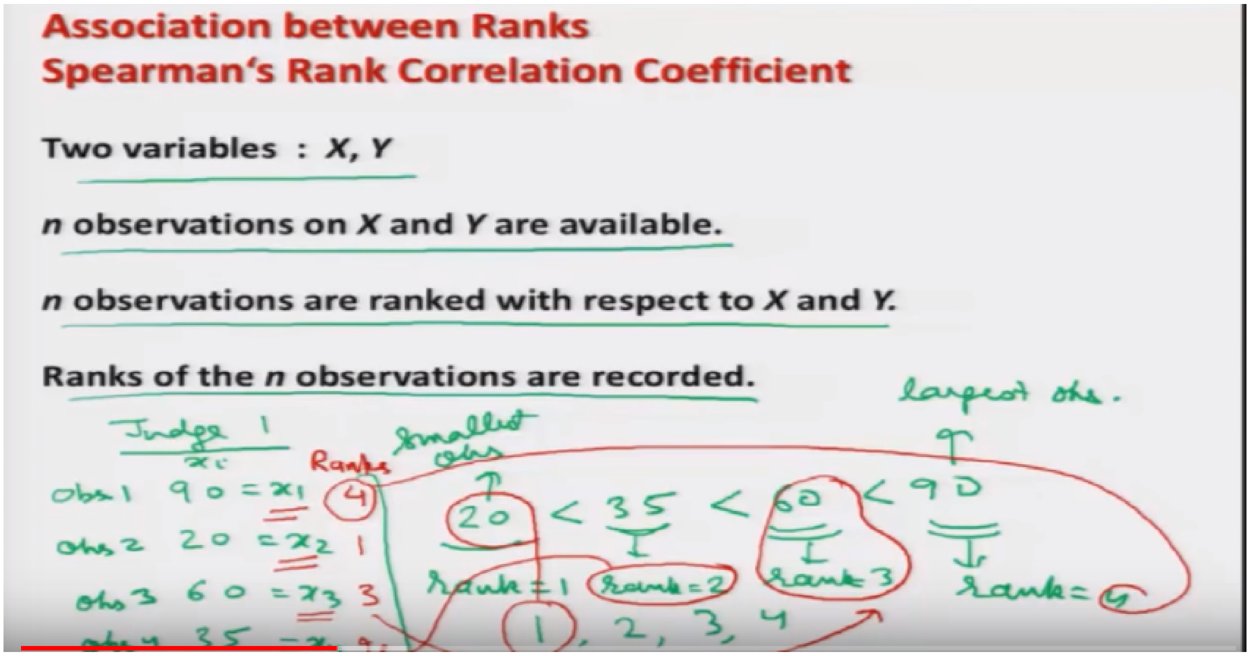
Lecture – 30
Association of Variables – Rank Correlation Coefficient

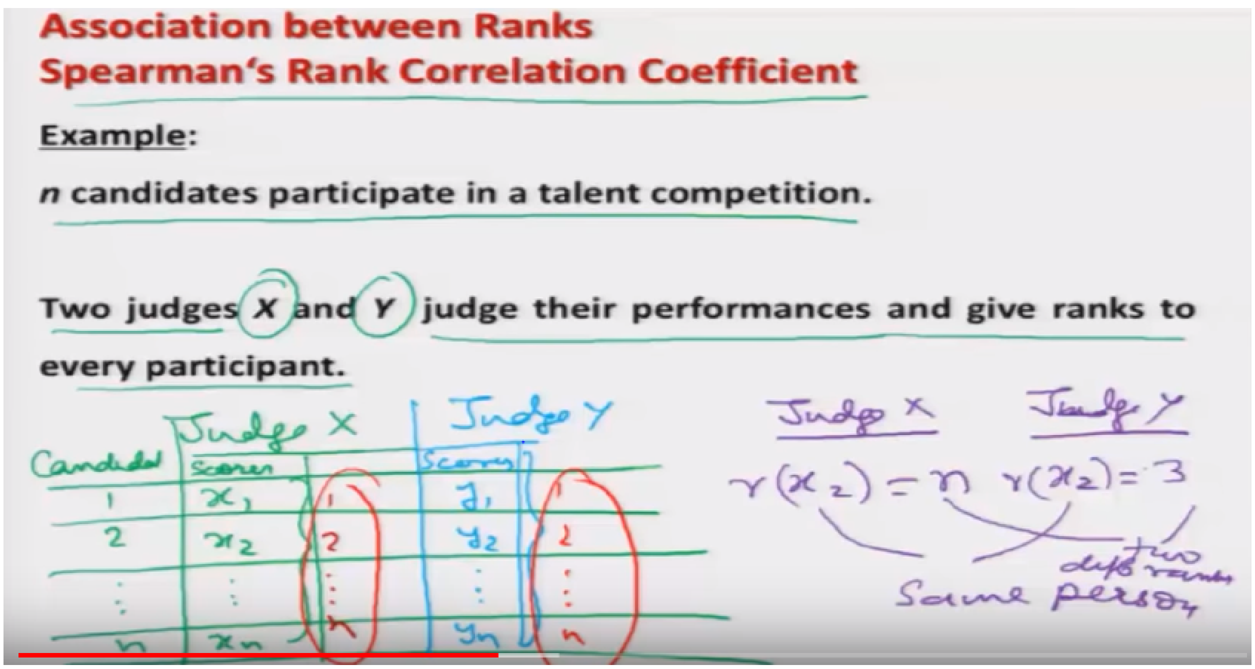
welcome to the lecture on the course descriptive statistic with R software, you may recall that in the last two lectures, we had discussed the concept of correlation coefficient and we learned that how to compute it in the R software. So, you may recall that when we started the discussion on the topic of Association of two variables, we had discussed three possible situations where we would like to measure the degree of Association. First was, when the variables are continuous, and the data is collected on some continuous scale. Second situation is where the data is collected as ranks, that means the observations are obtained and they are converted into the ranks and then we need to find out the correlation coefficient, and the third one is where the data is say categorical, where you try to count the number or the frequency. So, in this lecture we are going to consider that now the second case where the data is obtained as the ranks of the observations. Now, first question comes where such situations can occur, although I had explained you in the earlier lecture, but I will take a quick review, Suppose there is some fashion show that is going on and suppose a model comes on the stage and suppose there are two persons who are looking at the performance and they are giving some scores. Now, what do you expect, you expect that in case if the performance is good, then both the judges should give some higher score and in case if the performance is bad, then you expect that both the judges should give lower scores. Now, in practice it is very difficult that the judges are giving exactly the same is scores, suppose their scores are to be given in the say between zero and hundred, it is very difficult that they are giving the same scores, if the person is good, they may give a score of say 80 or say 85. So, now the question is this, how to measure the association between the opinions expressed by those marks by those two judges, what we can do that, we can rank the its scores of different candidates, suppose they are ten candidates. So, those ten candidates are just by the two judges judge, one has given some his scores to the ten candidates and just two also has given their scores to the ten candidates. Now instead of using the scores, we will try to find out the ranks, that which of the candidate got the highest rank, which of the candidate got the second highest rank and which of the candidate got the lowest rank and these ranks will be calculated for both the judges, and then we would like to find out the association and direction of the association between the ranks given by those two judges and this can be achieved by using the concept of rank correlation coefficient, what do we expect on the basis of what we have learnt in the case of correlation coefficient, if both judges are giving the similar scores or they have the similar opinion, if a person is good, then it is good for then he or she is good for both, then the correlation should be positive and it should be say quite strong and incase if for there is an opposite opinion that means the judges just feel opposite, one judge just said good and say another judges judge says bad, then in that case, we expect that the correlation should be negative and now how close or how strong or say how weak this depends on the magnitude of the correlation coefficient. So, you will see that the similar type of concepts will also be there in case of rank correlation coefficient as in the case of Pearson's correlation coefficient.

Refer Slide Time: (05:49)



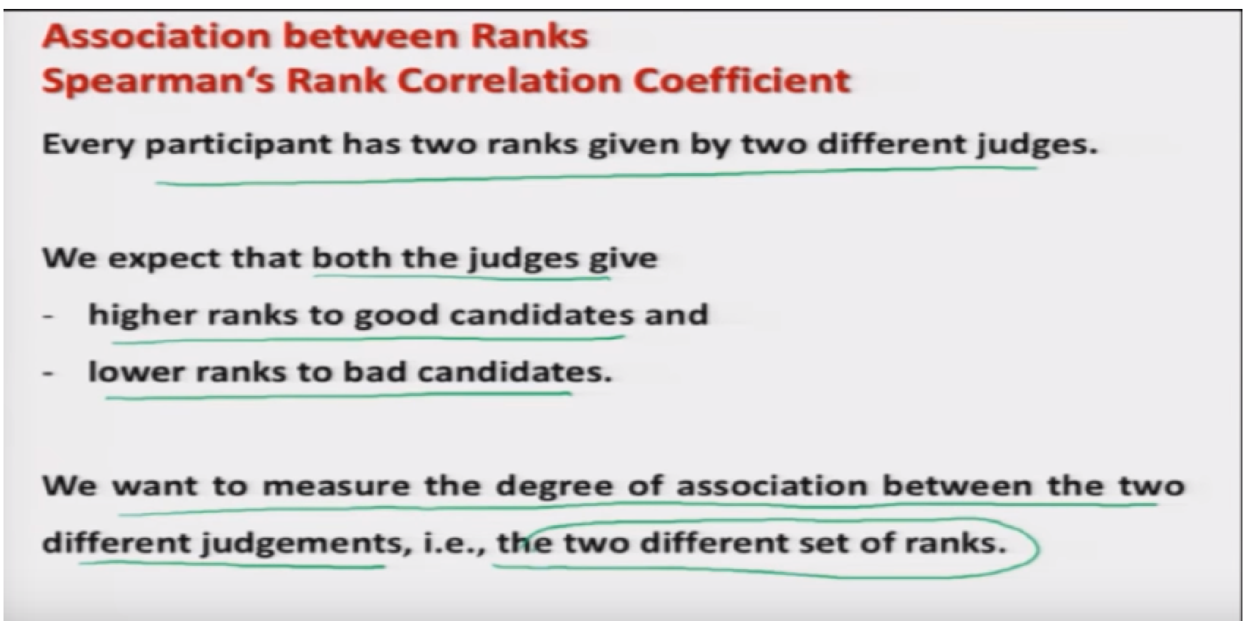
So, now we discuss our lecture and we will simply assume that here that we have here two variables x and y and and observations on x and y are available, Right! and after this whatever are the observations they are ranked with respect to x and y and ranks of those and observations are recorded. What does this mean, suppose if I say in the example which I have given, suppose I say there is a judge one and he has given the scores say here 90, 20 say 60 and say here 35. So, this is my observation number one, observation number two, observation number three and observation number four. So, these are now essentially the values of X's. So, this is my here x_1, this is x_2, this is x_3 and this is x_4. Now, what I can say I will find out the ranks, if you try to order these observations, you can see here that the smallest observation here is 20, that is the smallest observation and after this we have 35, then we have 60 and then we have 90. So, 90 is the largest observation, Right! So, if you try to give the rank, so the smallest observation will be getting a rank equal to 1 and there are 4 observation. So, the largest observation will be getting a Rank for and the second largest observation which is here 60, it will be getting ranked 3 and third largest observation, it will give here the Rank dash. So, you can see here now 1, 2, 3 and here 4, these are the ranks given to these observations. Now, if I try to write down here the ranks, what is the rank of this x_1, which is having the value 90, this is here rank is 4 in which is written in red color. Now, second observation is here x_2 equal to 20, what is the rank of this observation 20, this is here 1, you can see here and this I can write down here 1, Similarly, how this 4 was coming for was coming from here and next if you try to see x_3, the value of x_3 is 60 and what is its rank, this rank here is 3. So, this comes over here and I write here this 3 and similarly here x_4 is equal to here 35 and 35 has got the rank 2. So, this comes here and we write it here too. So, now you can see here that whatever are the its scores given by the judge they have been converted into the ranks and similarly there will be a second judge and we will try to convert the scores of judge two also in the order of ranks and then we will try to find the correlation coefficient between the two ranks. Now, incase if you think mathematically, that how to obtain the value of the correlation coefficient, in this case, you may recall that in the case of Pearson correlation coefficient, we have taken this x_i and y_i to be the values on the continuous scale, now the values are here as integers and because they are the ranks 1, 2, 3, 4 and so on, now in case if you choose x_1, x_2, x_n in the case of Pearson correlation coefficient to be 1, 2, 3, 4 up to n and so on, and similarly y_1, y_2, y_n also to be the integers 1 to n, then just computing the correlation coefficient or the Pearson correlation coefficient with the two sets of values 1 to N and 1 to n, you will get the value of the correlation coefficient and this is the idea that how the expression for the rank correlation coefficient has been obtained, well I'm not going to give you here the derivation between the two, but definitely I think that you should known, and this correlation coefficient was given by a Spearman. So, that is why this is also called as Spearman's rank correlation coefficient.

Refer Slide Time: (11:12)



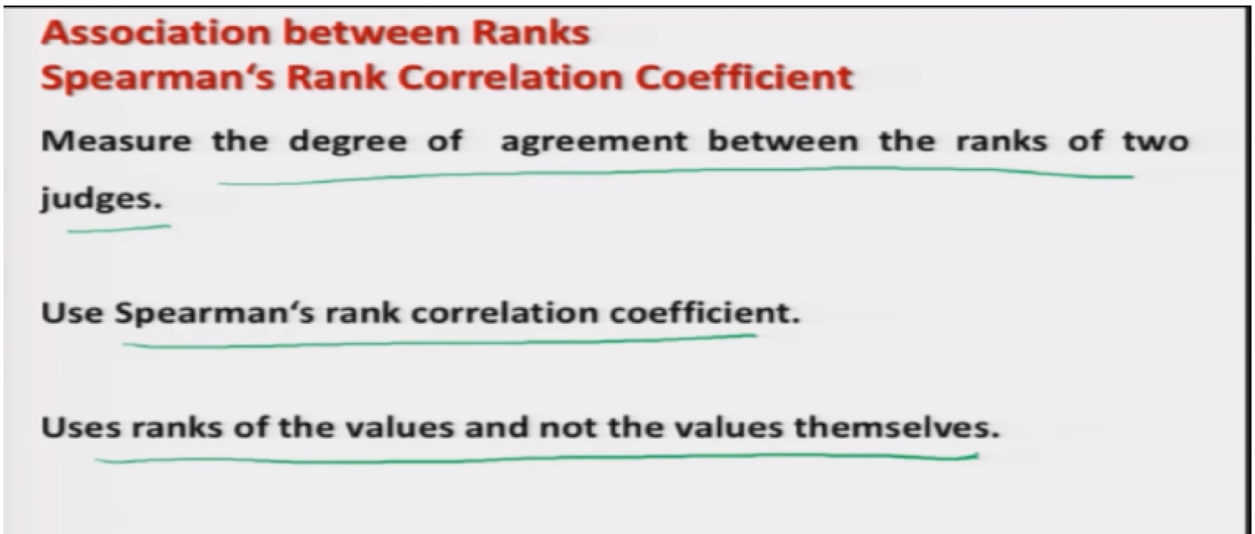
So, now let us try to continue with our discussion and now we try to understand it that how are we going to do it. So, we discuss about how to compute the experiments rank correlation coefficient. Now, suppose there are n candidates and who participated in a talent competition and suppose there are two judges and these two judges are indicated by X and Y and what these judges, they judge the performance of n candidates and they try to give the ranks to every participant or they try to give the scores which are finally converted into ranks. So, I will say here now we have a situation like this, one say judge is here X, it is trying to give us the scores and they are supposed candidates. So, you must know that how this data will look like, so that so there are 1, 2 up to here n candidates, and judge X gives them the scores x_1, x_2, x_n and then these scores are converted into the ranks. So, these ranks are going to be indicated by the numbers say here 1, 2 up to here n and similarly there is another hare judge which is here judge Y, this judge Y also gives the scores as y_1, y_2 suppose here y_n and these scores are once again converted into ranks say 1, 2 up to here n, and now we are going to find out the correlation coefficient between the two sets of observation 1 to n, definitely these numbers 1 to n, will not occur in the same sequence, for example, suppose judge X says that in his opinion the this candidate is x_2. So, he has given the rank x_2 to be here, see here n and we're suppose judge Y has given the rank to x_2, see here 3, because he think that he is only at the third order. So, you can see here that the person here is the same, same person, but he has been given two different two different ranks and then three, and then this is what we want to find that either the ranks given by the two judges whether they are similar or they are different or very different. So, that is the objective to introduce this correlation coefficient.

Refer Slide Time: (14:21)



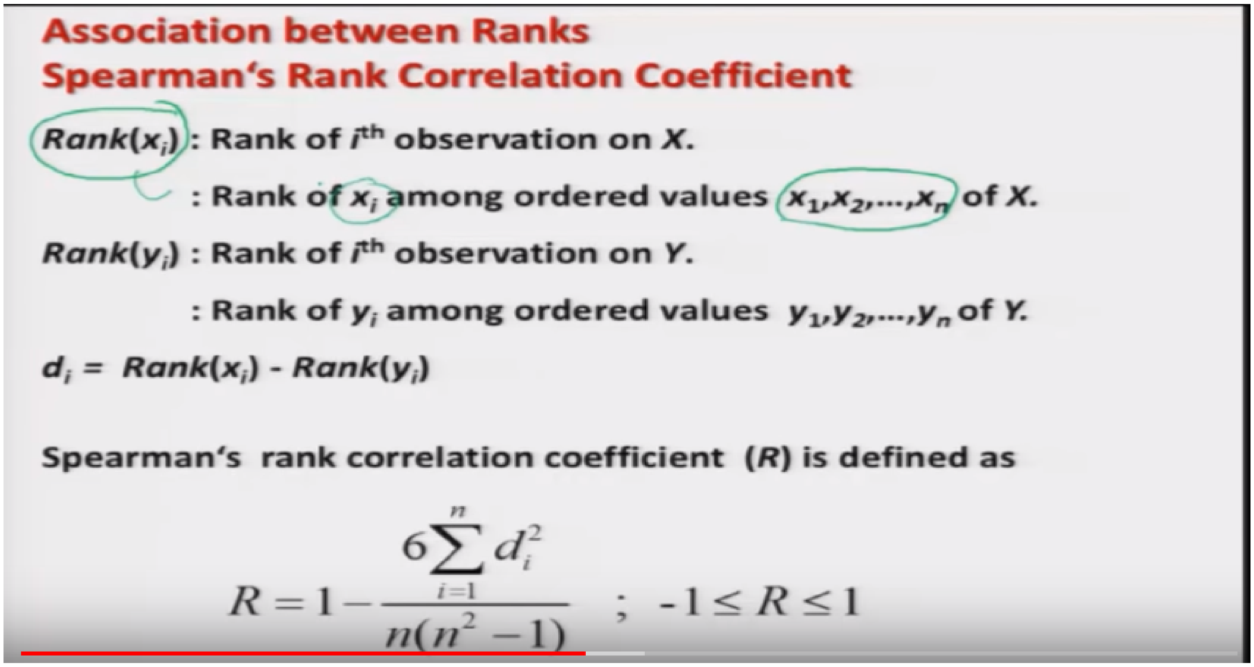
So, now if you see what I'm going to do here judge X gives ranks to the n candidates, and suppose he says that whosoever is the worst candidate he gives the rank one, and this rank 1 is given to that candidate, who had scored the lowest score and similarly, whosoever got the second lowest say score out of this x_1, x_2, x_n he has given the rank 2 and similarly whosoever is the best candidate the one who has got the highest score has been given the rank n, and similarly judge Y also has given the rank to the same n candidates, remember the candidates are same and whatever score he has given based on that he has converted the scores into ranks as 1 to n, builds on the scores y_1, y_2, y_n exactly in the same way as a judge has judge X has done it. Now, what we have to do that, now we understand that every participant has got two ranks which are given by two different judges, what do we expect? We expect from both the judges to give higher ranks to the good candidates and lower rank to the bad candidates. Now, our objective is this, we want to measure the degree of association between the two different judgments through the ranks and we would like to find out the measure of degree of Association for these two different sets of rank.

Refer Slide Time: (16:07)



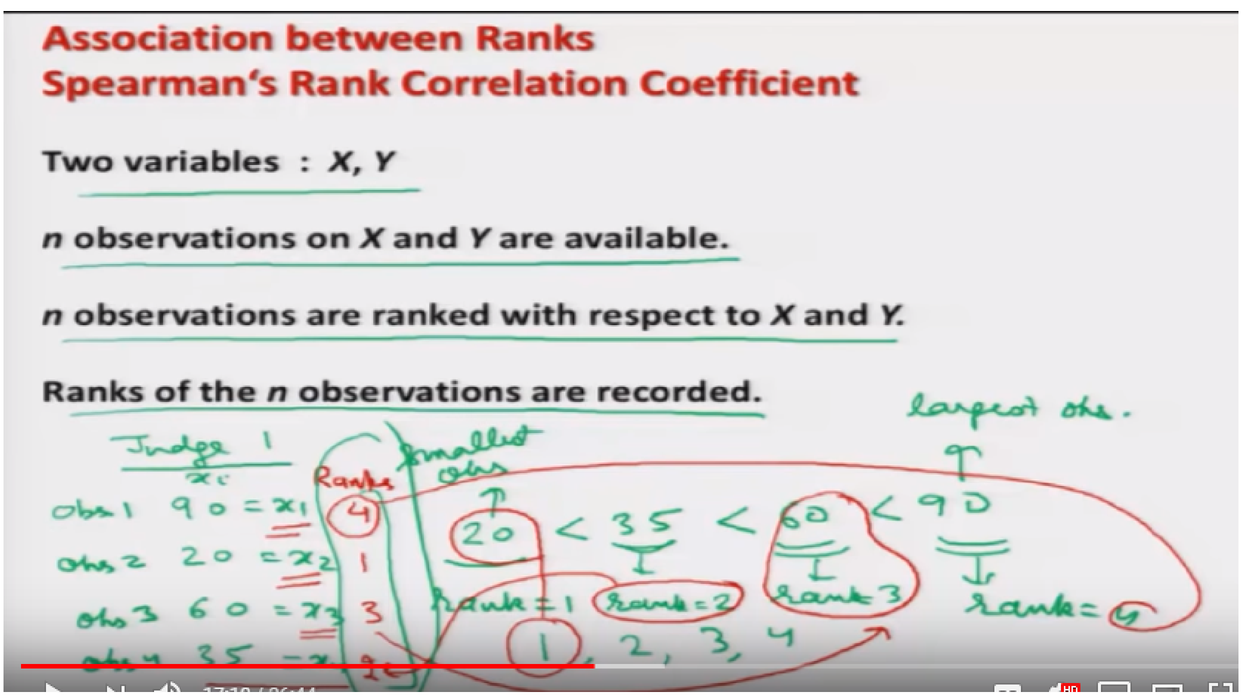
Now, the question is how to do it and what will it will indicate. So, in order to measure the degree of agreement between the ranks of two judges or two data set. In general, we use the Spearman rank correlation coefficient. So, one thing what you have to notice and always keep in mind, we are not using here the original observation, but we are using only their ranks.

Refer Slide Time: (16:35)



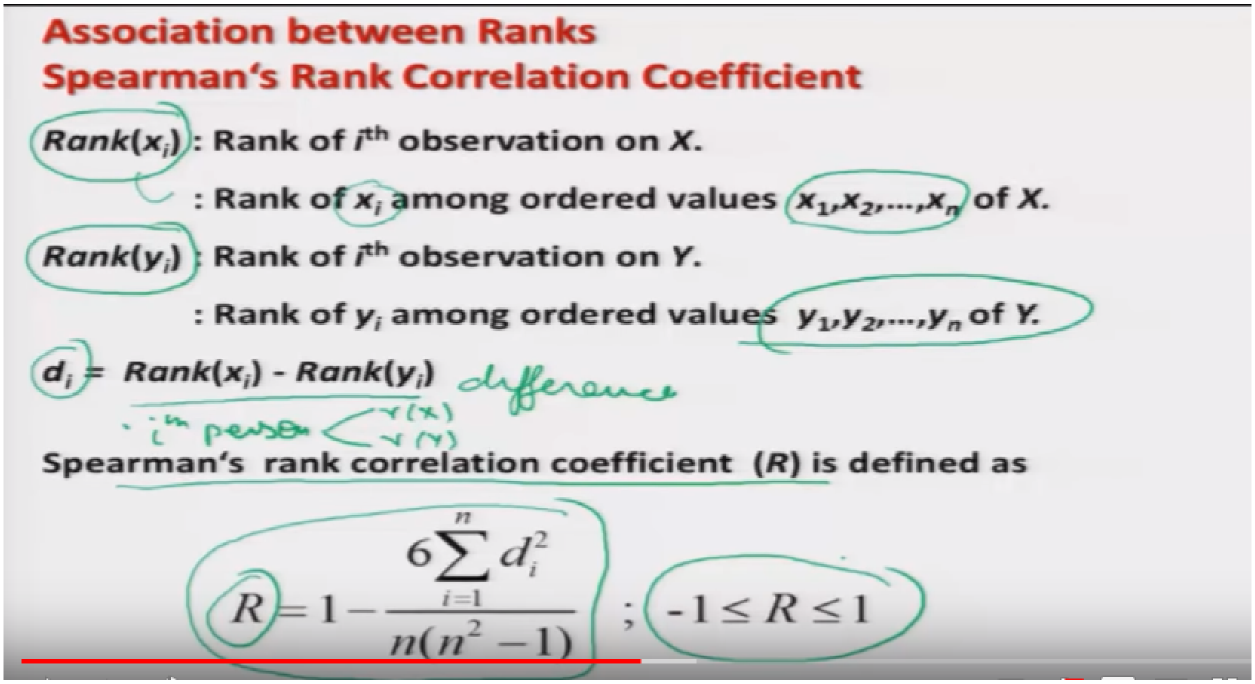
So, now let me develop or tell you that how do we do it. So, first we try to define here the rank of an observation, suppose rank of x_i, is denote as say here rank of x_i mean rank and inside the argument x_i. So, how it has been obtained that all the observation x_1, x_2, x_n have been ordered and then from there whatever is the numerical value of the rank, this has been recorded, Right!

Refer Slide Time: (17:07)



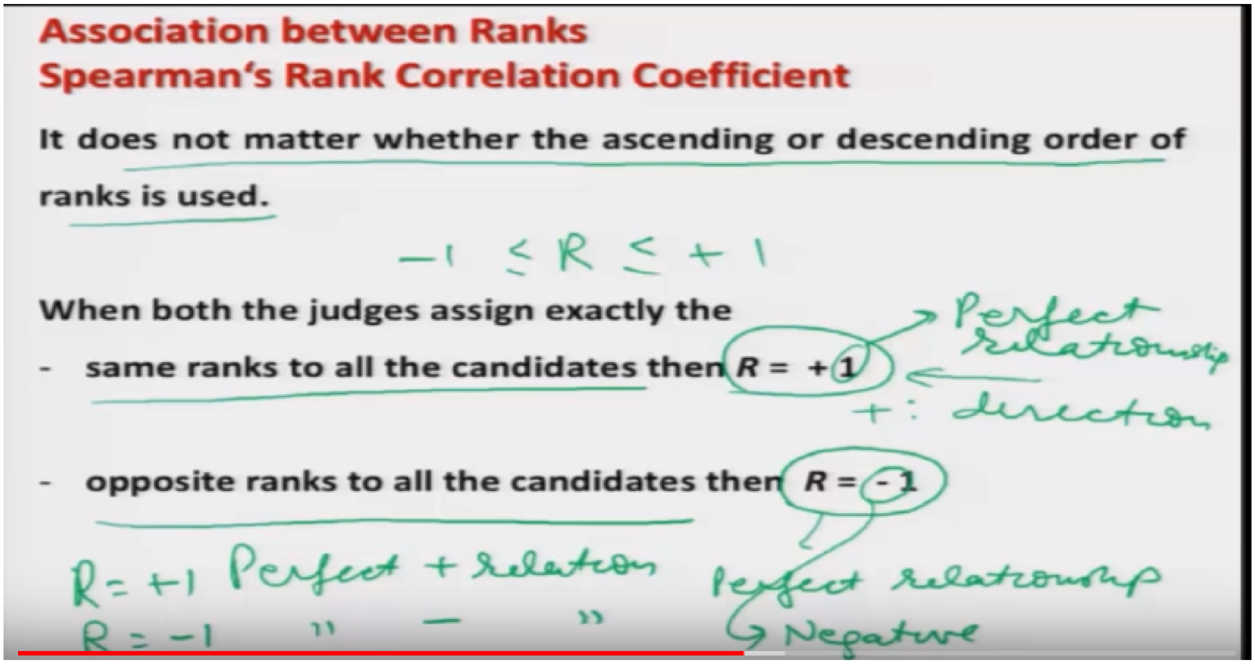
So, as I shown you here, if you see here,
I have shown you that how these ranks have been obtained, Right!

Refer Slide Time: (17:12)



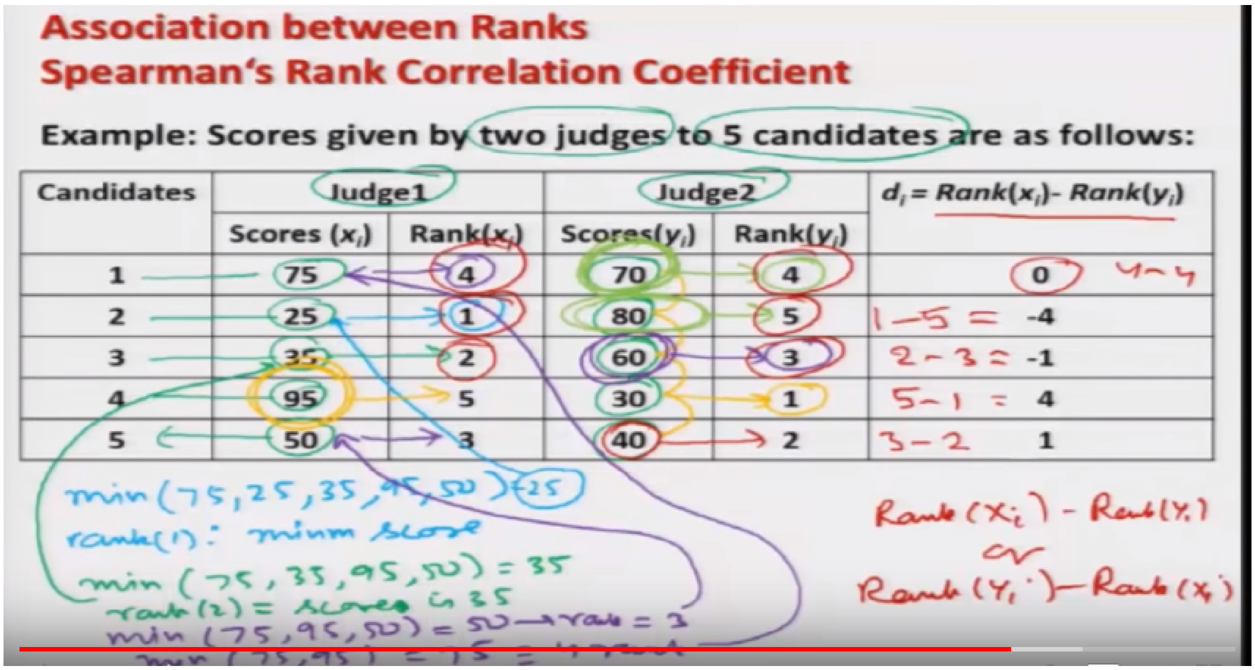
So, exactly in the same way these ranks have been obtained for the data in X, and similarly these ranks have been obtained for the data in Y and the ranks in Y, are denoted by simple statement rank of y_i. Now, what we do that I try to find out here the rank between x in y_i and we try to find out the difference, we consider the ranks of x_i and y_i and we try to find out the difference. So, what is the, if you try to see what I am doing, I try to take here the i^{th} person and I see what is the rank given by judge X and what is the rank given by judge Y and what, what is their difference I try to compute here and denoted by d_i. Now, the expression for the Spearman's rank correlation coefficient denoted as capital R, is given by this expression R is equal to one minus six summation i goes from 1 to n d_i square divided by n into n square minus 1, and this R is called as Spearman rank correlation coefficient and this lies between minus 1 n plus 1,

Refer Slide Time: (18:31)



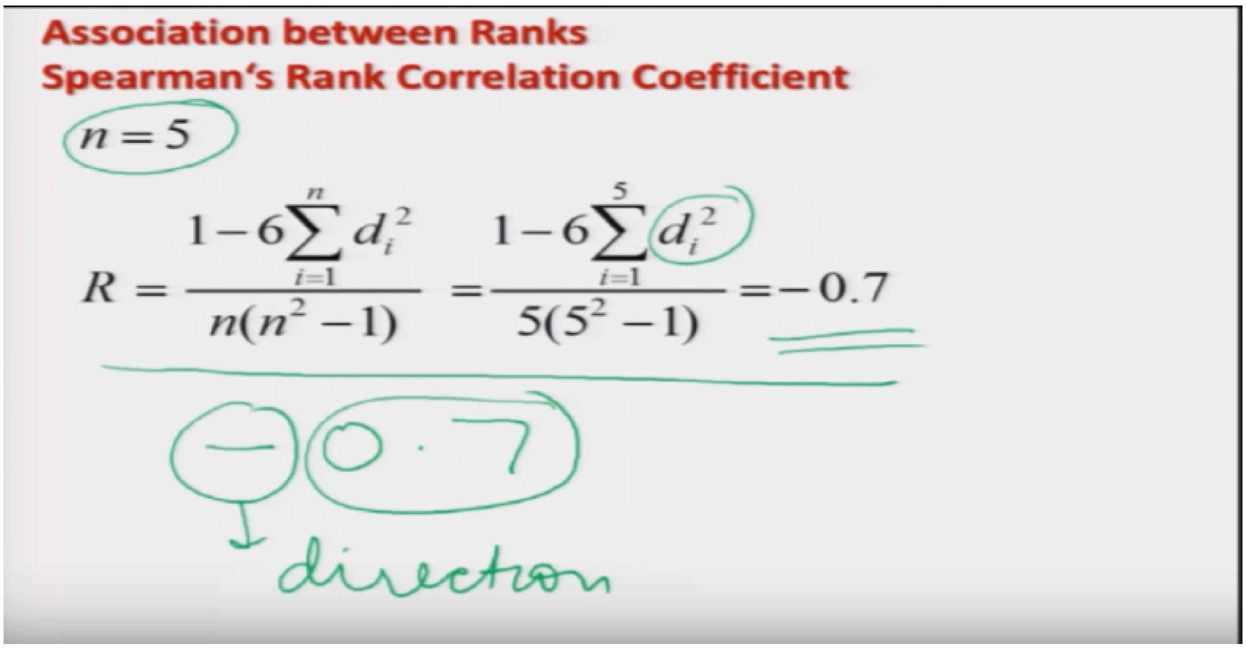
one thing which I would like to address here, that it does not matter whether the observations have been arranged in ascending order and their ranks are computed. So, what I am trying to say that if rank 1 is given to the worst candidate or to the best candidate it will not change the rank correlation coefficient provided both the judges have used the same criteria, if judge X has given the lowest rank with the worst candidate and judge Y also has given the lowest rank with the worst candidate and judge X has given the highest rank to the best candidate and same has been followed by the judge Y also, that he also has given the maximum rank to the best candidate, then in both the cases if you try to compute the correlation coefficient or the rank correlation coefficient, their values will come out to be same, but the only thing is this the ordering has to be the same in both the cases either ascending or descending. So, you give the ranks 1, 2, 3, 4 up to n or you try to give the ranks n, n minus 1 and minus 2 up to here 1, Right! Okay? Now, what is the interpretation of a, the values of R. So, we have seen that this R will lie between minus 1 and plus 1. So, when I say that R is equal to plus 1, this means that same ranks have been given to all the candidates by the to the two judges, exactly the same rank, there are n candidates and whatever ranks have been given by judge X to a particular candidate the same rank is given by the judge Y and this is true for all the candidates. So, in this case both the judges are going to give that exactly the same opinion and that is reflected by R is equal to plus 1. So, here this plus is trying to indicate the direction, direction that both the judges are given the opinion in the same direction or that direction of the opinion or both the judges is the same and one is trying to give us the value of which is indicating the perfect relationship as we have done in the case of simple collision or the Pearson correlation coefficient. Similarly, incase if the two judges give just opposite ranks to all the candidate that is another extreme, that means whoever is the best candidate in the opinion of judge X, is the worst candidate in the opinion of judge Y, then in this case the value of rank correlation coefficient will come out to be minus 1. So, this is the case of say perfect relationship once again and this relationship is negative. So, this negative sign is indicating that direction. So, R equal to plus 1 means perfect positive relation and R is equal to minus 1 means perfect negative relation, this is the interpretation and

Refer Slide Time: (21:58)



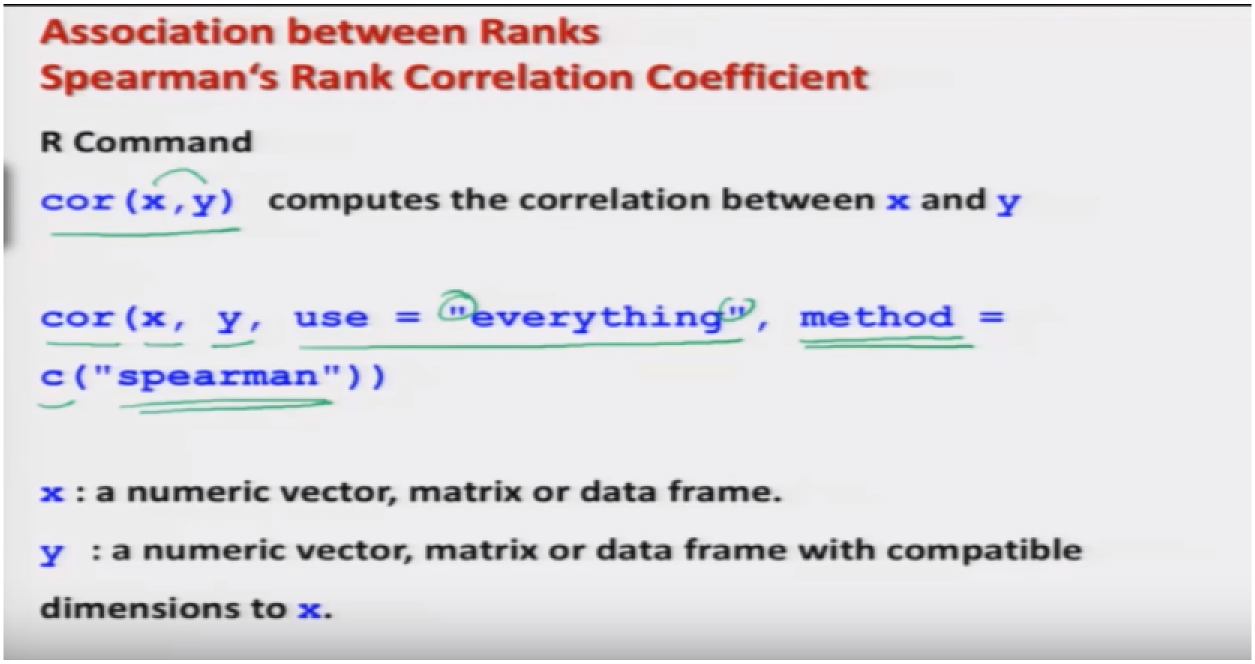
now in case if you try to choose any other value of R between minus 1 and plus 1, similar to the interpretations of Pearson correlation coefficient, we will have the similar interpretation in the case of rank correlation coefficient, for example, if the correlation coefficient is suppose rank correlation coefficient is 0.95, then I would say that, Okay. Both the judges are giving more or less the similar opinion, but definitely, if the rank correlation coefficient is supposed 0.02, then I would say well the the opinions given by the two judges are not dependent on each other, but definitely if the correlation coefficient is supposed minus 0.9, then I would say that well both the judges are giving just the opposite observations, and they have got just the different nearly the opposite opinions, whatever judge X thinks good, other judges just thinking opposite of that that is bad and so and vice versa. Now, let me take here an example and I try to show you that how to compute the rank correlation coefficient and then I will show you that how to compute it on the R software, this is important for me here to show you this minor calculations, because many times people try to compute the rank correlation coefficient not on the basis of the ranks, but simply on the basis of the of the observed values, but here you have to be careful, when you are trying to implement that the concept of rank correlation coefficient. Whenever you are given a data, try to see whether you have been given the original scores or you have been given the ranks. So, if you are given the original scores, then first you need to convert them into ranks and then try to compute the value of rank correlation coefficient, but incase if you have been given the ranks, then you can directly use them and compute the value of correlation coefficient, Right! So, in this example, I am trying to consider the scores and we will try to convert them in to ranks, now I'm taking a very simple and small example, so that I can show you the calculation, suppose there are 5 candidates and they have been just by two judges, now the judge one has given the score 75 to candidate number 1, 25 to candidate number 2, 35 to candidate number 3, 95 to candidate number 4 and 50 to candidate number 5 and similarly judge two has given 70 to candidate number 1, a score of 80 to candidate number 2, a score of 60 to candidate number 3, a score of 30 to candidate number 4 and a score of 40 to the candidate number 5. So, you can see here, we do not have the ranks. So, first we need to find out the ranks. So, first let me find out here the ranks, in the case of judge one, what is that the minimum value among all these values 75, 25, 35, 95 and 50. So, you can see here, this value here is 25. So, now I decide that my ordering will be that we will give rank equal to one to the candidate having the minimum score. So, this candidate who has got the score 25, which is here this he or she gets the rank 1. Now, once again I try to find out the minimum or maximum, whatever operation you want to do among the remaining values. So, I try to take his 75, 35, 95 and 50. So, this comes out to be 35. So, rank of 2 has to be given to a candidate who the score is 35. So, you can see here, now this is here the candidate who has been given the rank 2. Similarly, I try to compute the minimum value out of the remaining value which is 75, 95 and 50 and this comes out to be here 50 and then, I try to give rank equal to 3 to the candidate whose score was 50. So, you can see here, this I am doing it here. Now, once again I try to find out the minimum value with the remaining values which is here 75 and 95, Right! and this is comes out to be 75. So, I try to give the rank 4 to the candidate who has got 75 marks. So, you can see here this is here and this candidate has been given the rank 4 and similarly, now the value which is the maximum or the value which is left here is the maximum value, and maximum value here is 95. So, this will give the maximum rank 5 and the same operation is done on the scores given by judge 2. So, you can see here out of this 70, 80, 60, 30 and 40 values, the smallest value here is 30 and so, this has been given the rank 1. Now, after this I try to find out the second largest value second largest value here is 40. So, this candidate has been given the rank 2. Now, I try to obtain the third largest value, third largest value here is 60. So, this candidate has been given the rank 3, and similarly try to obtain the fourth largest value this is here 70, and this candidate has been given the rank 4 and finally the maximum score is here by 80, and this candidate has been given the rank 5, Right! Now, I try to find out the difference between the rank of x_i and rank of y_i actually here, I have both the options either I try to consider the rank of x_i minus rank of y_i or rank of y_i minus rank of x_i, and this will give you the same correlation coefficient Y, you can see here in the formula of this rank correlation coefficient here, we are using here d_i square. So, this sign of d_i will not make any difference. So, now if you try to take here the value here, I will try to highlight here 4 and here 4, this difference here is 0, this is 4 minus 4, second value here is, try to observe my pen this is 1 and here 5. So, this is 1 minus 5 is equal to minus 4. Similarly, next value here is 2 and 3, this is 2 minus 3 is equal to minus 1, then the fourth value here is 5 minus 1, is here plus 4 and similarly that difference between 3 and 2, which is 3 minus 2, is here plus 1. Now, after obtaining the value of this d_i is,

Refer Slide Time: (30:01)



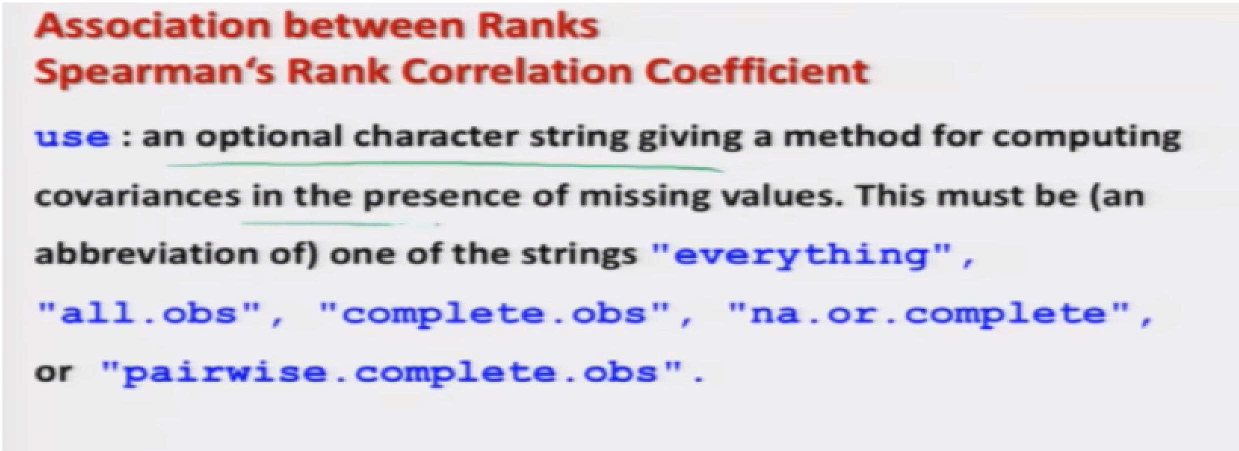
I will try to implement them on the expression for the rank correlation coefficient, which is given here. So, this d_i is we have obtained and here the number of observations are here 5. So, n is equal to here 5 and this value comes out to be minus 0.7, what is this indicating? Now, you can see here this value here is minus 0.7. So, this minus is indicating the direction direction of the ranks given by the two judges, Right! So, it indicates that both the judges have got say negative opinion and the degree of the association between them is 0.7. So, this is indicating that both the judges have got say quite different opinion about the candidates who participated, Right!

Refer Slide Time: (31:00)



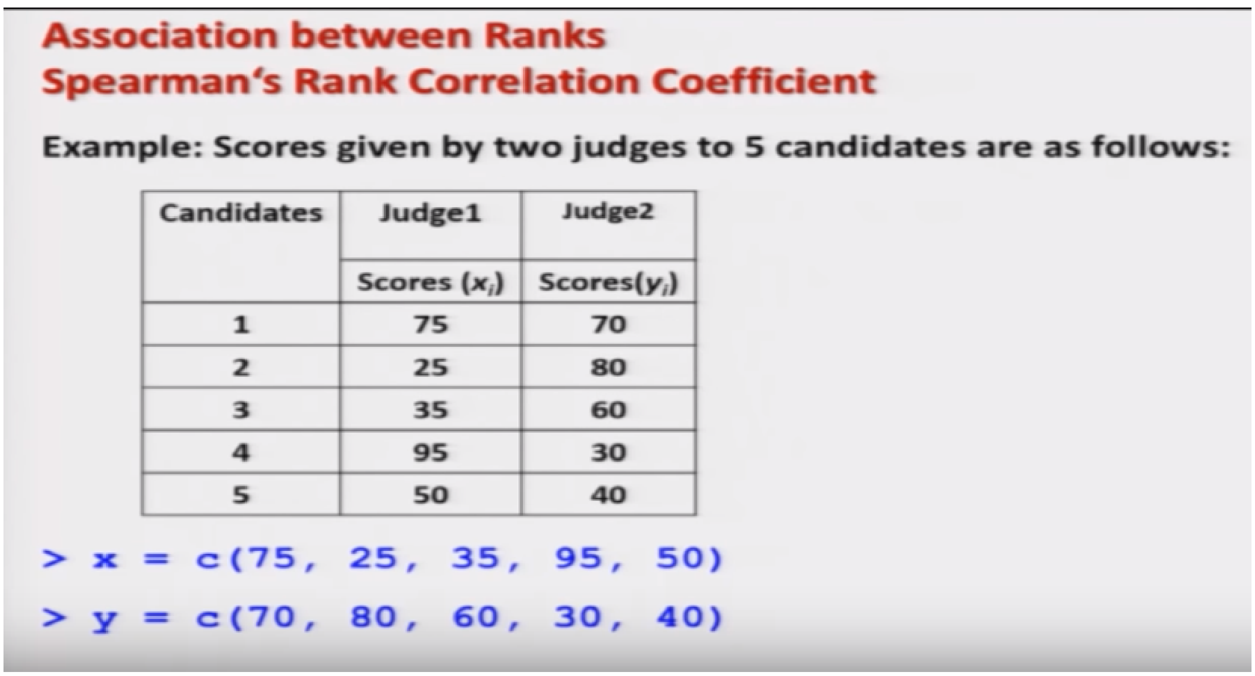
Now the p next question is, how to compute the rank correlation coefficient in the R software computation of experiments, rank correlation coefficient and the competition of Pearson correlation coefficient, they are similar. In R software, we use the same command for computing the both, the only difference is that when we try to specify the option method, then we have to be careful, if you remember when we discussed the computation of Pearson correlation coefficient in the last lecture then, then I had given the option, say method is equal to Pearson, at that time I also explained to you about the rank correlation now I am giving you more explanation that that we need simply need to specify the method is equal to Spearman, and then the entire computation procedure and all other R commands, they will remain the same they will have the same interpretation. So, now let us try to understand it here. So, the R command to compute the Spearman rank correlation is the cor and inside the argument you have to give the data vectors here x and y, and in this case there are some other options which have to be used in case if you want to compute the correlation coefficient based on rank, that is the rank correlation coefficient cor will remain the same, x and y will remain the same as the data vector, the option of use, to use all the data by giving the the command everything inside the double quotes, will remain the same as in the case of Pearson correlation coefficient, the only change will occur here that now, the method is going to be a Spearman which has to be given inside the double quote with the c command and once you try to do it here, then you will get the value of this rank correlation coefficient.

Refer Slide Time: (33:11)



So, this is about the use that handles the presence of missing value that, that is the same thing what we discussed in the case of Pearson correlation coefficient.

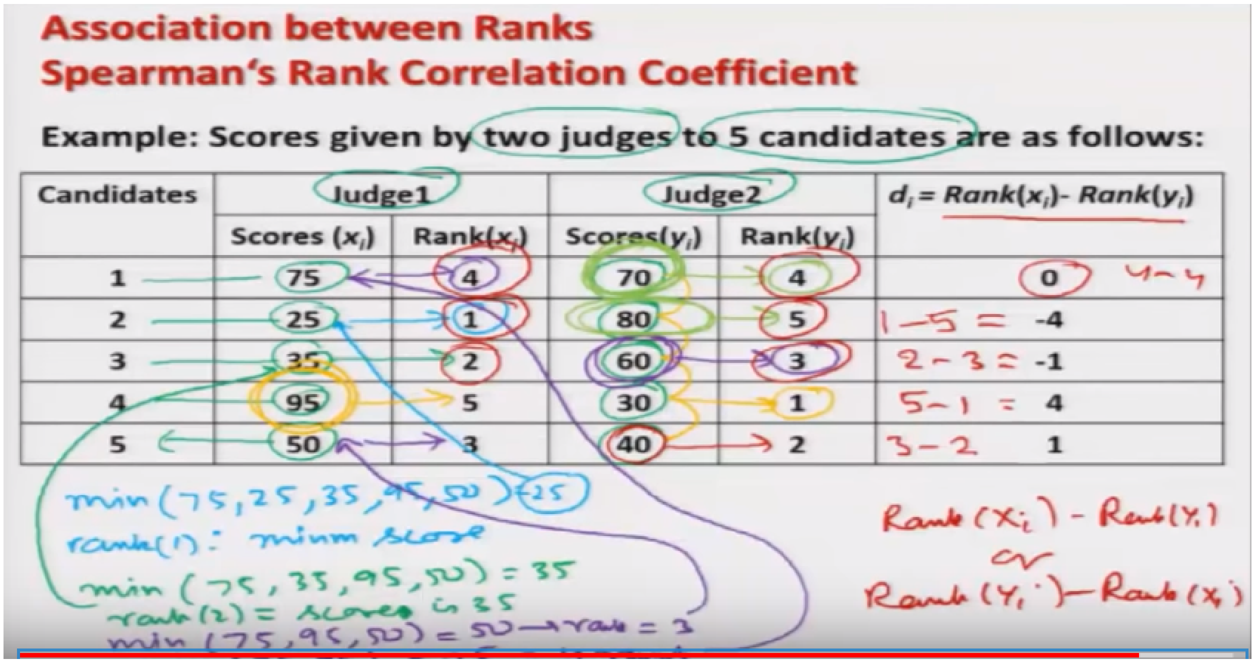
Refer Slide Time: (33:20)



Candidates	Judge1 Scores (x_i)	Judge2 Scores(y_i)
1	75	70
2	25	80
3	35	60
4	95	30
5	50	40

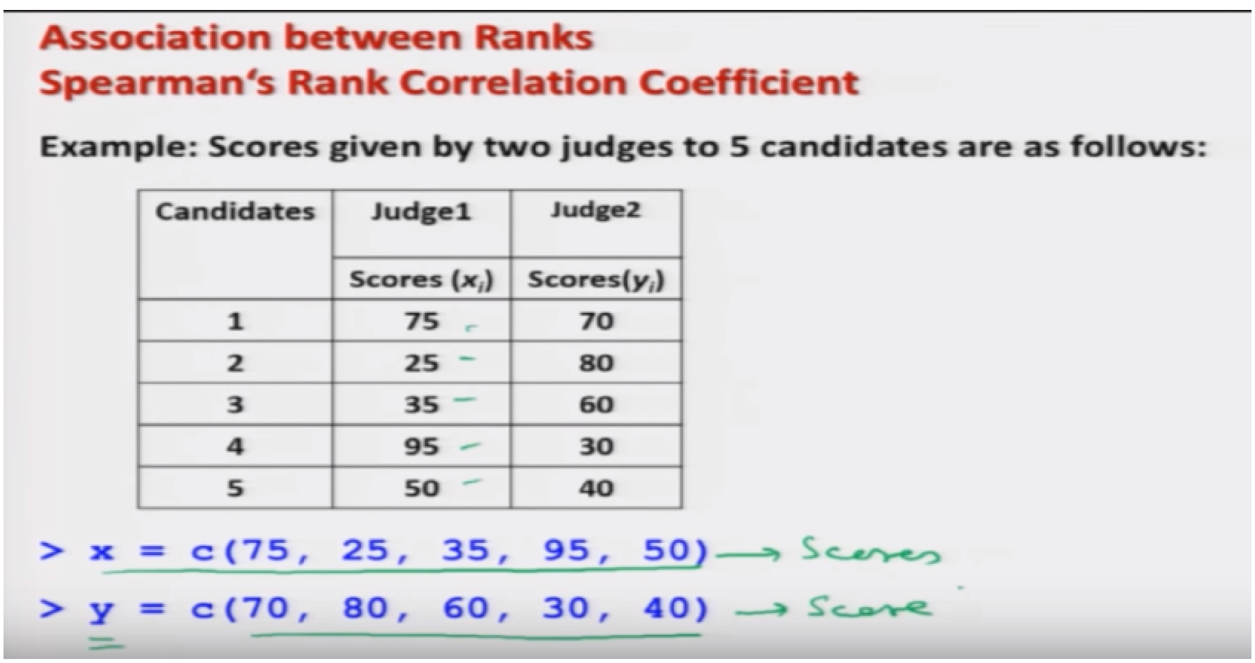
So, now I try to take the same example in which I have computed the rank correlation coefficient manually, you can see here that two values are here the same

Refer Slide Time: (33:32)



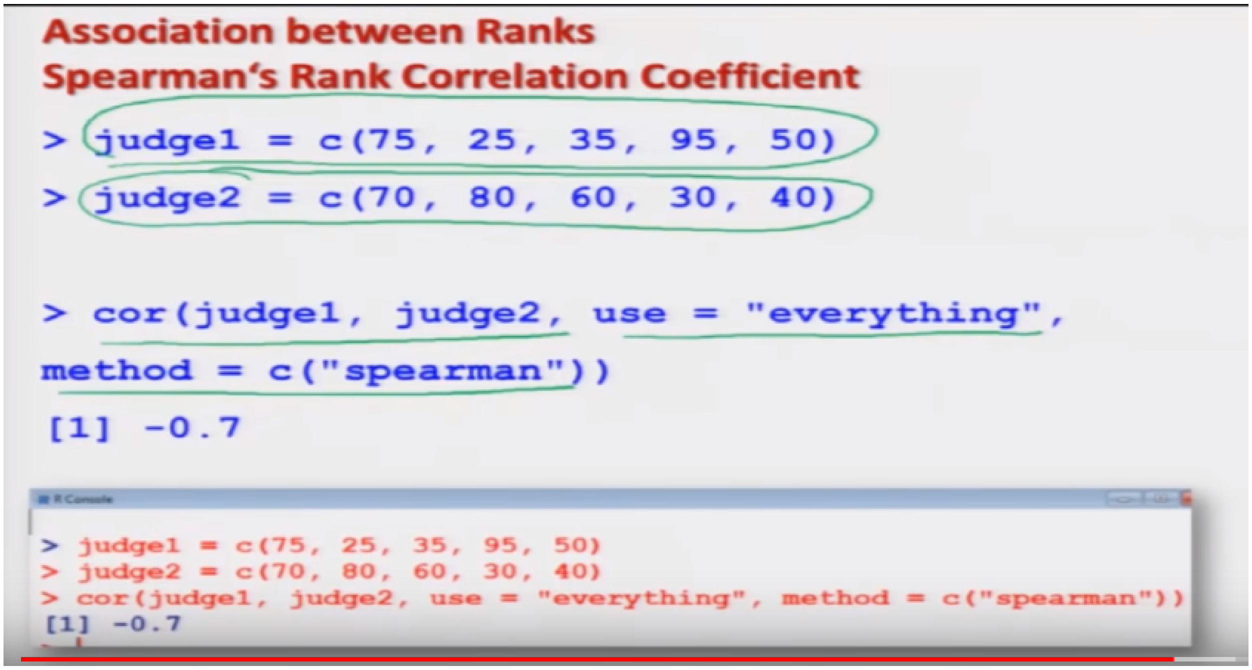
75, 25, 35, 95,50 in the first case,

Refer Slide Time: (33:40)



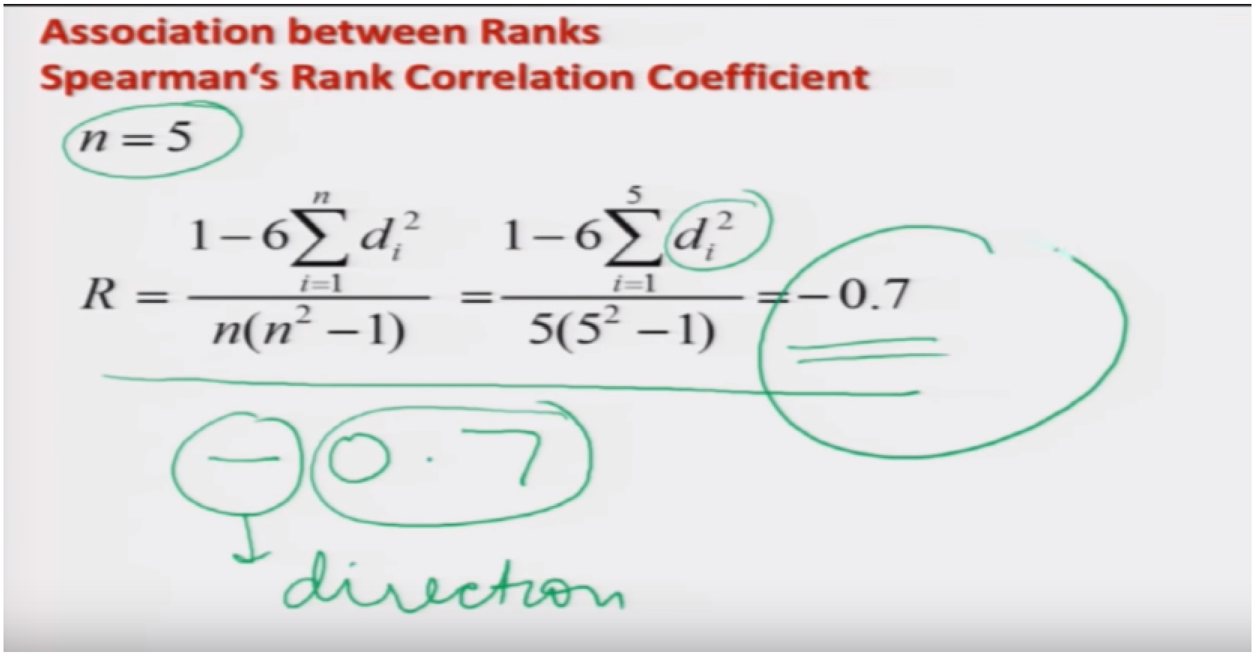
and there's 75, 25, 35, 95, 50 in the first case. So, similarly the other values are same. So, I try to give here the data X equal to the scores given by the judge 1 and Y here as the scores given by the judge 2, remember that I am not giving here the ranks, but I am simply giving this scores and R software will automatically convert them into ranks.

Refer Slide Time: (34:05)



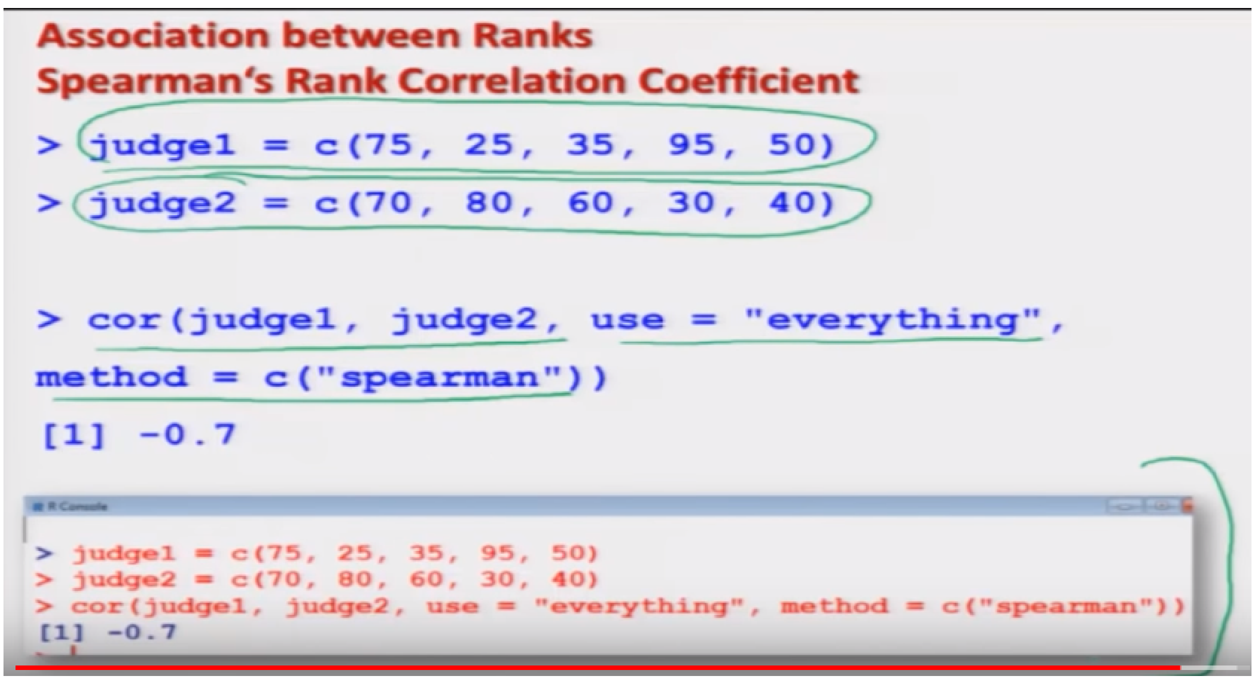
So, I try to name this data for the sake of convenience, as a judge 1 the data of given by judge 1 and judge 2 for the data given by second judge and now I try to find out the correlation coefficient between judge 1 and just to using the option everything and now my method becomes a Spearman and this you can see here, that this value comes out to be minus 0.7,

Refer Slide Time: (34:33)



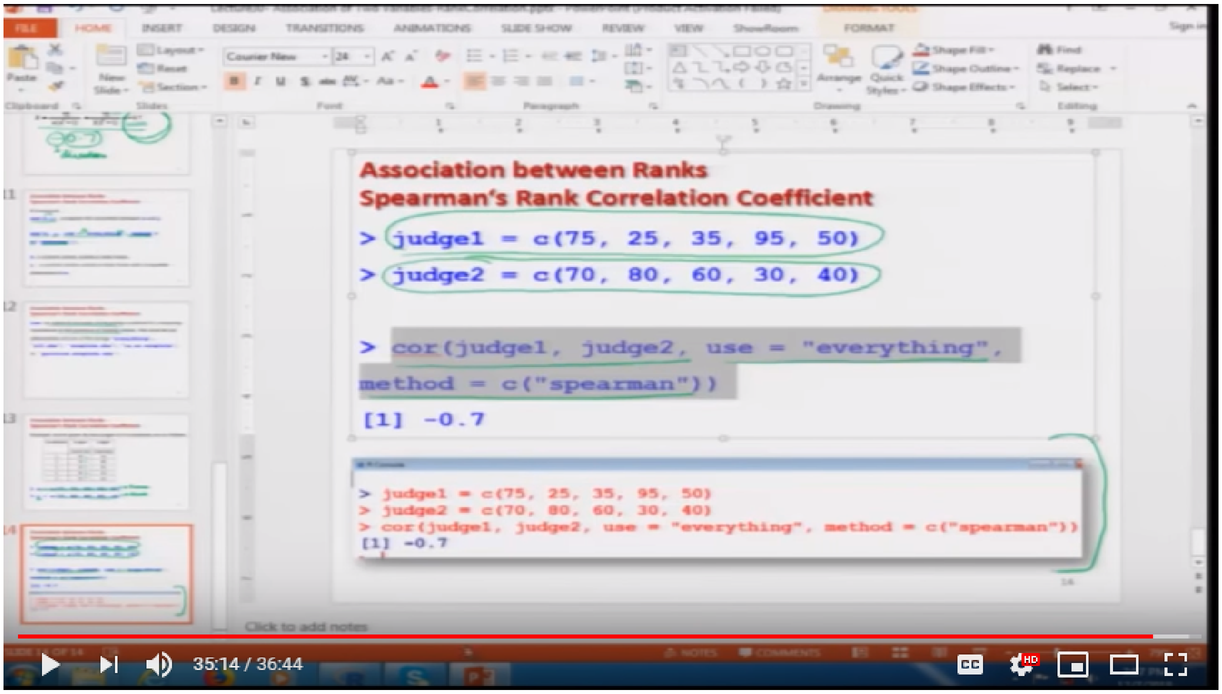
and this is the same value which you had obtained manually

Refer Slide Time: (34:37)



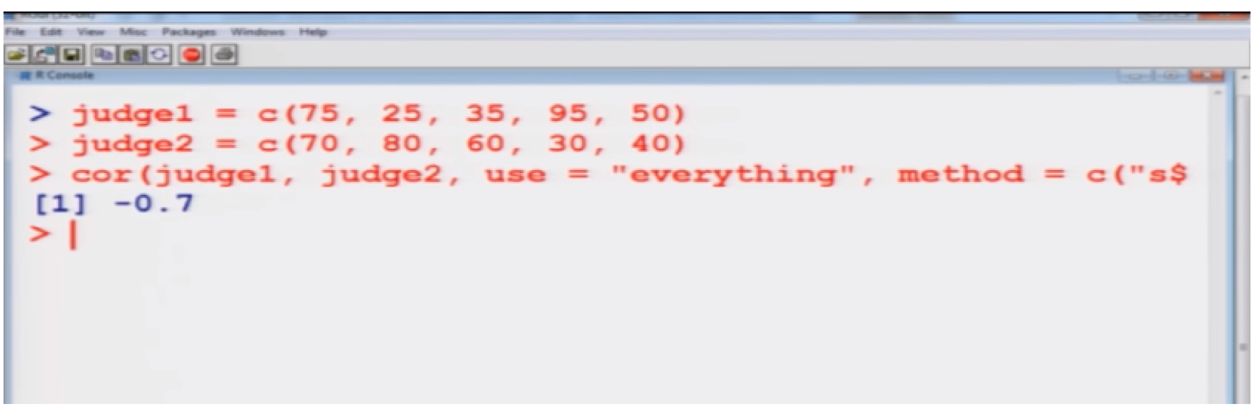
and this is here the screenshot of this operation, but I will try to show you the same operation on the R console also.

Refer Slide Time: (34:44)



So, first I try to copy the data of judge 1, see here judge 1 and the data of here, judge 2 and then I try to take the command of finding or the correlation coefficient using the method Spearman

Refer Slide Time: (35:17)



and you can see here this comes out to be minus 0.7. Now, I will stop here and now we have learned that how to compute that the Association, when the data is in the form of ranks. So, once again as usual, I will request you that you, please try to look into books, try to study more about this rank correlation coefficient, try to take some examples and try to execute them on the R software and you practice it more you practice, more you learn and I will see you in the next lecture, where we will discuss about, how to find out the or how to measure the Association, when we have counting variables, till then goodbye!