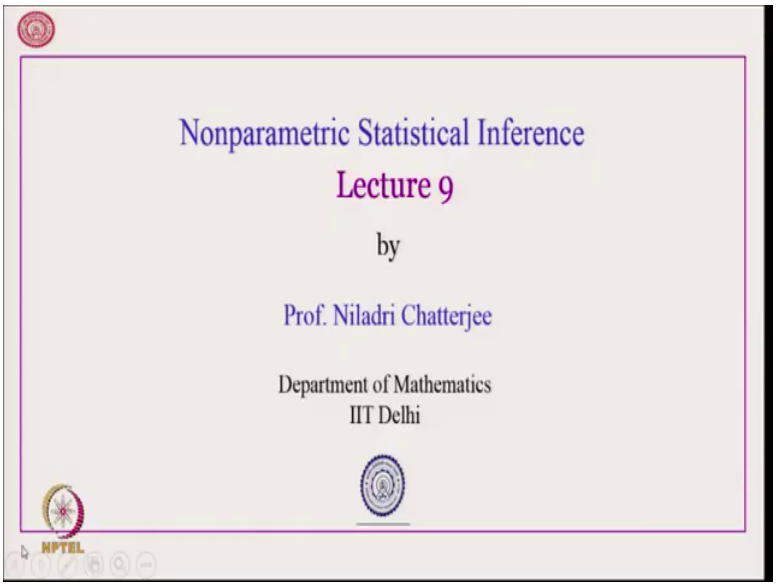
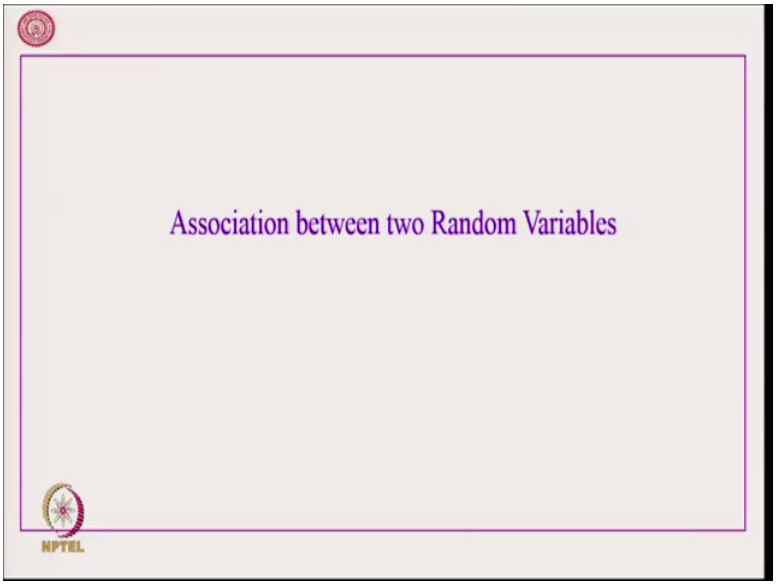
Nonparametric Statistical Inference
Professor Niladri Chatterjee.
Department of Mathematics
Indian Institute of Technology, Delhi
Lecture - 9
Nonparametric Statistical Inference

(Refer Slide Time: 0:18)



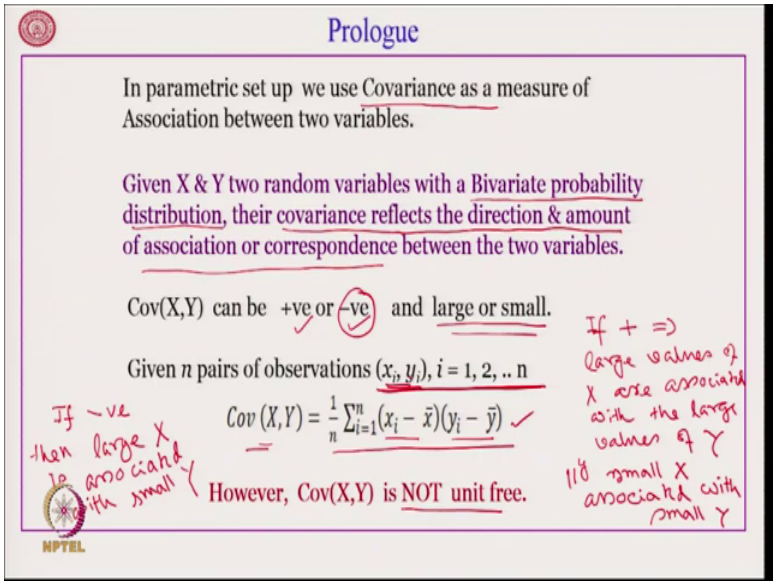
Welcome students to the MOOC's series of lectures on nonparametric statistical inference, this is lecture number 9.

(Refer Slide Time: 0:40)



As I said at the end of the last class that in this class we shall study association between two random variables.

(Refer Slide Time: 0:48)



I hope all of You know how we study the association of two random variables in a parametric setup and the corresponding measure is covariance. So, the situation is like this, given X and Y two random variables, we take bivariate probability distribution, this is very important because if X and Y are independent, then they do not have any covariance, that value is going to be 0. But if there is a joint distribution, which we call bivariate probability distribution, then their covariance reflects the direction and amount of association or correspondence between the two random variables.

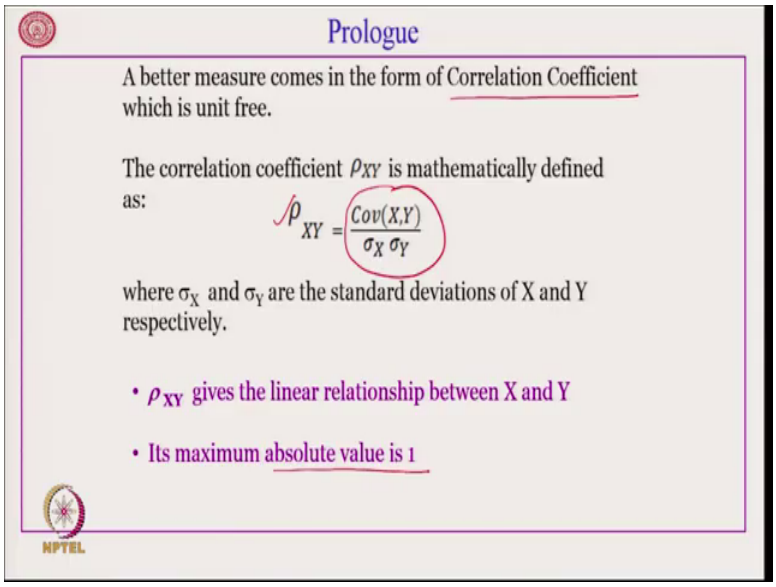
So, we know that covariance of X, Y can be positive or negative, that is direction is plus or minus. If it is plus, implies that large values of X are associated with the large values of Y and similarly, for small values, that is small x is associated with a small y. On the other hand, if it is negative, then we know that if the covariance is negative, then large value of X is associated with small values of Y. And similarly, large values of Y are associated with small values of X and it can be large or small, depending upon how much space is there for X and Y.

And in particular, if X1, Y1, X2, Y2, Xn, Yn is n pairs of observation, the covariance of X, Y is calculated using this following formula. It is

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Therefore, depending upon these magnitudes, the overall value of the covariance will be determined and note that it is not unit free. That means that suppose X and Y are height and weight of boys, then if height is measured in meters and weight is measured in kilograms, then these values will be smaller in comparison with when height is measured in foot and weight is measured in say pounds. And therefore, the magnitude of the covariance may change.

(Refer Slide Time: 4:33)

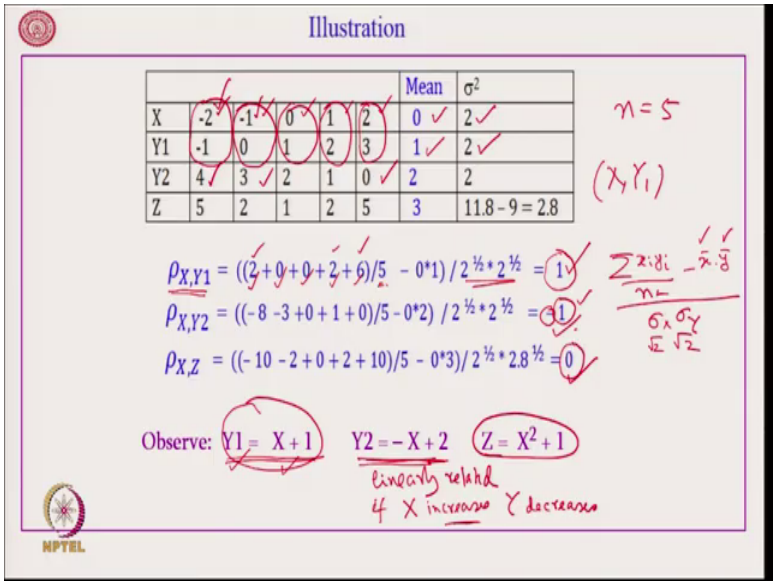


Hence, a better measure comes in the form of correlation coefficient, which is unit free and it is typically denoted by the symbol ρ and $\rho_{X,Y}$ is defined as

$$\frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

where σ_X and σ_Y are the standard deviations for X and Y. Note that the $\rho_{X,Y}$ gives a measure of the linear relationship between X and Y and like covariance, it can be plus or minus, but the maximum absolute value is 1.

(Refer Slide Time: 5:19)



So, let me illustrate. Consider 5 points, that is n is equal to 5 and we have X1, X2, X3, X4 and X5, we are checking its association with three other random variables Y1, Y2, and Z. So, let us consider X and Y1 that means

You are looking at these pairs of observations (-2 , -1), (-1, 0), (0, 1), (1, 2) and (2,3). Therefore, what is going to be the correlation between X and Y1, we look at this table the mean of X is 0 and its variance is 2 the mean of Y 1 is 1 and variance is 2.

Therefore, $\rho_{X,Y1}$ is $(\Sigma Xi\ Yi)/n - \overline{X} * \overline{Y}$, whole divided by $\sigma X\ \sigma Y$. And therefore, it is coming out to be - 2 into -1 which is 2, - 1 into 0 which is 0, 0 multiplied by 1 which is 0, 1 multiplied by 2 which is 2 and 2 multiplied by 3 which is 6, that divided by 5 minus $\overline{X}$ is equal to 0, $\overline{Y}$ is equal to 1 and their standard deviations are $\sqrt{2}$, $\sqrt{2}$. Therefore, together we are getting the value is coming out to be 1. Because 2 + 2 is 4, 4 + 6 is 10, 10 / 5 is 2, therefore, 2 / 2 = 1.

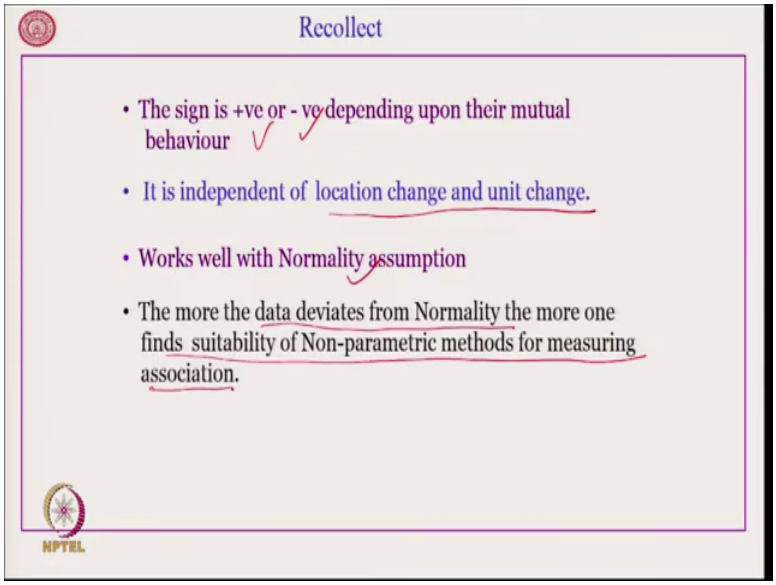
In a similar way, $\rho_{X,Y2}$ coming out to be -1 and $\rho_{X,Z}$ is coming out to be 0. Now, let us look into that data, we can see that they have the following relationship, Y1 is equal to X plus 1, that means Y and X has a positive relationship as X increases Y1 increases as a decreases Y1 decreases and they are perfectly linear. Therefore, the correlation is coming out to be positive and because it is perfectly linear, its value is 1.

On the other hand, if we consider Y2, it is nothing but -X + 2, therefore for -2 it is coming out to be 4, for -1 it is coming out to be 3 and for 2 it is coming out to be 0. What is there in this data? Firstly, they are linear or linearly related and secondly if X increases then Y decreases.

Because they are perfectly linear, the magnitude of the correlation is 1, but because increasing X means decreasing Y, we get a negative sign there that is it is negatively correlated, that is very clear from this value.

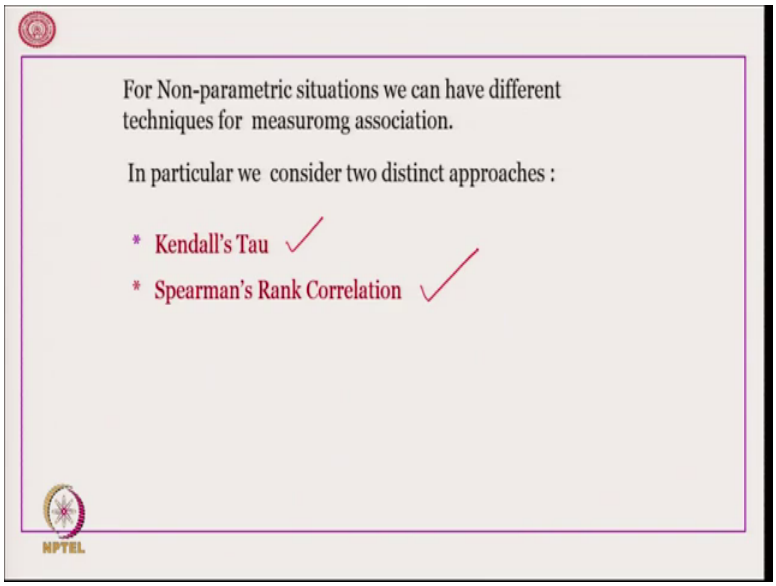
Now, if we consider the correlation of X and Z we can see that their relationship is Z is equal to $X^2 + 1$. Therefore, there is no linearity between them and therefore, we can see that the correlation between them is 0. So, this is some intuition that we need to know when we talk about the association of 2 random variables.

(Refer Slide Time: 9:52)



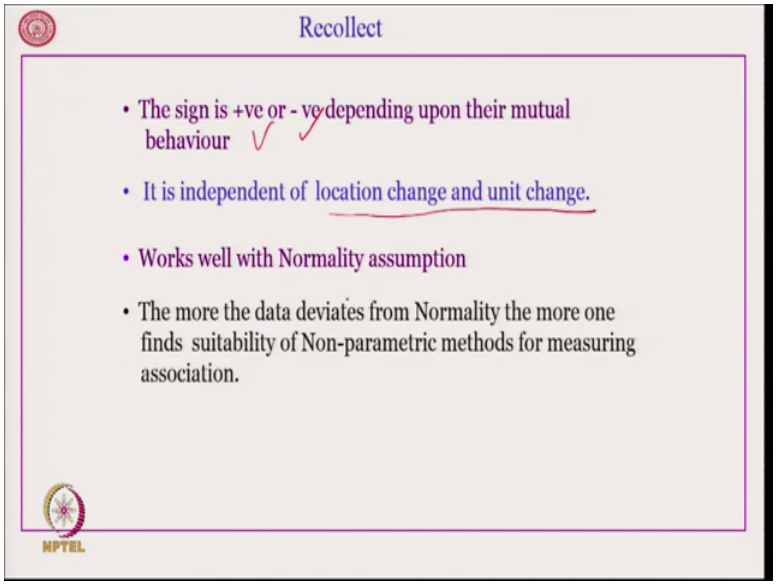
Therefore, let us recollect that the sign is positive or negative depending upon their mutual behavior, but it is independent of location change and unit change, it works well with normality assumption. As we said before, that the parametric statistical inference is based on a normality assumption. Therefore, the correlation coefficient that we discussed works well under normality assumption. However, the more the data deviates from normality, the more one finds the suitability of nonparametric methods for measuring Association. That is, if the data is non normal, then rather we should look at non parametric way of measuring their associations.

(Refer Slide Time: 10:48)



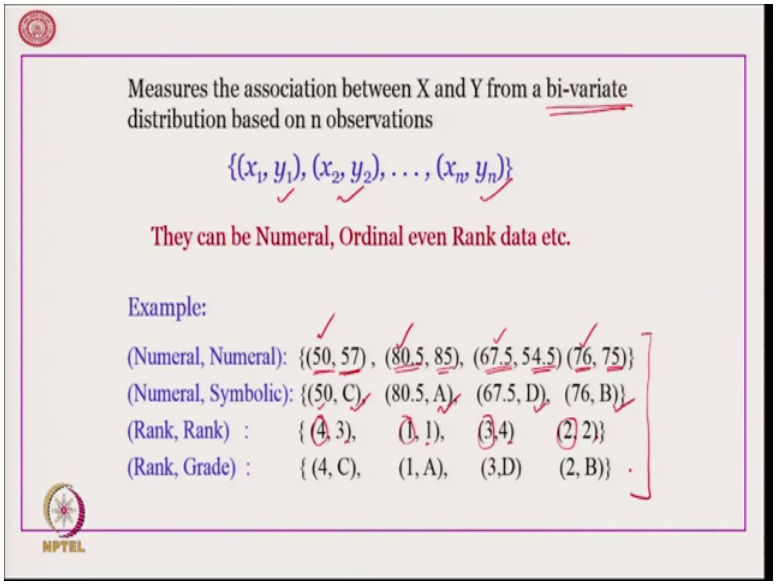
So, in this class, we should study two important non parametric tests for association, one is called Kendall's Tau and Spearman's Rank correlation.

(Refer Slide Time: 11:05)



So, let us first look into Kendall's Tau, often it is written like this from the Greek alphabet τ.

(Refer Slide Time: 11:10)



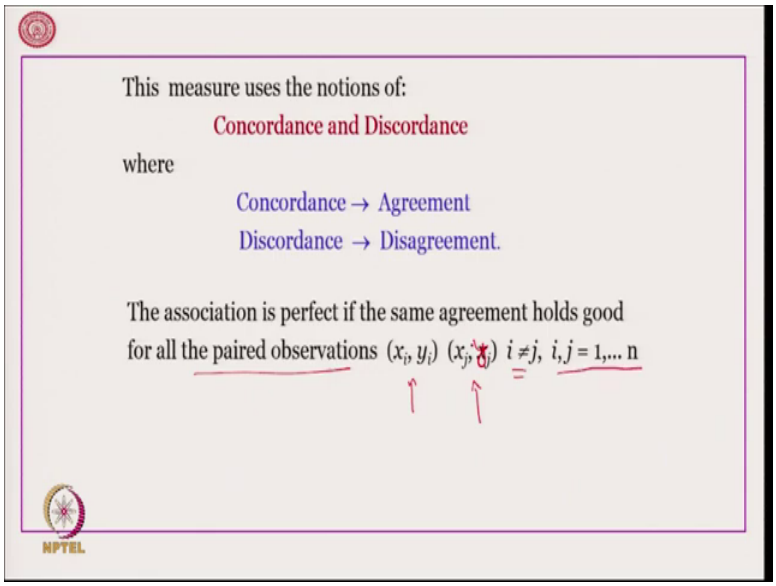
So, what it is? It measures the association between X and Y from a bivariate distribution based on n observations. So, let X1, Y1, X2, Y2, up to Xn, Yn be the n pairs of observation taken from the bivariate distribution. They can be numeral, they can be ordinal or they can be even rank data.

Say for example, we are looking at marks of 4 students in 2 subjects. So, one way of looking at them is both of them are numerals. So, first student got 50 in one subject and 57 in the other subject, second subject got 80.5 and 85. Similarly, 67.5 and 54.5 for the third student and these 2 are for the fourth student.

Again, one of them can be numeral, but the other can be symbolic. Many institutions they give grades instead of marks and suppose the grades obtained by that four students in the second subject are C, A, D and B respectively. One can even represent that data using ranks, that means with respect to the first subject, the ranks of the people for students are fourth, first, third, and second, as we can see that this is the highest one, this is the second highest one, this is the third highest and this is the fourth highest.

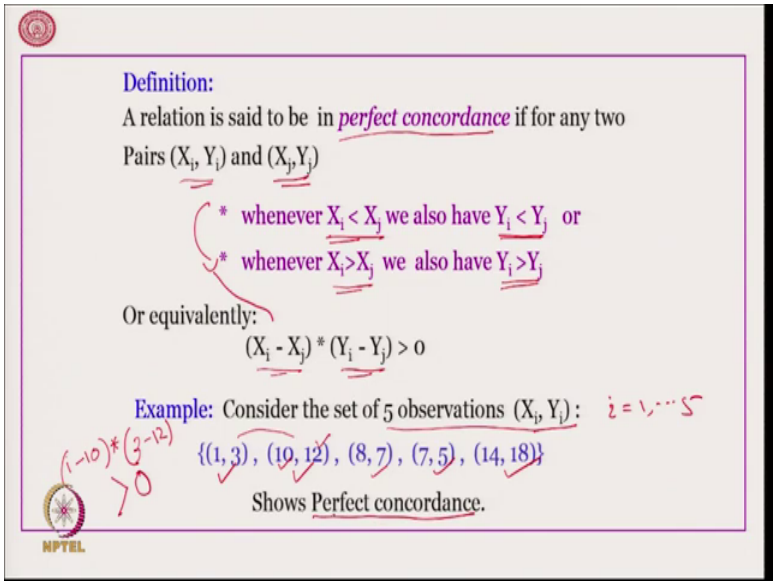
And similarly, the ranks for the second subject they have given below. Or the data can be something where one is rank, other one is symbolic. In short, we want to say that the data can be in many different forms, they need not be numeric. And therefore, in such cases, the correlation coefficient that we discussed cannot be computed, because they need arithmetic manipulation of data with the help of the formula that we have given earlier.

(Refer Slide Time: 13:35)



Then the question comes how to measure their association. In Kendall's Tau, it uses 2 notions, one is called concordance or that is called discordance, where concordance means agreement and discordance means disagreement. So, what does it mean? It means that the association is perfect for the ith and jth data point namely (xi yi) and (xj yj) for all paired observations when i ≠j and i, j is equal to from 1 to n.

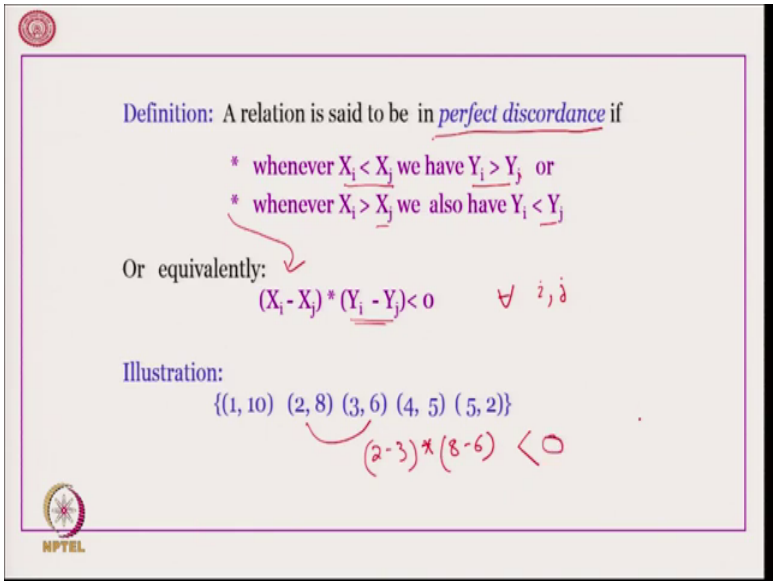
(Refer Slide Time: 14:25)



So, by definition, a relation is said to be in perfect concordance if for any two pairs Xi, Yi and Xj, Yj, whenever, Xi < Xj, we also have Yi < Y j or whenever, Xi > Xj, we also have Yi > Yj. Together we can write that Xi - X j multiplied by Yi - Yj > 0. For example, consider the set of 5 observations, Xi Yi, i is equal to 1 to 5, (1, 3), (10, 12), (8, 7), (7, 5), (14 and 18). Then we can see that it shows perfect concordance.

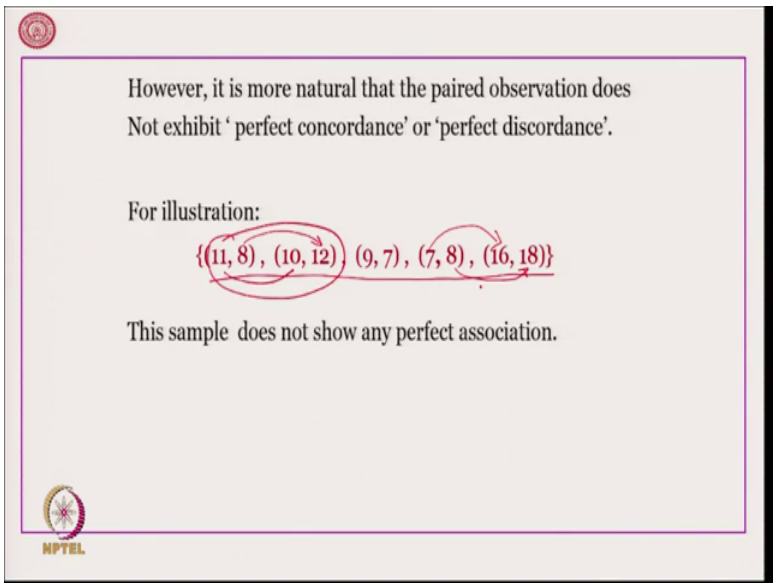
For example, if we consider these 2 pairs, here the first observation the value is increased and the second observation the value is also increased. Therefore, we are looking at 1 minus 10 multiplied by 3 minus 12. And we can see that this is going to be greater than 0. Similarly, we can check with respect to all the possible pairs, that is ${}^5C_2$ many pairs and we can see that for all of them, this value is coming out to be greater than 0 and therefore, this is a perfect concordance.

(Refer Slide Time: 16:11)



At the same time, a relation is said to be perfect discordance if whenever Xi < Xj, we have Yi > Yj, that is, if the value of X is increased, the value of Y is decreased or if value of X is decreased, then value of Y is increased. In other words, these two together we can write as Xi - Xj multiplied by Yi - Yj has to be less than 0 for all i, j. For example, let us choose arbitrarily these two pairs, therefore, we get 2 - 3 multiplied by 8 - 6 and we know that this is coming out to be less than 0. The same one can compute with respect to all the $^5C_2$ many pairs.

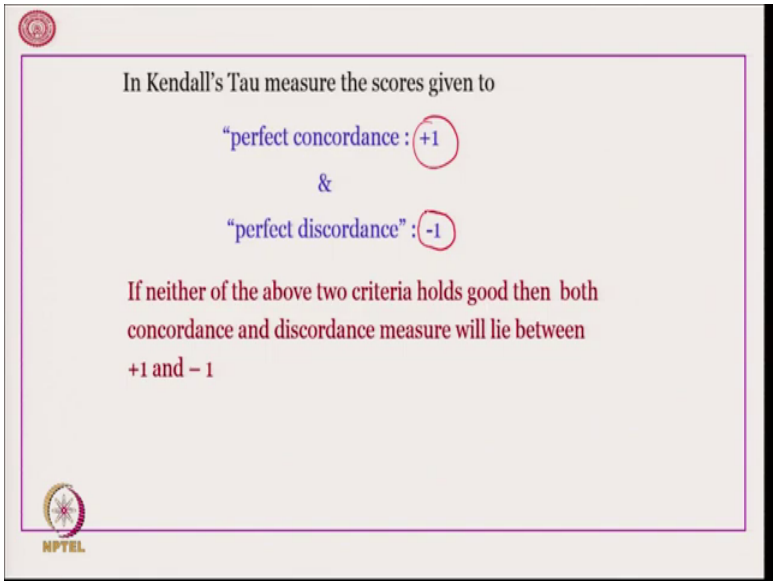
(Refer Slide Time: 17:20)



However, it is more natural that the paired observation does not exhibit perfect concordance or perfect discordance with respect to all the pairs. Again, for illustration consider these 5

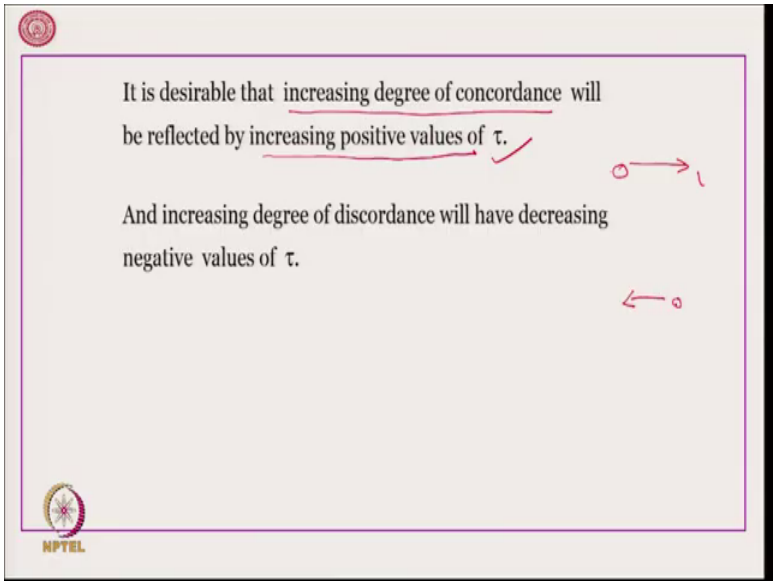
tuples. If we consider these two, (11, 8) and (10, 12), we can see that value of X is decreased, but value of Y is increased therefore they are in discordance. However, if we consider these two pairs, the value of X is increased and value of Y is also increased. Therefore, there is a concordance between these two pairs. Therefore, this sample does not show any perfect association. And that is very common, that is what in general we would expect.

(Refer Slide Time: 18:30)



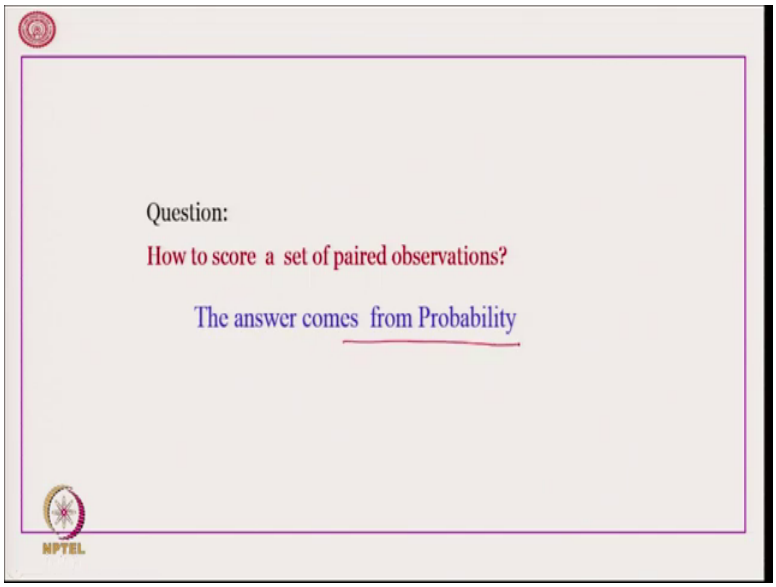
And therefore, we need to give a measure which resembles a degree of association between them. So, in Kendall's Tau, the scores are given as follows. If it is perfectly concordant pair or they are in perfect concordance, then the value given will be plus 1, if perfect discordance then value will be given minus 1, if neither of the above two criteria holds good, then both concordance and discordant measure will lie between plus 1 and minus 1. Thus, we can see a similarity with respect to correlation coefficient.

(Refer Slide Time: 19:13)



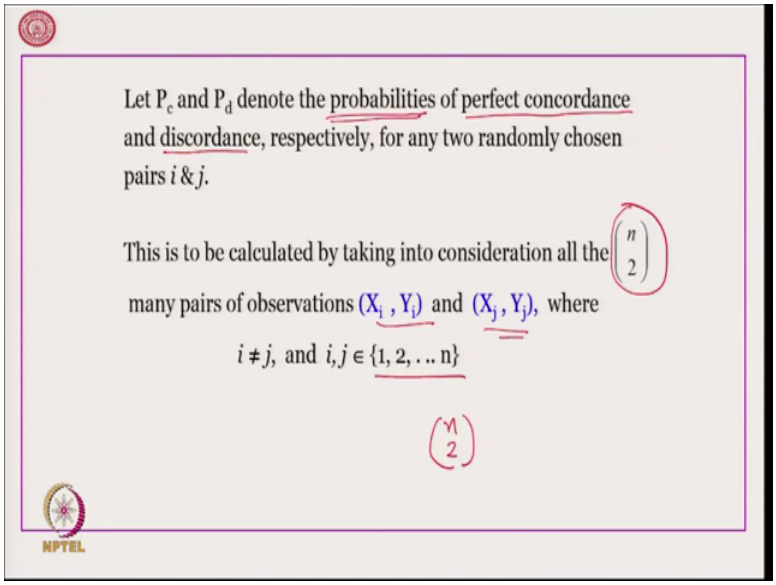
It is desirable that increasing degree of concordance will be reflected by increasing positive values of τ. That means, the more is the degree of concordance then the value of τ will go from 0 to 1. And similarly, increasing degree of discordance will be shown by the negative value as it is going from 0 to - 1.

(Refer Slide Time: 19:44)



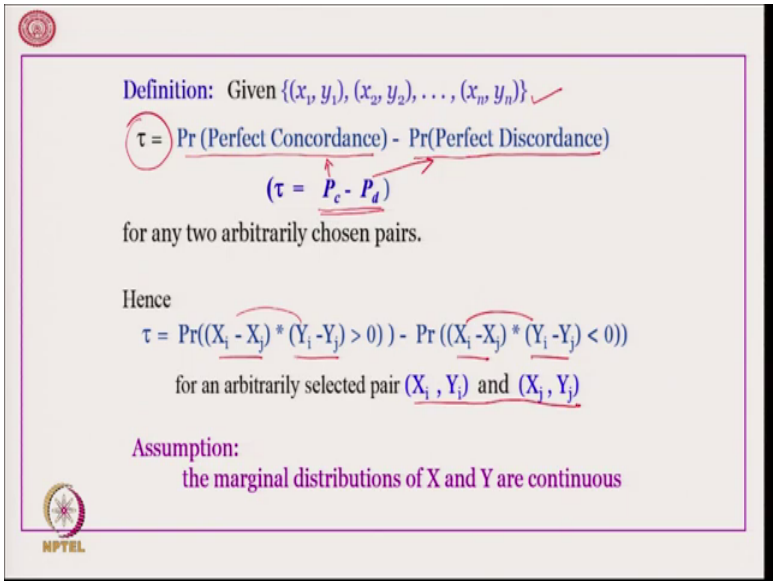
But the main question is still there. How to score a set of paired observations? We understood what is going to be for perfect concordance or perfect discordance, but how to score in between? this answer comes from probability.

(Refer Slide Time: 20:04)



So, let Pc and Pd denote the probabilities of perfect concordance and perfect discordance respectively, for any two randomly chosen pairs i, j. Now, this is to be calculated by taking into consideration all the $^{n}C_2$ many pairs (Xi , Yi) and (Xj, Yj), where i ≠j and they are running from 1 to n. Or in other words, what we are saying we will consider all the $^{n}C_2$ many pairs and we shall try to see how many of them are concordant and how many of them are discordant. So, from there, we shall try to get a measure of the probability of concordance and discordance.

(Refer Slide Time: 21:05)

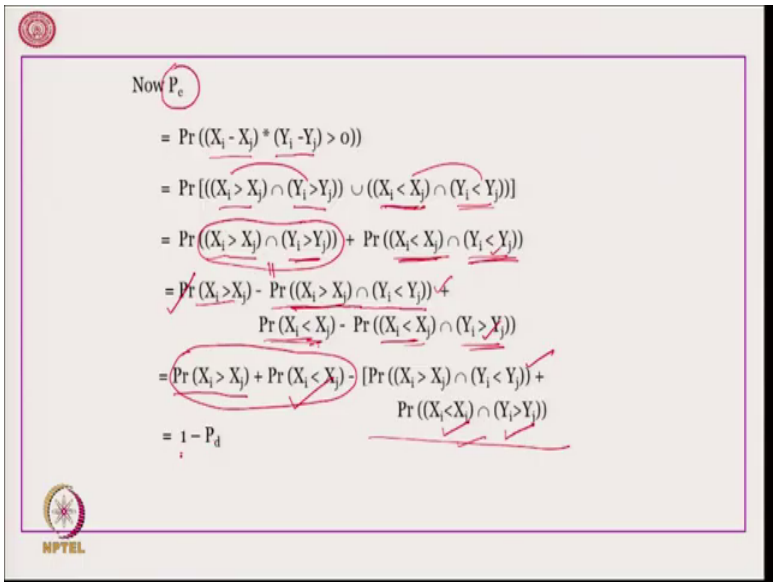


So, given X1, Y1, X2, Y2 and Xn, Yn that tau value or Kendall's τ is probability of perfect concordance minus probability of perfect discordance. Or in short, we write tau is equal to Pc

minus Pd were Pc is the probability of perfect concordance and Pd is the probability of perfect discordance. Hence, Tau is probability Xi - Xj multiplied by Yi - Y j is greater than 0 minus probability Xi - X j multiplied by Yi - Y j < 0.

As we have already discussed that if these two pairs are discordant, then this product is going to be less than 0. And if they are concordant, then this product is going to be greater than 0. And this we are telling for an arbitrarily selected pair Xi, Yi and Xj Yj. However, Kendall's Tau has a basic assumption that we have seen with respect to most of our nonparametric tests that the marginal distributions of X and Y are going to be continuous.

(Refer Slide Time: 22:29)



Now, what is Pc, it is the probability of perfect concordance. Therefore, it is the probability that Xi - Xj multiplied by Yi - Y j > 0. That means, it is the probability that Xi > Xj and Yi > Yj union Xi < Xj and Yi < Yj that means, both of them will give positive or both of them will give us negative values, so that the product will remain positive.

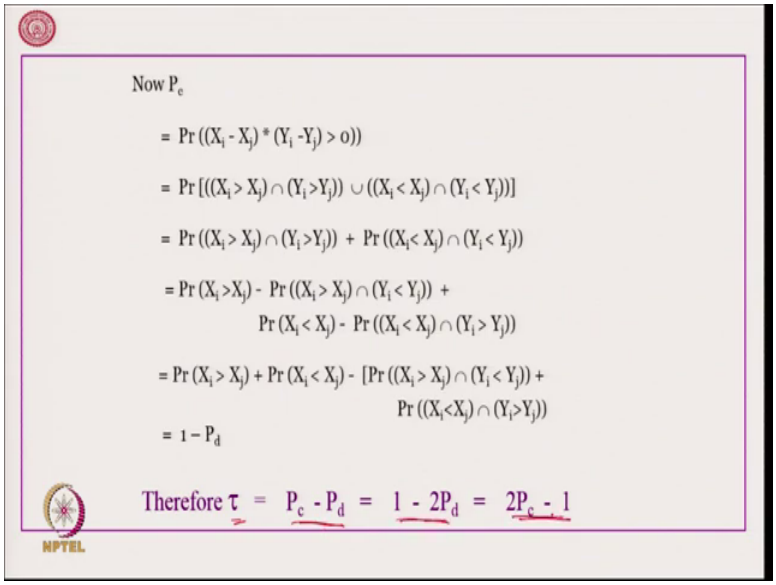
Since these two are disjoint events, we can write as probability Xi > Xj and Yi greater than Yj plus probability Xi Xj and Yi less than Y j. Now, this event Xi greater than Xj and Yi > Yj can be written as probability Xi > Xj minus probability Xi > Xj and Yi < Yj. This is very straightforward, and we can see it very easily from the Venn diagram. So, I am not going to explain that, you convince yourself that probability of this event can be written as the difference of these two probabilities.

And in a very similar way, probability Xi < Xj and Yi < Yj, this can be written as probability Xi < Xj minus probability Xi < Xj and Yi > Yj, here it was less than now, it becomes greater

than. Therefore, this whole thing can be written as probability Xi > Xj plus probability that Xi < Xj minus probability Xi > Xj and Yi < Yj plus because minus we have taken common probability Xi , Xj and Yi > Yj.

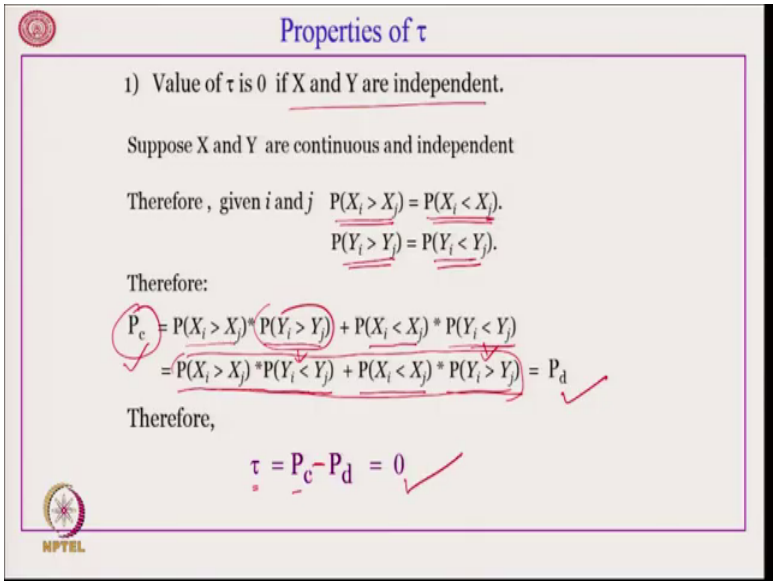
Now, this part is equal to 1, because under continuity, either Xi > Xj or Xi < Xj, therefore, their sum is going to be 1. And this is nothing but the probability of discordance between Xi, Yi and Xj Yj. Therefore, we can see that probability of concordance is equal to 1 minus probability of discordance.

(Refer Slide Time: 25:25)



Therefore, that tau or Kendall's τ can be written as Pc minus Pd, which we have already defined can also be written as 1 – 2* Pd or 2* Pc - 1. So, these are different formulae for computing Kendall's τ.

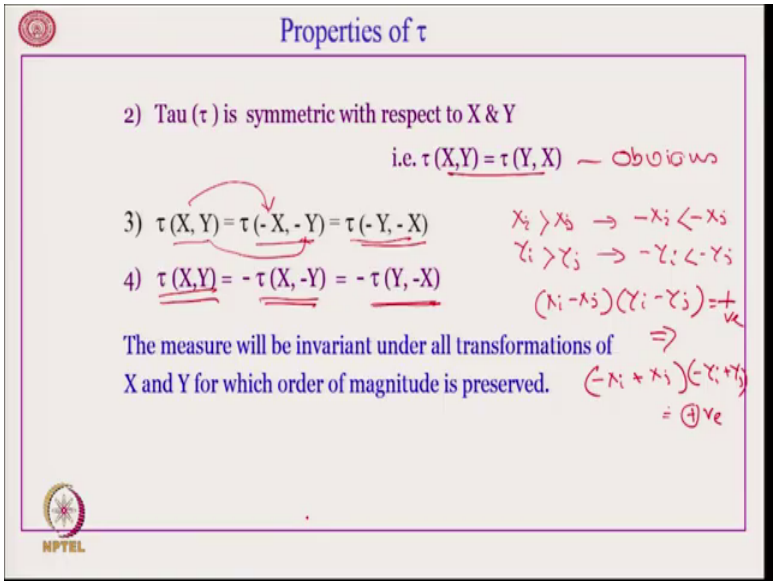
(Refer Slide Time: 25:51)



So, let us first examine a few properties of tau. The value of tau is 0 if X and Y are independent. Because X and Y are continuous and independent, therefore, given i and j probability Xi > Xj same as Xi < Xj and similarly, probability Yi > Yj is same as probability Yi < Yj. Therefore, probability of concordance is equal to probability Xi greater than 0 multiplied by probability Yi > Yj plus probability Xi < Xj multiplied by probability Yi < Y j.

Now, because of this equality, we can write this as probability Yi < Yj and similarly, this as probability Yi > Yj. Therefore, we find that Pc is same as probability Xi > Xj multiplied by Yi < Yj plus probability Xi < Xj multiplied by probability Yi > Yj. And we know that this entire thing gives us the probability of discordance. Therefore, if X and Y are independent, probability of concordance is same as probability of discordance. Therefore, the value of tau is equal to Pc minus Pd is equal to 0. Therefore, if X and Y are independent, the value of tau is 0.

(Refer Slide Time: 27:41)

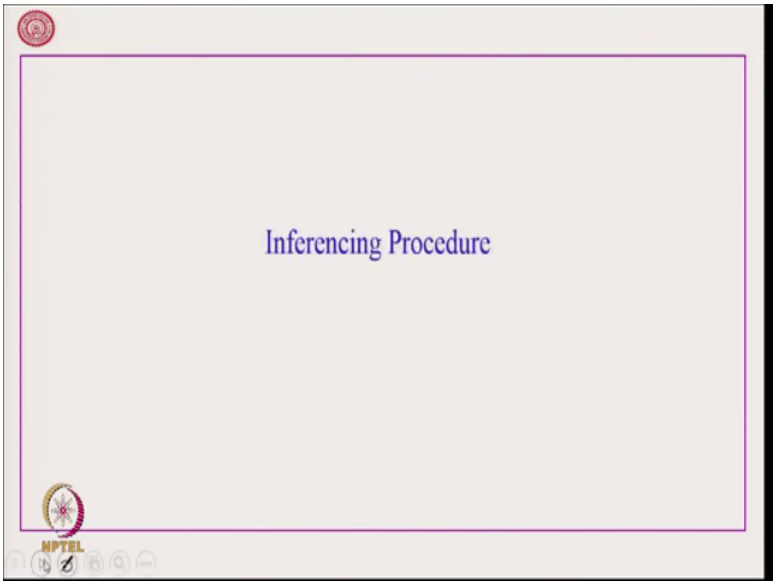


Let us examine some other properties of tau. Tau is symmetric with respect to X and Y. That is $\tau(X, Y) = \tau(Y, X)$, which is pretty obvious, I am not going to prove anything here. You can easily convince yourself $\tau(X, Y) = \tau(-X, -Y)$, this is again very obvious, because we are considering the negative values of X and negative values of Y.

Therefore, the relative values if Xi is greater than Xj will now change to - Xi is less than -Xj and similarly, Yi greater than Yj implies - Yi is less than -Yj and therefore, when we are taking the product whenever this was positive then -Xi plus Xj multiplied by -Yi plus Yj will also be positive. Similarly, with respect to negative, therefore, $\tau(X, Y) = \tau(-X, -Y)$, which is again by this observation is $\tau(-Y, -X)$.

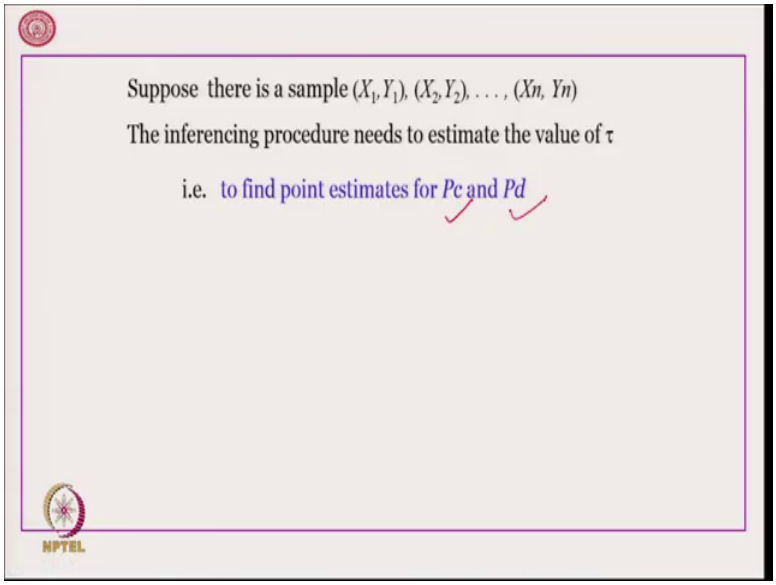
In a similar way, $\tau(X, Y)$ is $- \tau(X, -Y)$ or it is equal to $-\tau(Y, -X)$. Or in other words, the measure will be invariant and that all transformations of X and Y for which order of magnitude is preserved. Because, we are not looking at the exact value, we are looking at whether it is positive or negative. Therefore, if the relative orders are preserved, then there is no change in the value of the Kendall's Tau.

(Refer Slide Time: 30:17)



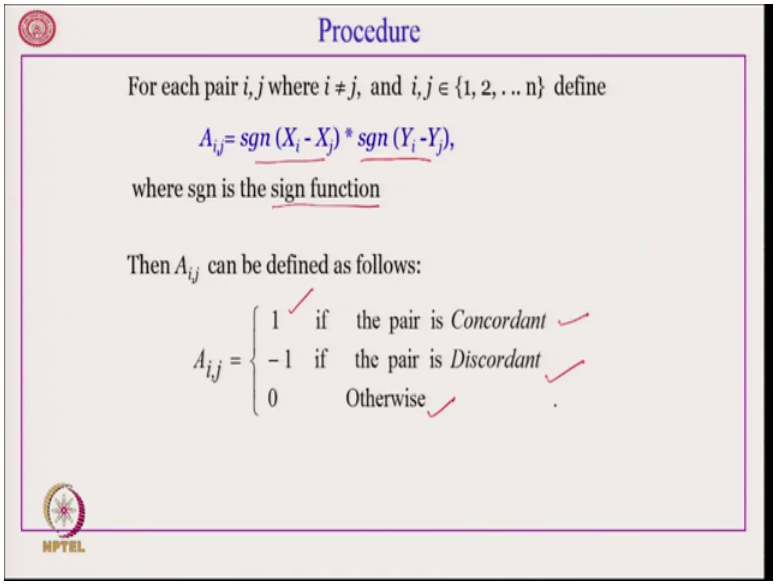
Inferencing procedure.

(Refer Slide Time: 30:19)



Suppose, there is a sample (X1, Y1), (X2, Y2), (Xn, Yn) the inferencing procedure needs to estimate the value of tau. That is to find point estimates for Pc and Pd, that is probability of concordance and probability of discordance.

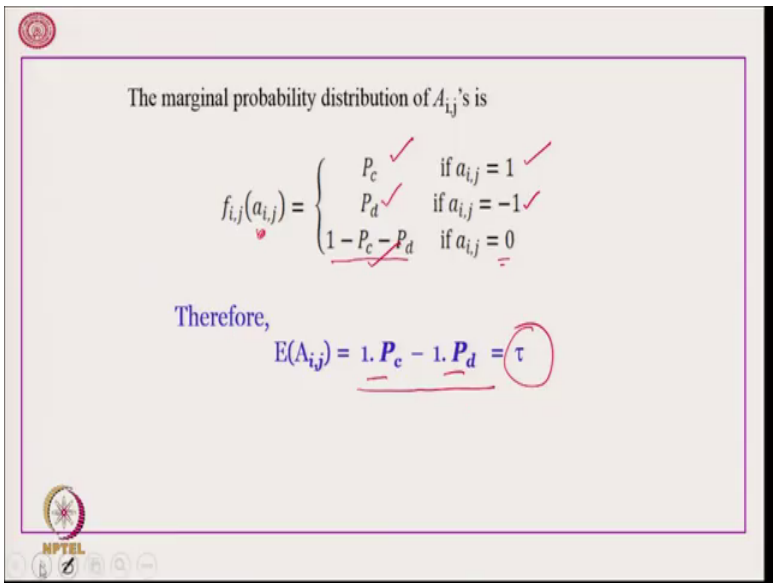
(Refer Slide Time: 30:40)



So, for each pair I, j, when i ≠j, define A i j is equal to sgn of X i - X j multiplied by sgn of Yi - Yj for sgn is the sign function that means, it is +1 or -1 depending upon whether it is positive or negative. Then Aij can be defined as follows, it is +1 if the pair is concordant, it is -1 if the pair is discordant, and it is 0 otherwise. That means, when a pair is neither concordant nor discordant and since it is a pair, it means that there will be some equality.

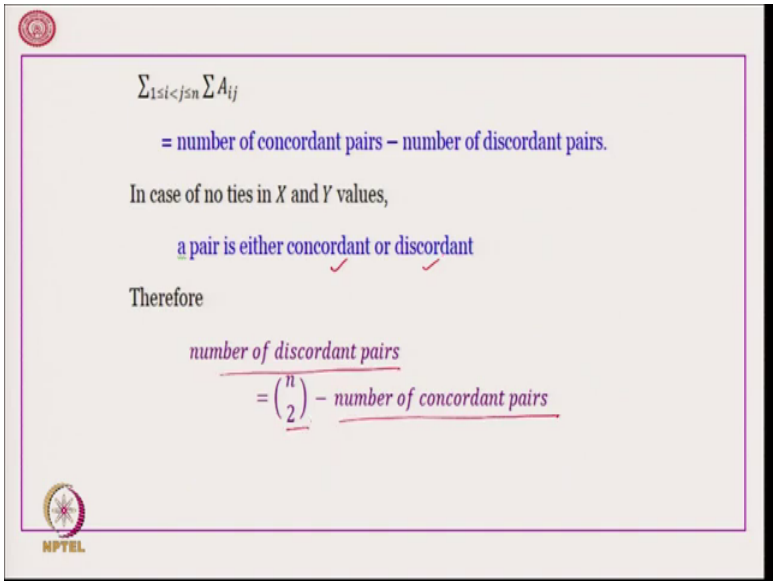
Although under continuity assumption, equalities should not happen, but in reality when we collect data, discrete data, there may be equalities, therefore, we need to exclude those data points from further calculation.

(Refer Slide Time: 31:49)



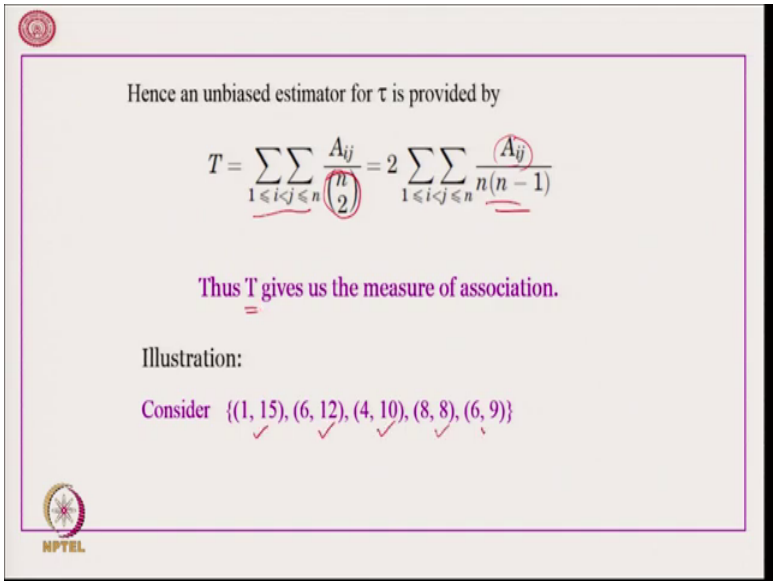
Therefore, the marginal probability distribution of Aij is this, it is Pc, if Aij is equal to 1. That means, probability that Aij will take value 1 that probability is probability of concordance that is Pc, Aij will take a value minus 1 that probability is Pd, and if Aij is equal to 0, that value is 1- (Pc – Pd). Therefore, expected value of Aij is equal to 1 times Pc minus 1 times Pd plus 0 times this quantity which we can ignore or in other words, what we have is 1 times Pc minus 1 times Pd, which is equal to the value of tau. Therefore, this Aij helps us in estimating the value of tau.

(Refer Slide Time: 32:43)



Now, if we take summation over Aij the number of concordant pairs minus number of discordant pairs that is what we are getting. If there is no ties, each pair will be either concordant or discordant as I have mentioned. Therefore, the number of discordant pairs is going to be ${}^{n}C_2$ minus the number of concordant pairs.

(Refer Slide Time: 33:10)

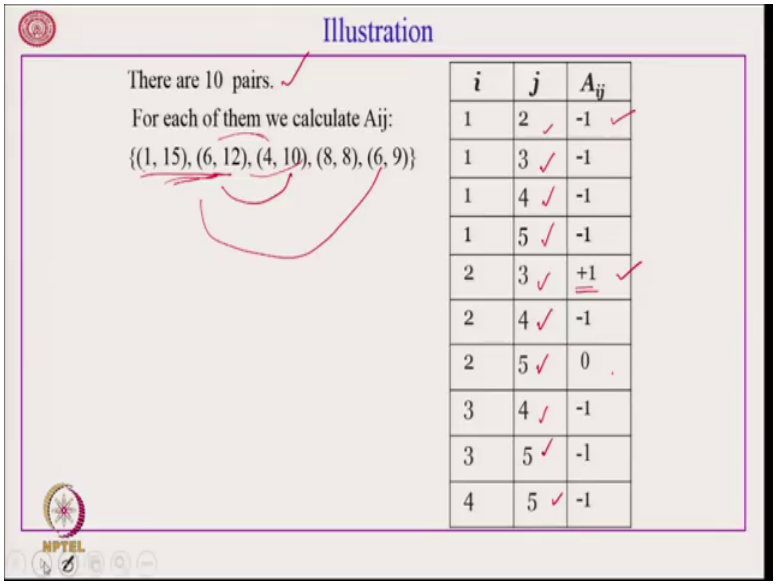


Hence, and unbiased estimator for tau is provided by sigma over i and j, i, j going from 1 to n and i less than j, A ij divided by ${}^{n}C_{2}$, because there are ${}^{n}C_{2}$many possible pairs, which is coming out to be

$$2 \sum_{1 \leq i<j \leq n} \frac{A_{ij}}{n(n-1)}$$

So, this T value we compute from the sample and it gives a measure of association between X and Y. For illustration, consider these pairs (1,15), (6, 12), (4, 10), (8, 8) and (6, 9).

(Refer Slide Time: 33:57)

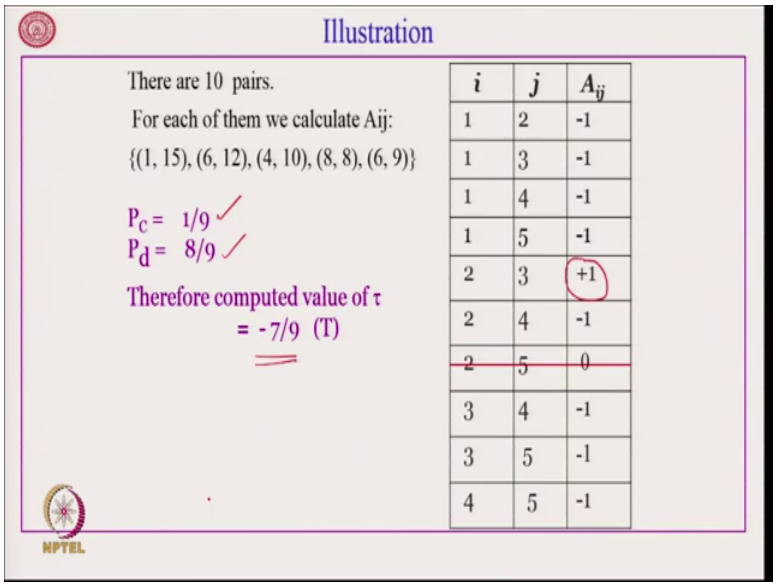


i	j	A_{ij}
1	2	-1
1	3	-1
1	4	-1
1	5	-1
2	3	+1
2	4	-1
2	5	0
3	4	-1
3	5	-1
4	5	-1

Since, there are 5 observations, therefore, we have $^5C_2$ that means 10 pairs. So, let us consider these 10 pair (1, 2), (1, 3), (1, 4), (1 ,5), (2, 3), (2, 4), (2, 5), (3, 4), (3, 5) and (4, 5). Now, we have to check how many of them are concordant, how many of them are discordant, so, we go pairwise. So, 1 and 15 as we can see that has 1, X values are 1 and 6 and 15 and 12 therefore, this is a discordant pair, therefore, the value of A ij is equal to minus 1. Similarly, You can calculate for all others, let us just check the pair 2 and 3, it is 6, 4 and 12, 10.

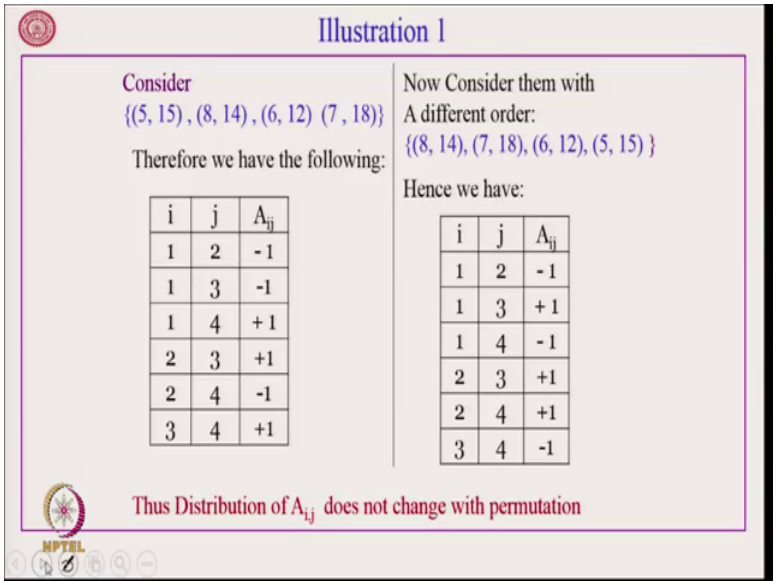
That means, as I go from second sample to the third sample, value of X is decreasing and value of Y is decreasing therefore, that is a concordant pair. And therefore, as we see, we have got a plus 1 there. What is with respect to 2 and 5? Since the X values are 6, for both of them, therefore, this pair is neither concordant nor discordant, therefore, we get the value 0. So similarly you can calculate for all other pairs, I am not going to do that.

(Refer Slide Time: 35:32)



But now, what is see that we have to discard this pair, because it is neither concordant nor discordant and then we find that there is only 1 pair, which are concordant and remaining 8 pairs, they are all discordant. Therefore, probability of concordance is equal to 1/9, probability of discordance is equal to 8/ 9. Therefore, the computed value of tau from the sample is - 7 / 9. So, that is the estimate of the Kendall's Tau for this sample.

(Refer Slide Time: 36:14)

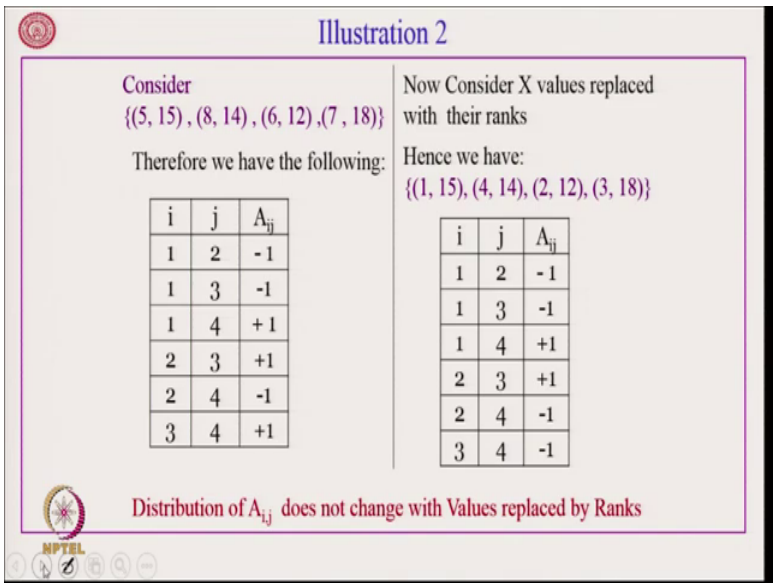


i	j	A_{ij}
1	2	-1
1	3	-1
1	4	+1
2	3	+1
2	4	-1
3	4	+1

i	j	A_{ij}
1	2	-1
1	3	+1
1	4	-1
2	3	+1
2	4	+1
3	4	-1

Now, let us see a few things considered this pair (5, 15), (8, 14), (6, 12) and (7, 18). Therefore, we have the following pairs $^4C_2$ which is 6 and in fact, we can see that (1, 2), (1, 3), (1, 4), (2, 3), (2, 4) and (3, 4) and these are the values of Aijs. Now, consider them with a different order, the same pair of observations, but we have permuted them differently, that means, the sample might have been taken in a different order. But what we see that the number of positive and number of negative still remains same. Or in other words, we can say that, the distribution of Aij does not change with permutation. That is an important property of Kendall's Tau, that if You shuffle the data, that does not alter the value of the measure of association that is tau.

(Refer Slide Time: 37:22)



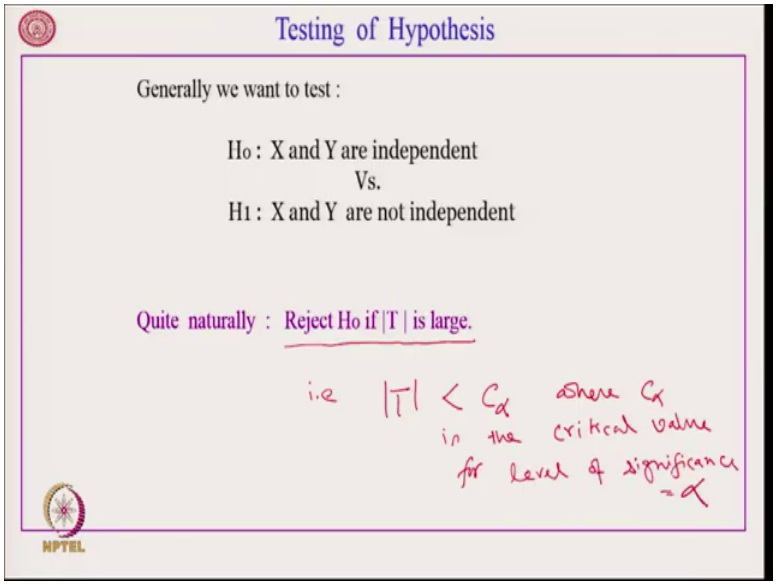
i	j	A_{ij}
1	2	-1
1	3	-1
1	4	+1
2	3	+1
2	4	-1
3	4	+1

i	j	A_{ij}
1	2	-1
1	3	-1
1	4	+1
2	3	+1
2	4	-1
3	4	-1

Let us consider another example. Again, we have taken this 4 pair of observation, we have computed this table of Aij. Now, consider X values are replaced with their ranks. That means, if we look at these the values are 5, 8, 6 and 7. So, suppose we rank them as 1, 2, 3 and 4 then now my data is transformed into this. Therefore, again if we calculate A ij, we find that there is no change in the computation of A ij between these 2 set of data. So, what does it tell us?

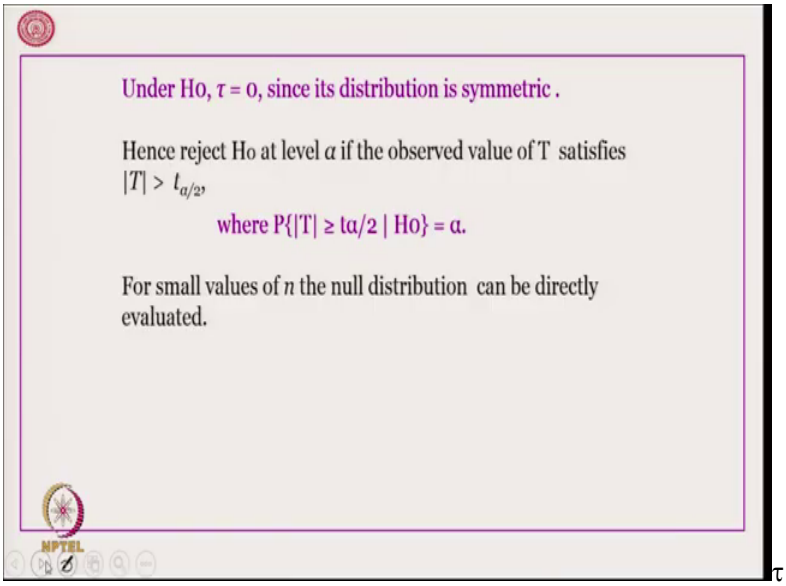
It tells us that the distribution of Aij does not change with the values replaced by ranks. What is the advantage? The advantage is that instead of numeric value, if we add A, B, C, etc and they can be ordered perfectly, then there is no change in the value of the measure of association namely Kendall's Tau.

(Refer Slide Time: 38:42)



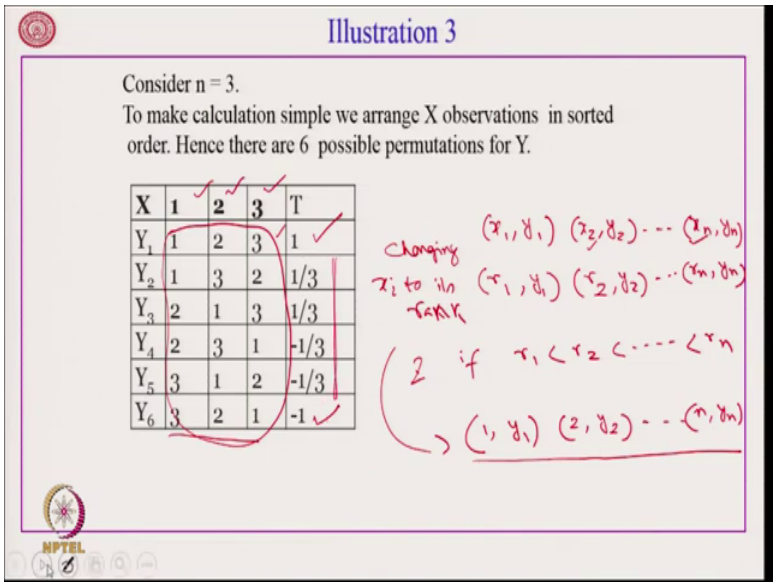
So, generally, we want to test if X and Y are independent versus X and Y are not independent. So, what do we try to do? We try to compute the value of the T and we reject Ho if T is large. That means the sample observation need not give us the exact value 0, but the maser of T should be less than some critical value Cα, where Cα is the critical value for level of significance is equal to α.

(Refer Slide Time: 39:42)



That is what we are saying under Ho: $\tau = 0$ since the distribution is symmetric, hence reject Ho at level α, if the observed value of tau satisfies the modulus of tau is greater than $t\alpha/2$. For small values of n the null distribution, of course, can be directly evaluated. And the way we have done many problems through complete enumeration, we can also do that when the value of n is small.

(Refer Slide Time: 40:18)

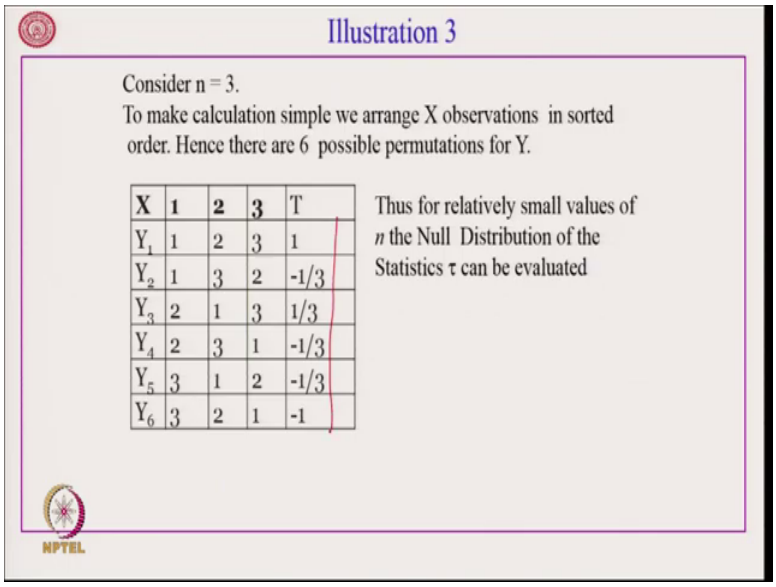


Now, let me give you another illustration. To make the calculation simple, we arrange the X observation in sorted order. So, we have arbitrary pairs (X1, Y1), (X2, Y2), up to (Xn, Yn). Suppose, we convert them to ranks, therefore, what we get rank of X1, Y1, let me write it down for you changing Xi to its rank we get (r 1, y 1), (r 2, y 2), (rn, yn) and if r1 is less

than r2 less than r3 less than rn, then this gives us of the form (1 y1), (2 y2), (n yn) where 1 is the rank of X1, 2 is the rank of X2 and n is the rank of X n.

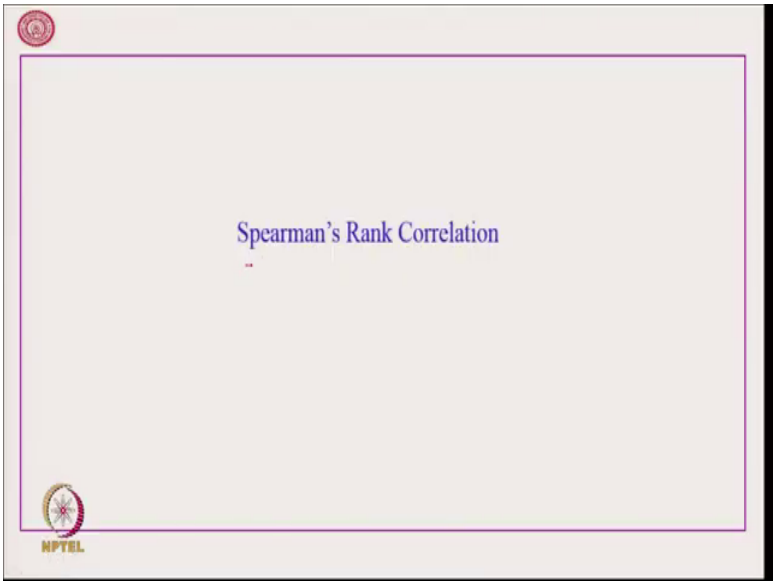
Or in other words, the calculation becomes simple if we arrange one observation, maybe the X in sorted order. Now, we have 3 pairs of observation let us sort X in increasing order 1, 2, 3 and then for Y we can have all the 6 possible permutations (1, 2, 3), (1, 3, 2), (2, 1, 3) etc and then we can compute the tau with respect to each permutation. And we will find that for this it is coming out to be 1 because they are perfectly concordant, for this order 3, 2, 1 there is perfect discordant and for others, we shall get the values like this.

(Refer Slide Time: 42:35)



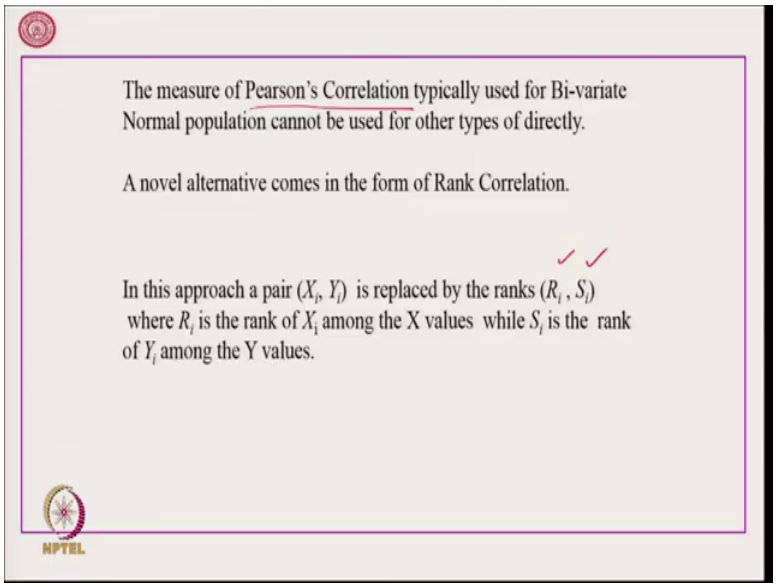
Thus, for relatively small values of n the null distribution of the statistic t can be easily evaluated because we are getting the probability distribution of t.

(Refer Slide Time: 42:55)



Let us now, consider the other metric that is Spearman's rank correlation, which also measures the association between 2 random variables which are jointly distributed.

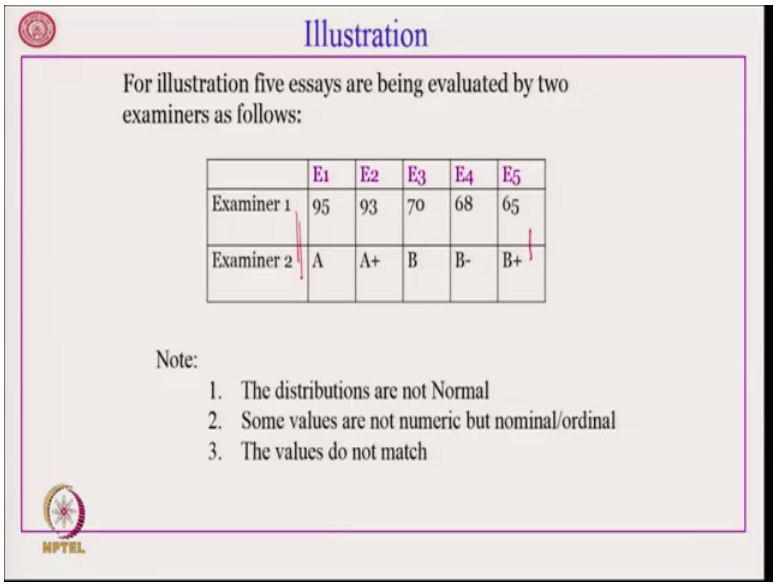
(Refer Slide Time: 43:09)



So, the measure of Pearson's correlation typically used for bivariate normal population cannot be used for other types of data directly. That is what we started with that when we are computing the correlation, which is called Pearson correlation, it can be used for normal population. But, if the data is not of that type, then we can use a subtle variation of that, which is called rank correlation.

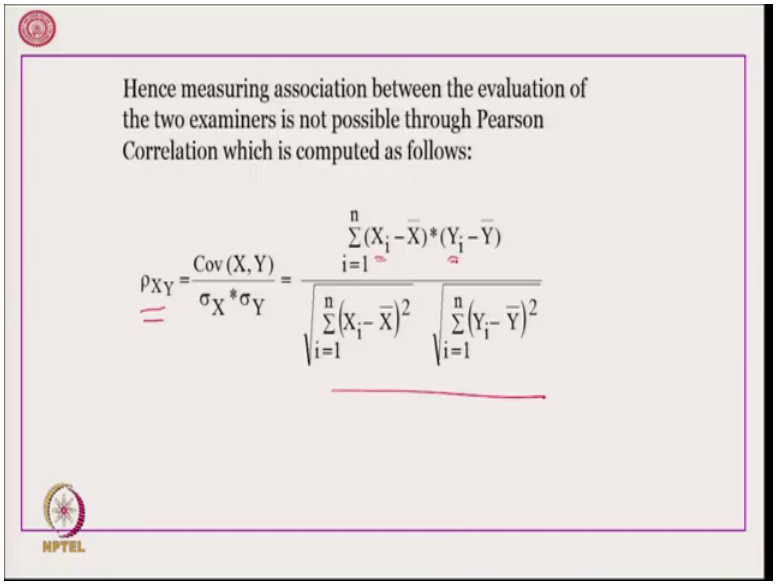
So, in this approach your pair (Xi, Yi) is replaced by the ranks (Ri Si), where Ri is the rank of Xi among the X observations and Si is the rank of Yi among the Y observations. We have already seen how we can replace a value with its rank, so we are not elaborating it much further.

(Refer Slide Time: 44:07)



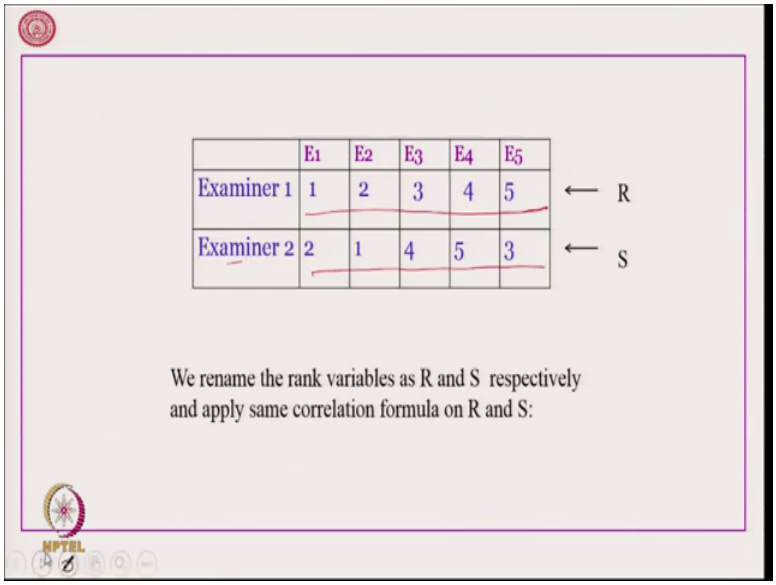
Suppose, we have the 5 essays graded by 2 different examiners, one has given numerical scores, others have given grades and we want to see to what extent the examiners agree with each other or in other words, the association between their evaluations. So, we can see that the distributions are not normal and there are values which are not numeric and also we can see that they are not having the similar opinion with respect to the 5 essays. Or in other words, there is a disagreement between the examiners.

(Refer Slide Time: 44:52)



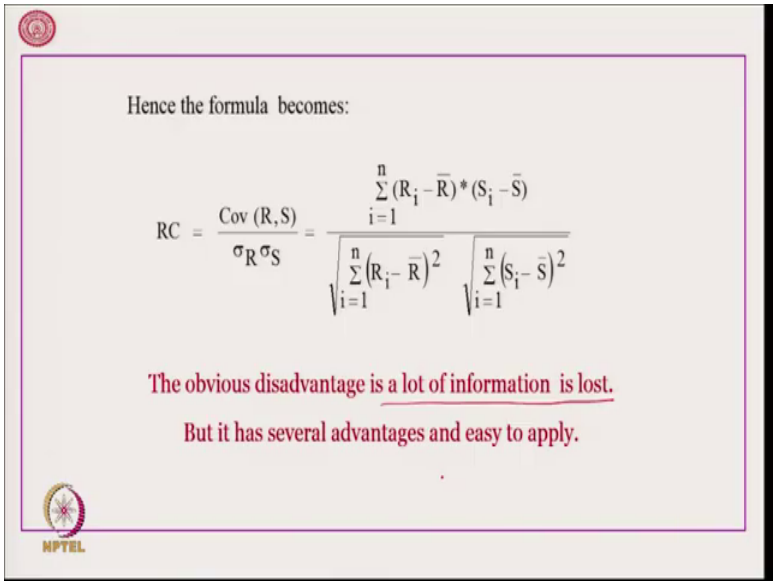
Hence, the question of measuring their association is important. Here we invoke our earlier definition of $\rho_{X, Y}$, which we have already defined, so I am not going into that detail here. But for Spearman rank correlation the Xi will be replaced by Ri, and Yi will be replaced by Si.

(Refer Slide Time: 45:14)



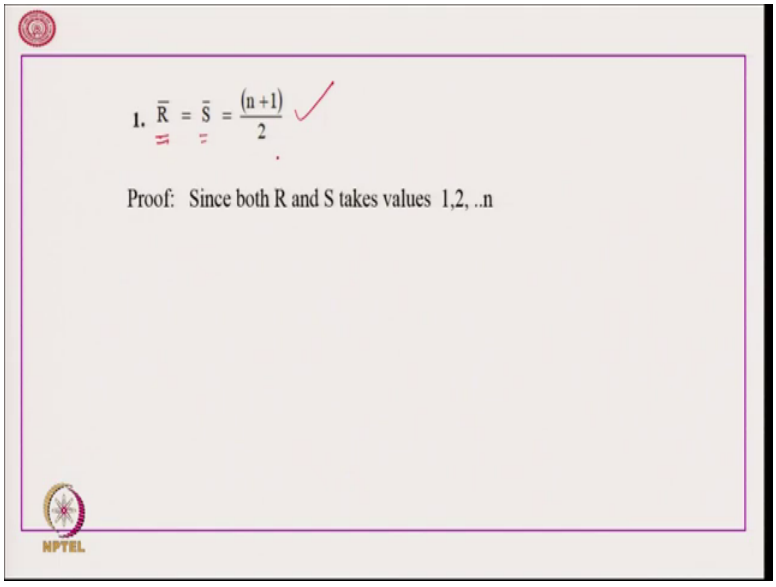
And therefore, from here we get the following. So, these are the ranks among the X's and these are the ranks among the Y's, which is the scores given by the examiner 2. Therefore, effectively by taking ranks we convert the data into numeric and this enables us to compute the correlation coefficient between R and S.

(Refer Slide Time: 45:44)



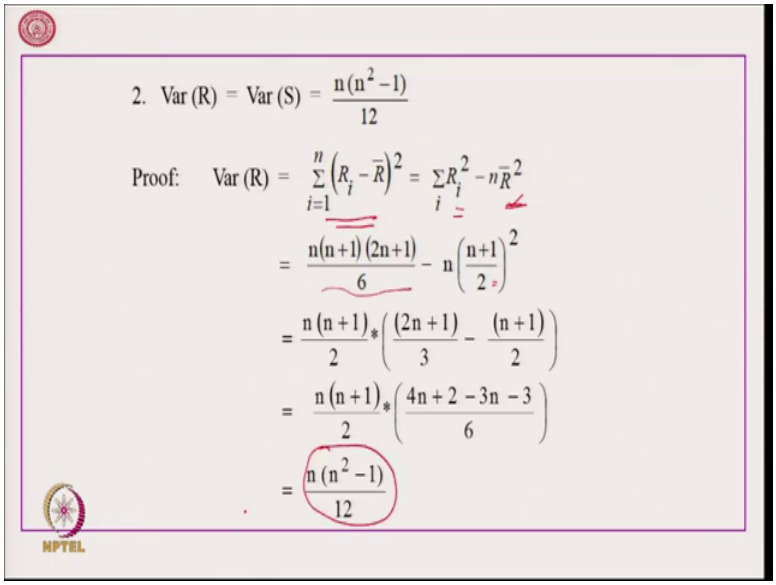
Which will now be of this formula and this is called rank correlation. It of course has certain disadvantage, because we are losing a lot of information, but it has certain advantages, because it is easy to apply.

(Refer Slide Time: 46:04)



Why it is easy to apply, because of the simple thing that R takes values 1, 2, 3 up to n therefore, mean of R value is going to be n plus 1 by 2. Similarly, the mean of S is also going to be (n + 1)/ 2.

(Refer Slide Time: 46:24)



What about the variance, the variance is coming out to be

$$\frac{n(n^2-1)}{12}$$

because it is very simple, because variance of R is equal to sigma over Ri minus R bar whole square

also we can write it as sigma Ri squared minus n times R bar square

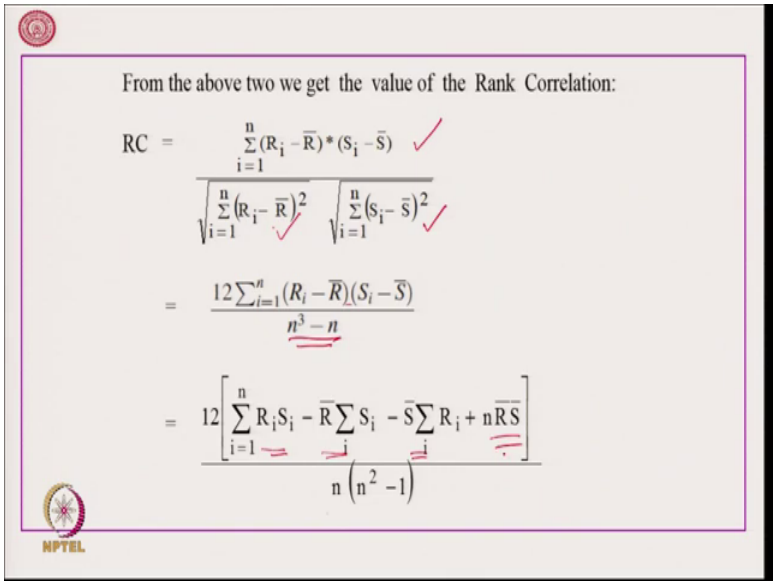
that is the alternative formula for variance. Now, if we simplify them, we find that the variance of R is equal to

$$\frac{n(n^2-1)}{12}$$

In a very similar way, variance of S is equal to

$$\frac{n(n^2-1)}{12}$$

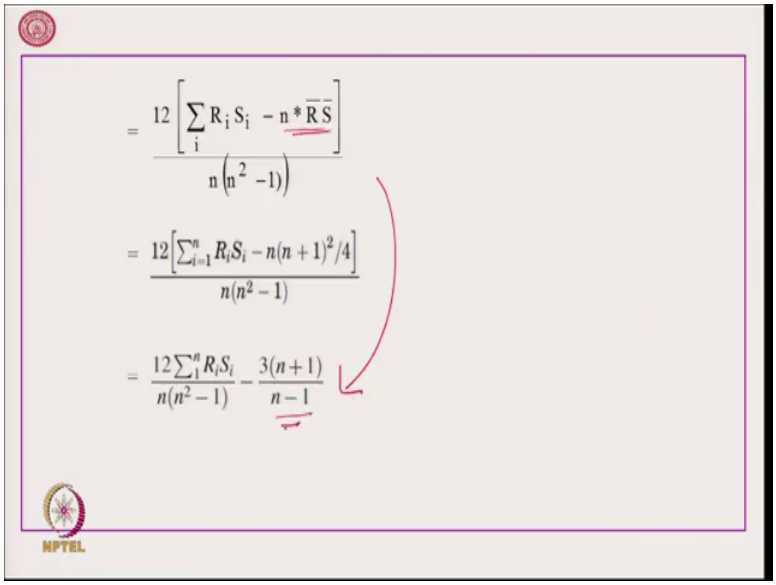
(Refer Slide Time: 47:06)



Therefore, when we put these values in the formula for rank correlation, what we get it, sigma Ri minus R bar into Si minus S bar summation over i is equal to 1 to n divided by their standard deviations. Now, because these 2 values are same, we get n cubed minus n and what we are getting eventually is this formula 12 upon n into n square minus 1 multiplied by, as we open up this product with the summation we get

$$\frac{\sum_{i=1}^{n}R_iS_i-\bar{R}\sum_i S_i-\bar{S}\sum_i R_i+n\bar{R}\bar{S}}{}=$$

(Refer Slide Time: 47:59)



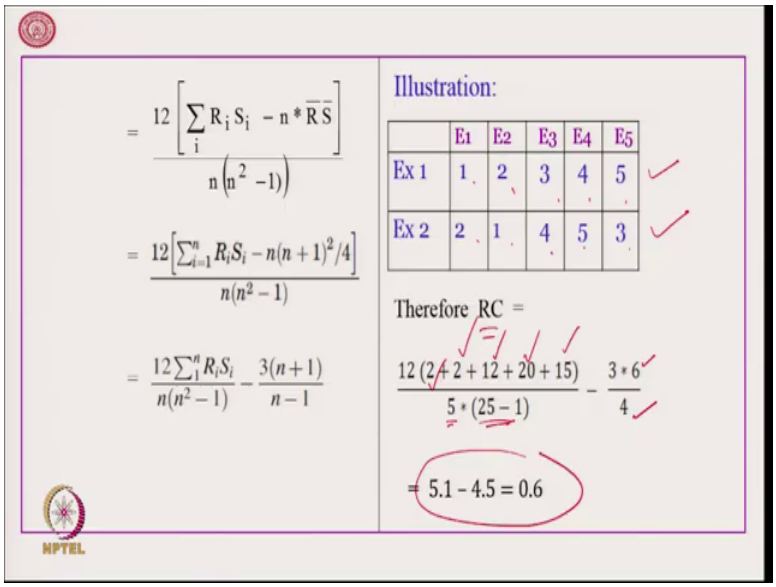
Which is coming out to be

$$\frac{12\left[\sum_i R_i S_i - n * \bar{R}\,\bar{S}\right]}{n\left(n^2 - 1\right)}$$

which can be further simplified to

$$\frac{12\sum_1^n R_i S_i}{n(n^2-1)} - \frac{3(n+1)}{n-1}$$

(Refer Slide Time: 48:22)

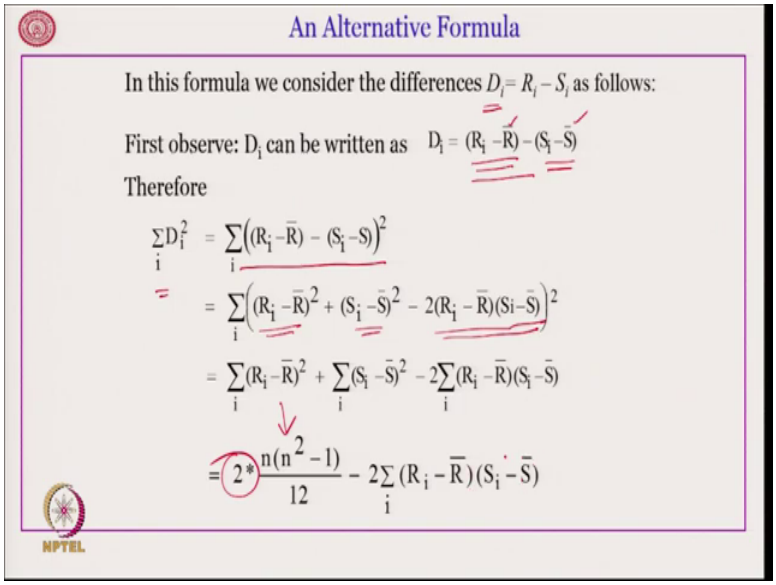


Therefore, consider the problem that we are dealing with. We have the ranks of R's like this and the rank of S like this, therefore the rank correlation is coming out to be

$$\frac{12\,(2+2+12+20+15)}{5*(25-1)} - \frac{3*6}{4}$$

which is coming out to be 0.6. So, that is the rank correlation or Spearman's rank correlation between these two pairs of observation.

(Refer Slide Time: 49:14)



Now, there is a alternative formula for that one. This is computed using the value of Di, which is the difference in ranks for Ri and Si. So, Di we can write it as

$$(R_i - \bar{R}) - (S_i - \bar{S})$$

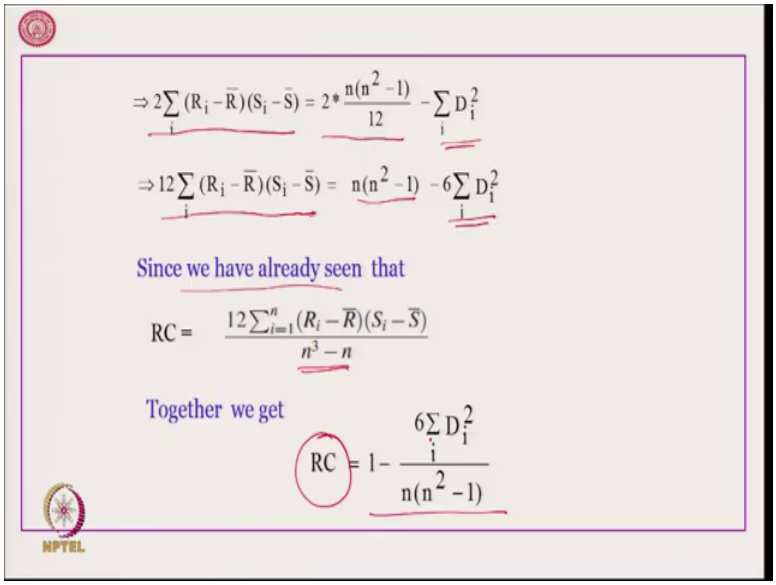
because, R bar is same as S by both are (n+1)/2. Therefore, summation of Di square is equal to summation of square of these quantities, which we can open up and we get

$$\sum_i (R_i - \bar{R})^2 + \sum_i (S_i - \bar{S})^2 - 2\sum_i (R_i - \bar{R})(S_i - \bar{S})$$

Now, this is giving you the notion of the variance and 2 of them that is how we get

$$2 * \frac{n(n^2 - 1)}{12} - 2\sum_i (R_i - \bar{R})(S_i - \bar{S})$$

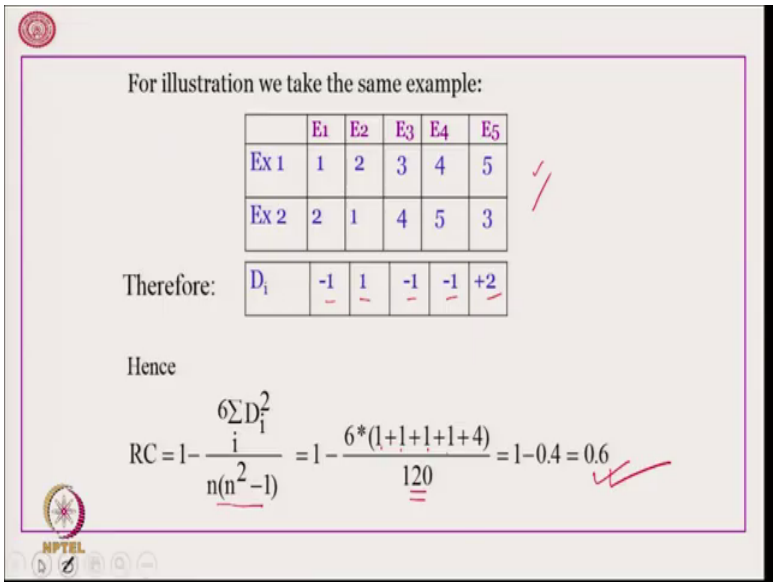
(Refer Slide Time: 50:15)



Therefore, sigma 2 into sigma Ri minus R bar into Si minus S bar we can write it as 2 star n into n square minus 1 by 12 minus summation over Di square or if we multiply this by 6 we get 12 into sigma Ri minus R bar into Si minus S bar is equal to n into n square minus 1 minus 6 into sigma Di square. Since, we have already seen that the rank correlation is equal to 12 into Ri minus R bar into Si minus S bar divided by n cube minus 1. Therefore, together we can write that the rank correlation is equal to

$$1 - \frac{6\sum_i D_i^2}{n(n^2-1)}$$

(Refer Slide Time: 51:12)

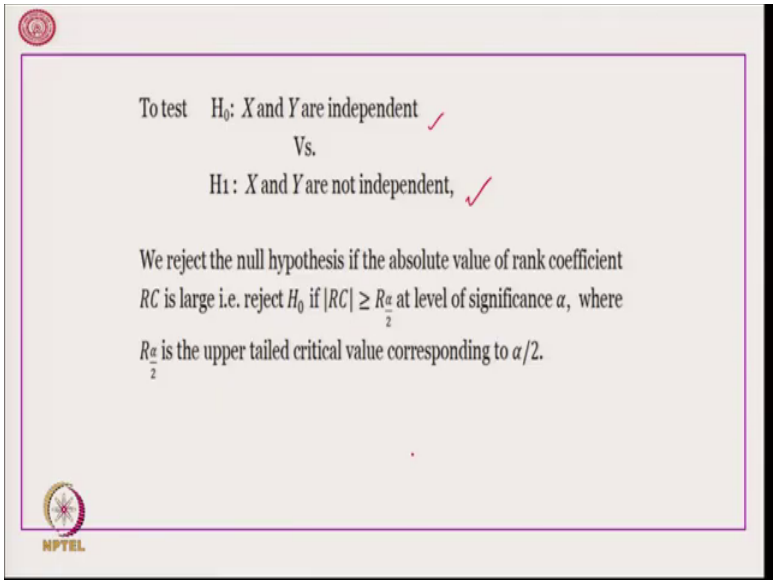


Therefore, if we apply the same formula for the same data, then we get Di is equal to -1, +1, -1, -1 and +2. Therefore, 1 minus 6 into sigma Di squared is coming out to be

$1 - 6 \times (1 + 1 + 1 + 1 + 4) / n \times n^2 - 1,$

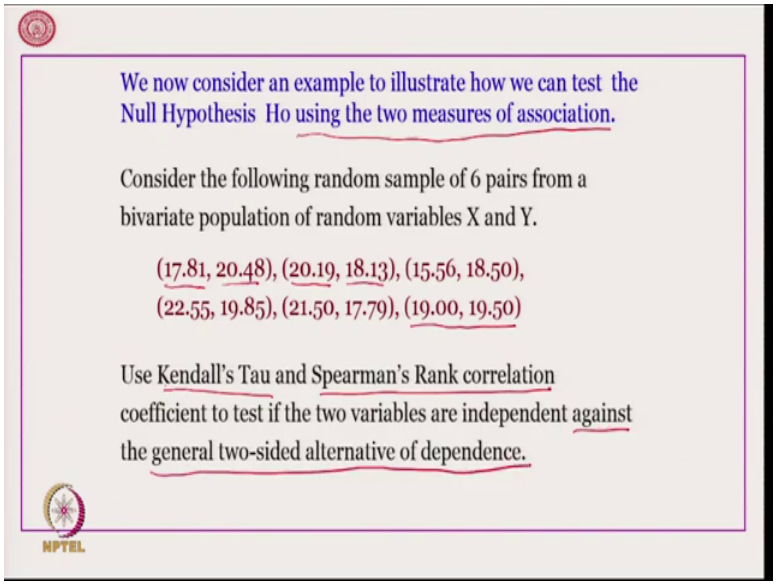
which is 5 into 24, that is 120 therefore, again we get the same value 0.6. So, that is how we calculate the Spearman's rank correlation for a given set of paired observations.

(Refer Slide Time: 51:59)



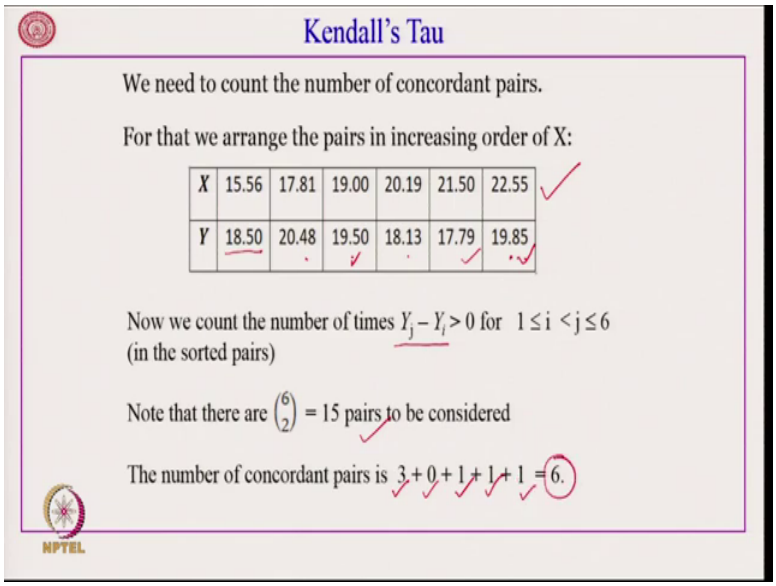
Again, we do the similar type of testing whether X and Y are independent or they are not independent, therefore we reject the null hypothesis. If the absolute value of the rank coefficient is somewhat large, that is it is greater than some threshold, otherwise we are going to accept that they are independent.

(Refer Slide Time: 52:22)



We now consider an example to illustrate how we can test the null hypothesis Ho using the two measures of association, that is Kendall's Tau and Spearman's rank correlation. So, consider the following random sample of 6 pairs from a bivariate population of random variables X and Y. So, the values are 17.81 and Y is equal to 20.48, second sample is X is 20.19 and Y is 18.13, like that this is the 6th sample X is 19 and Y is 19.50. So, the task is that to use Kendall's Tau and Spearman's rank correlation to test if the two variables are independent and alternative is that these two are not independent.

(Refer Slide Time: 53:36)

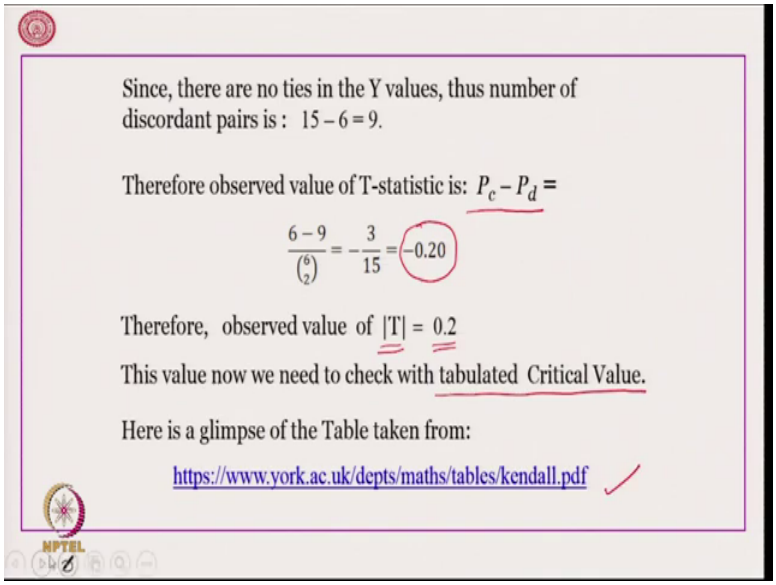


So, let us first discuss Kendall's Tau. We need to count the number of concordant pairs. So, for that, we first arrange the pairs in increasing order of X. So, you can see that 15.56, 17.81 like that up to 22.55 and Y's are kept accordingly as per the pairs to stop now, we count the

number of times Yj - Yi is greater than 0, because we always know that if j > i, then Xj > Xi. So, we are looking at $Y_j > Y_i$ for $1 \leq i < j \leq 6$.

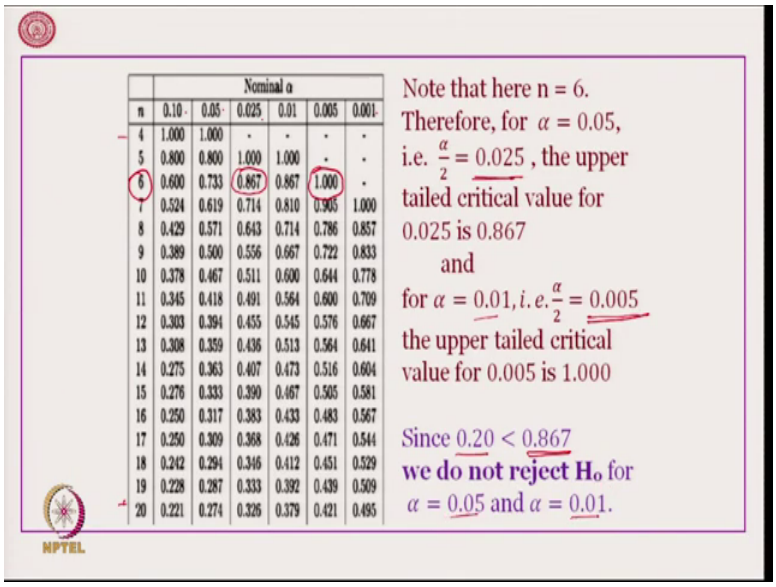
Now, there are ${}^{6}C_{2}$, that is 15 pairs to be considered and we can see that the number of concurrent pairs is 6. How? So, we look at for the first one Y is 18.50 there are 1, 2 and 3 of them are bigger than this value, therefore, corresponding to the first pair, we get the value 3. For the second pair, it is 20.48, the value of Y, but none of the other one is greater than that, therefore, we get 0. In a similar way, we get 1 for this one, namely from here, 1 for 18.13, and 1 for 17.79. So, the number of concordant pairs is 6.

(Refer Slide Time: 55:39)



Since there are no ties in the Y values, the number of discordant pair is going to be 15 minus 6, that is 9. Therefore, the observed value of T statistic, which is probability of concordance minus probability of discordance, which is coming out to be 6 minus 9 divided by ${}^{6}C_{2}$, that is 15 and this gives us - 0.20. Therefore, of sound value of the T, which is the absolute value of the Pc minus Pd is coming out to be 0.2. So, that is the statistic that we have to use for acceptance or rejection of the null hypothesis. Therefore, what we do, we check it with the tabulated critical value. So, we have taken the table from this side, I have given it for your benefit, and we get the following.

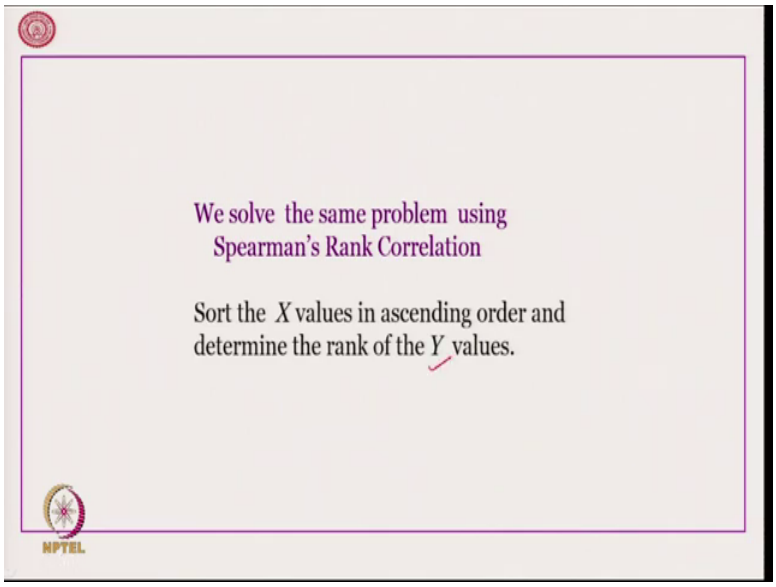
(Refer Slide Time: 56:44)



So, this is the sample from the table where corresponding to different n's, I am showing some 4 to 20 and different values of α 0.1, 0.05, 0.025 etc. up to 0.001 we have been given the critical value. So, for our example, n is equal to 6, therefore, when α is equal to 0.05, α by 2 is 0.025. Therefore, the critical value that we can see is 0.867. Again, if we look at 1% level of significance, then α is equal to 0.01, therefore α by 2 is equal to 0.005 and the corresponding value given here is 1.

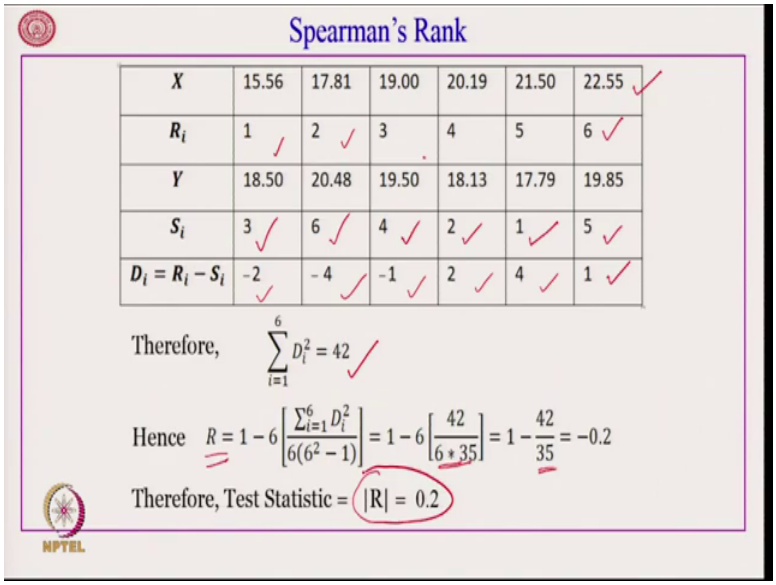
Now, our test statistic is 0.2 which is less than 0.867 and therefore naturally < 1, therefore, we cannot reject the null hypothesis for α is equal to 0.05 and α is equal to 0.01.

(Refer Slide Time: 58:09)



Now, let us solve the same problem using Spearman's rank correlation. So, we have already sorted the X values in ascending order and let us determine the rank for the Y values.

(Refer Slide Time: 58:29)



So, let us focus on this table X is in increasing order, therefore, the rank of the elements are R1 = 1 or R2 = 2 up to R6 = 6. Now, these are the corresponding Y values and let us see what are their ranks. So, this is rank 1 that is S5 is equal to 1, S4 is equal to 2, S1 is equal to 3, S3 is equal to 4 and S6 is equal to 5 and S2 is equal to 6. Therefore, what we get Ri - Si is equal to -2, - 4, - 1, 2, 4 and 1.

Therefore, as we know we compute sigma Di square which is coming out to be 42 and therefore, the value of R is equal to

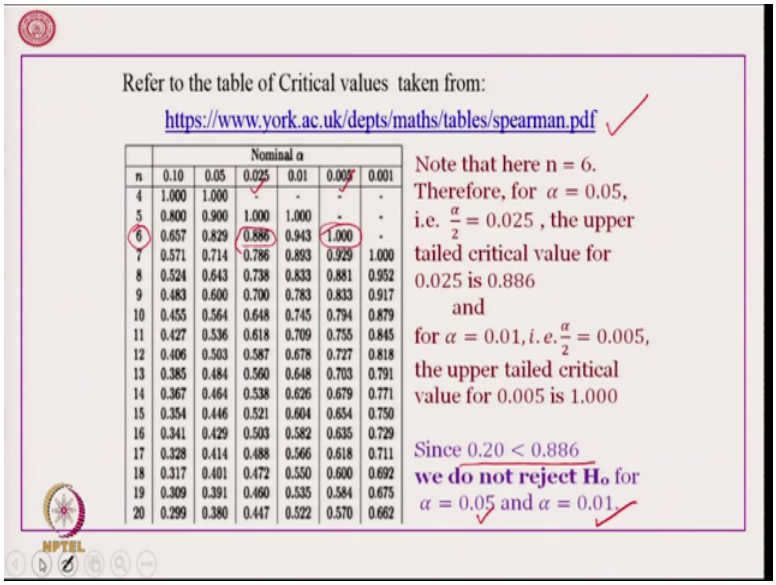
1 – (6 *42 / 6 *35)

that is

1 - 42 / 35 = -0.2.

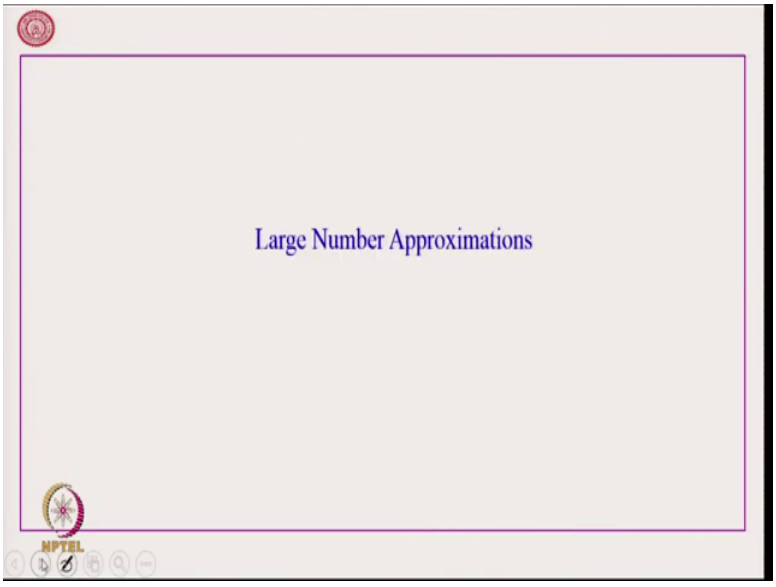
Therefore, the test statistic R is also 0.2, because we take the absolute value of this.

(Refer Slide Time: 59:59)



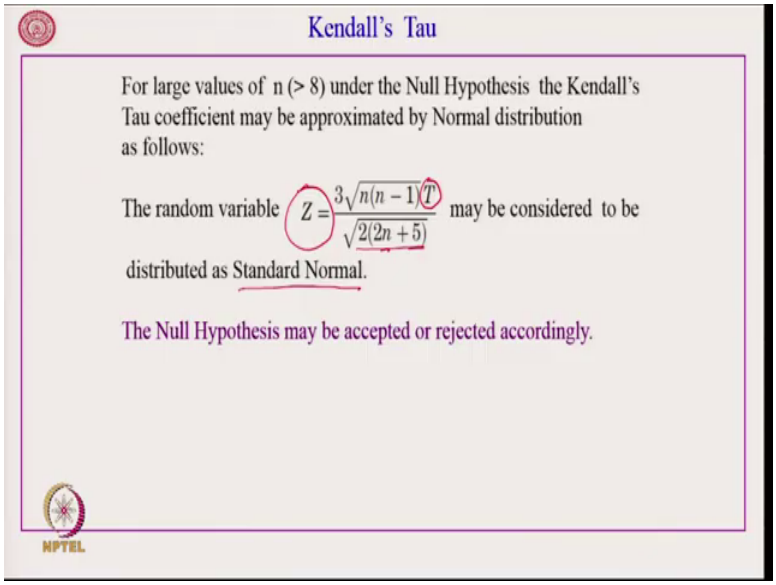
Again as we before we look at the corresponding table, again I have given you the link for the table. And as before we look at for 0.025 and 0.005, the values that we can see corresponding to n is equal to 6 are 0.886 and 1. And therefore, since the observed value is less than 0.886, we cannot reject the Ho for $\alpha = 0.05$ and $\alpha = 0.01$. So, that is how we use the Kendall's Tau or Spearman's rank correlation to test a hypothesis about whether the sample pairs are showing independence or not.

(Refer Slide Time: 61:03)



Large number approximations.

(Refer Slide Time: 61:06)

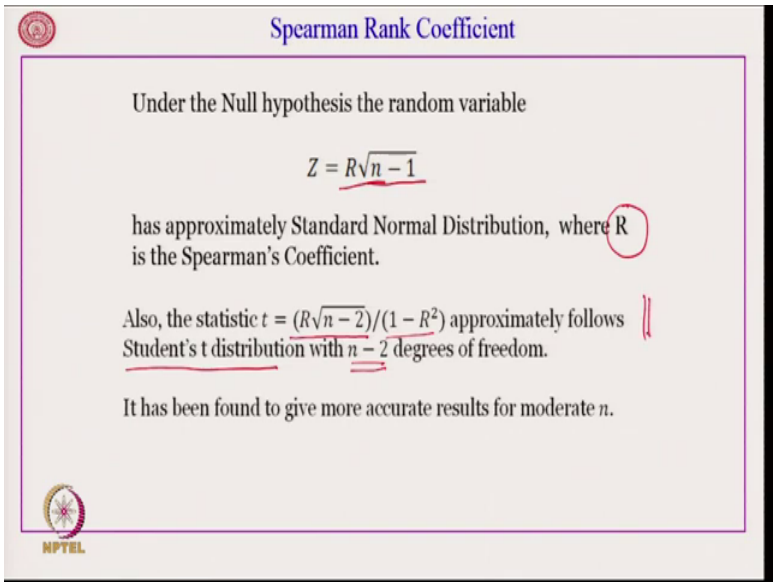


As usual, when n is slightly large, we can try to approximate that using standard normal distribution, with respect to Kendall's tau, what we find that if t is the observed value of the statistic, then

$$\frac{3\sqrt{n(n-1)}}{\sqrt{2(2n+5)}}$$

this whole multiplied by T may be considered as a standard normal distribution. We are not going to derive this thing, let us accept that and that null hypothesis will be accepted or rejected by using Z test with this value of Z that is obtained by a transformation of the value that we got from the sample for the statistic T.

(Refer Slide Time: 62:09)



With respect to Spearman rank coefficient, under the null hypothesis, the random variable Z is equal to R times root over n minus 1 has approximately a standard normal distribution, where R is the obtained value of the Spearman rank coefficient. However, it has been found that another statistic

$$(R\sqrt{n-2})/(1-R^2)$$

approximately follows Student's t distribution with n - 2 degrees of freedom.

And it has been found that for moderate values of n, that is n is not too large, then this T statistic gives a better result than the normal approximation. Okay friends, I stop here today. In the next class, I shall continue with some more discussions about different nonparametric tests. Thank You.