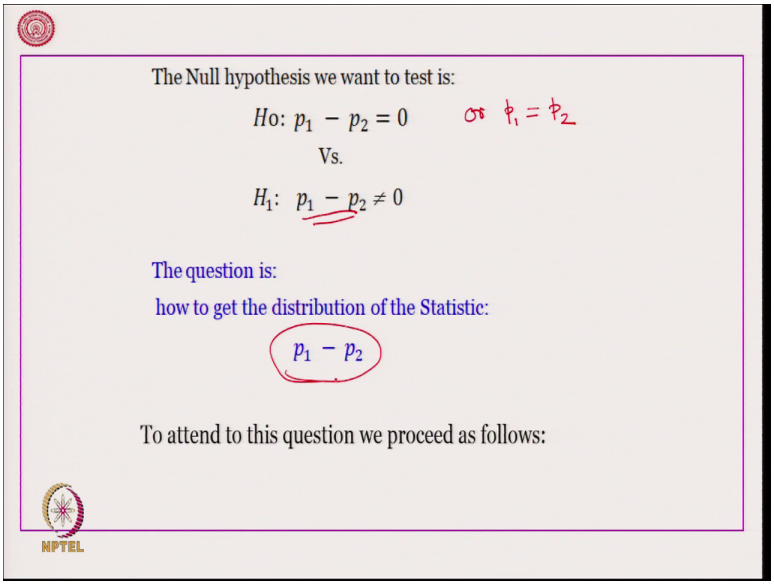
(Refer Slide Time: 49:09)



Let X1, X2, Xm and Y1, Y2, Yn be two independent samples. We classify each of the m plus n that is capital N many observations based on whether it is less than C or greater than equal to C. So, somewhere it has a similarity with the sign test that we have studied at the initial lectures of this course. So, let U and V denote the respective number of X and Y observations which are less than C.

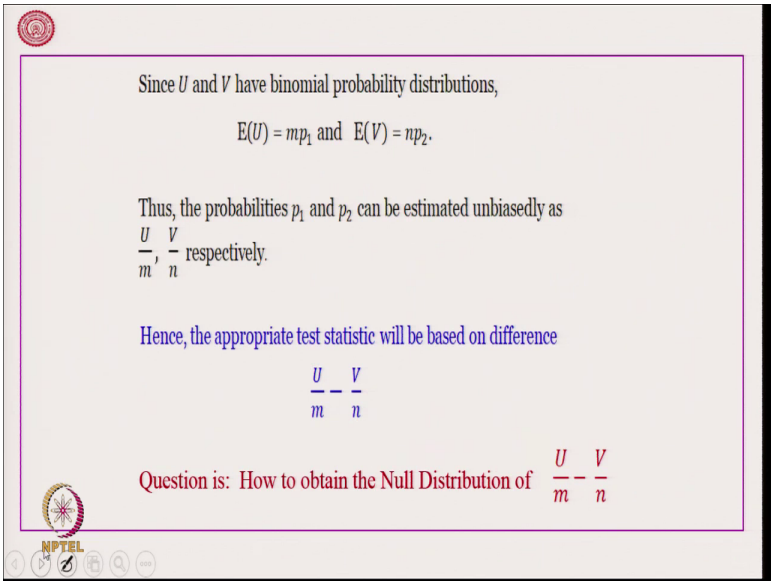
Therefore, both U and V follow the binomial probability distribution with the following parameters, U is equal to binomial m and p1 and v is binomial with n and p2. So, for your advantage let me write binomial here and binomial here so that you understand that these are binomial distributions where p1 is equal to probability X is less than c and p2 is equal to probability Y is less than C and this p1 and p2 are with respect to the respective distributions Fx and Fy. That is the cumulative distribution of X and Y.

(Refer Slide Time: 50:40)



Now we want to test if p1 minus p2 is equal to 0 that is p1 is equal to p2. That is the proportion as per the distribution these two should be equal versus p1 minus p2 is not equal to 0 that is they are unequal. Therefore the question is how to get the distribution of the statistic p1 minus p2. To attend to this question we proceed as follows.

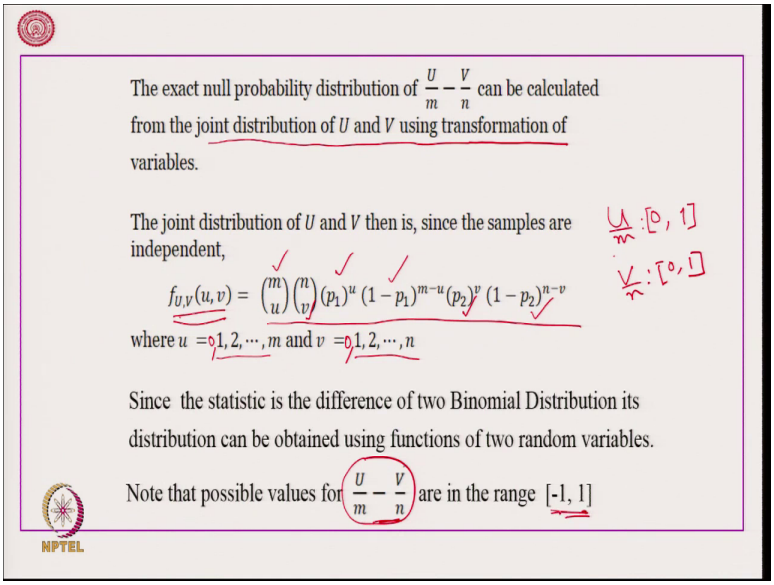
(Refer Slide Time: 51:14)



Since U and V have binomial probability distributions, expected value of u is equal to mp1 and expected value of v is equal to np2. Thus the probabilities p1 and p2 can be estimated unbiasedly as U by m and V by n where U is the number of observations coming from X which are less than C and V is the number of observations coming from Y which is less than C. Therefore, the appropriate test statistic will be U by m minus V by n. We need to test whether it is close to 0 or not.

Therefore, we shall reject hypothesis Ho if the obtained value of the statistic U by m minus V by n is greater than some Cα which is the critical value. But the question is how to obtain the null distribution of this statistic u/m - v/ m?

(Refer Slide Time: 52:54)



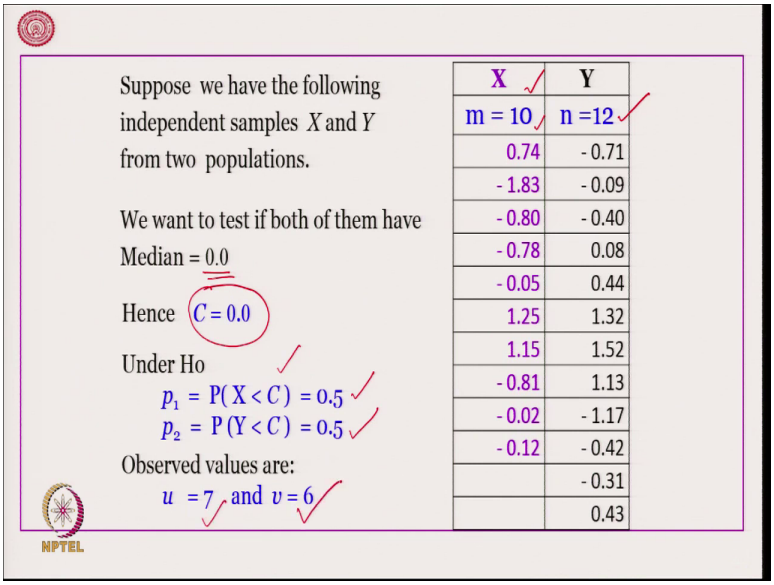
Now the exact null probability distribution of u/m – v/n can be calculated from the joint distribution of U and V using transformation of variables. I am not going into the detail of that because that is not within the scope of this course. But let us understand that the joint distribution of f(u, v) at a point small u small v can be written as the product of their individual mass functions because they are independent.

Therefore, it is coming out to be

$$\binom{m}{u}\binom{n}{v}(p_1)^u(1-p_1)^{m-u}(p_2)^v(1-p_2)^{n-v}$$

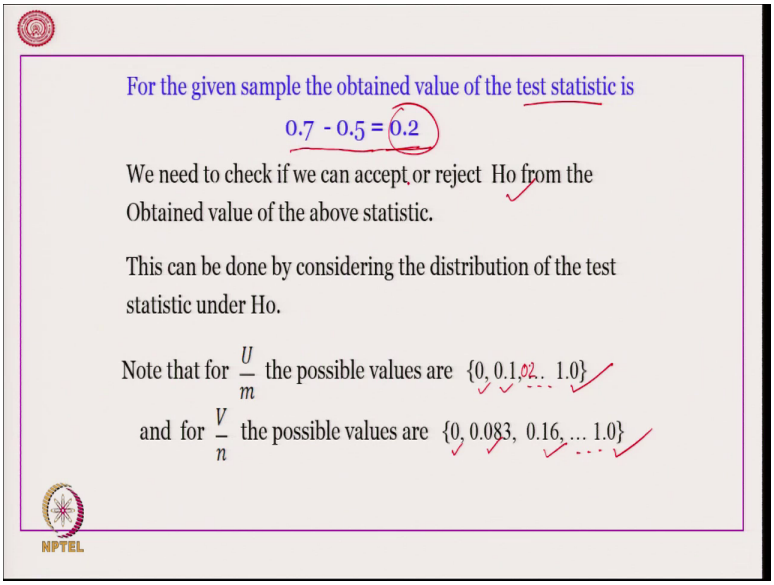
where u can be between 1 to m and v can be between 1 to n. And what is going to be the value of the statistic u /m - v / n? This is going to be between -1 to + 1 because u / m can take values 0 to 1. This will be 0 corresponding to 0 and this will be 1 corresponding to m in fact they can take the value 0 also. So, let me write that. Therefore, u by m can take the value 0 to 1 and similarly v by n can also take the value between 0 to 1 and therefore this statistic can be in the range -1 to +1.

(Refer Slide Time: 54:46)



So, let me illustrate with an example suppose we have the following independent samples of X and Y, m is 10 that means there are 10 observations from X, , and n is 12 that means there are 12 observations from Y and we are testing if the median is 0.0 that means the value of C that we are taking is 0 and under Ho p1 which is the probability of X is less than C that has to be half or 0.5. Similarly, p2 also has to be 0.5 but the obtained statistic u is equal to 7 and v is equal to 6.

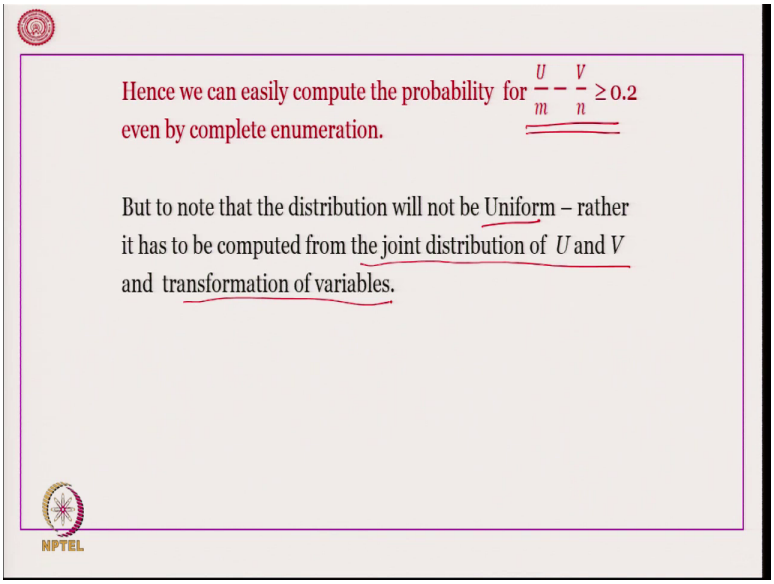
(Refer Slide Time: 55:35)



Therefore, for the given sample the obtained value of the statistic is 0.7 minus 0.5 that is 0.2. Therefore, we need to check if we can accept or reject Ho from the obtained value of the statistic

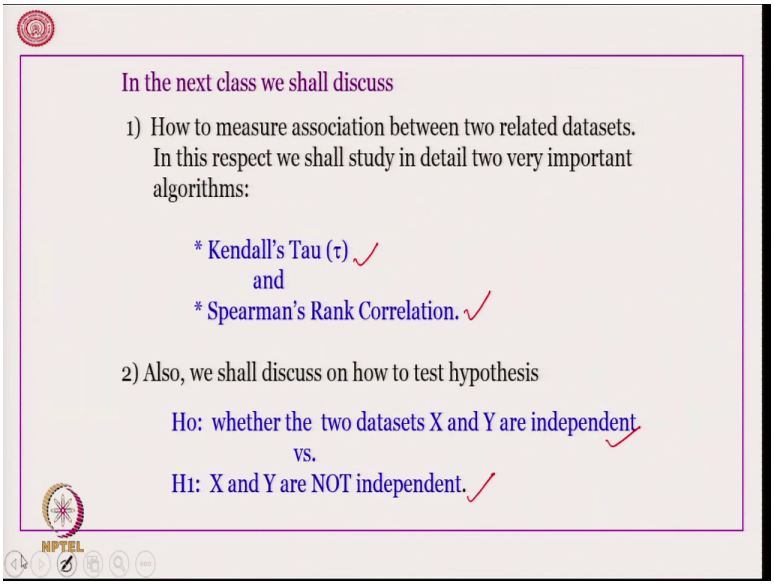
as I said before this can be done by considering the distribution of the test statistic under Ho. In this particular case U by m will take values 0, 0 point 1, 0.2 like that up to 1. And V by n will take values 0, 0.083, 0.16 like that up to 1 and therefore we can calculate the m n many pairs of values between minus 1 to plus 1. And from there we can count the probability of 0.2.

(Refer Slide Time: 56:33)



That is what we have written that we can easily compute the probability for U by m minus V by n is greater than 0.2. But note that the distribution will not be uniform. Because they are coming from binomial distributions therefore they have to be computed from the joint distribution of U and V and transformation of variables thus, I have already indicated. So, as I have said that we are not going to pursue it any further because it is a weaker test in comparison with two sample Kolmogorov Smirnov test. However, we can see that this can give us a quick decision with respect to the null hypothesis whether they are coming from the same distribution.

(Refer Slide Time: 57:23)



Okay friends I stop here today in the next class we shall discuss how to measure association between two related data sets. In this respect we shall study in detail two very important algorithms namely Kendall's Tau and Spearman's rank correlation and also we shall discuss on how to test hypothesis that is Ho whether the two delta says X and Y are independent versus X and Y are not independent. Ok friends that is for the next class for the time being thank you so much.