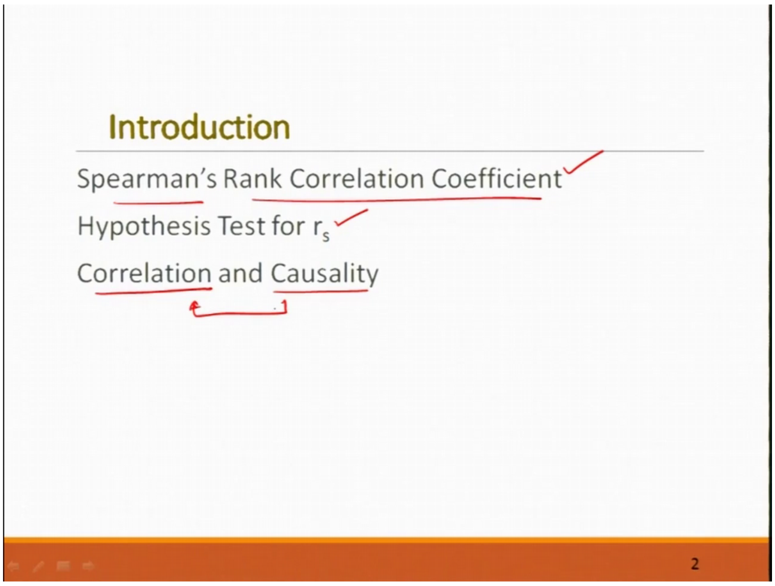
Practitioners Course in Descriptive, Predictive and Prescriptive Analytics

Prof. Deepu Philip
Dr. Amandeep Singh Oberoi
Mr. Sanjeev Newar
Department of Industrial & Management Engineering
Indian Institute of Technology, Kanpur
National Institute of Technology, Jalandhar

Lecture - 21
Correlation continued

Good morning welcome back to the course on Analytics here. So, I will continue my correlation lecture here.

(Refer Slide Time: 00:24)



Introduction

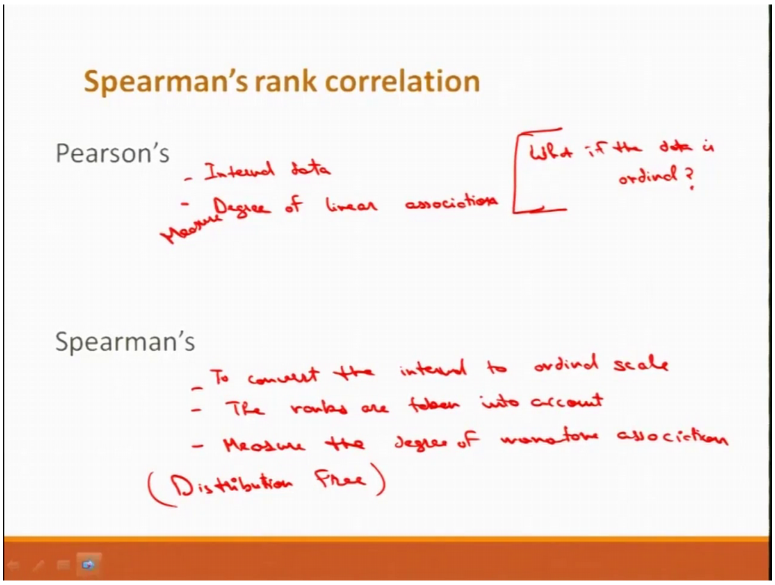
Spearman’s Rank Correlation Coefficient

Hypothesis Test for r_s

Correlation and Causality

So, previously we discussed the Karl Pearson’s product moment correlation in this lecture I will try to discuss Spearman’s rank correlation coefficient. First I will differentiate this from Spearman’s coefficient then we will test the hypothesis for r S then we will see does correlation institute causality and we will see how these related.

(Refer Slide Time: 00:53)



Spearman's rank correlation

Pearson's

- Interval data
- Measure Degree of linear association

What if the data is ordinal ?

Spearman's

- To convert the interval to ordinal scale
- The ranks are taken into account
- Measure the degree of monotone association

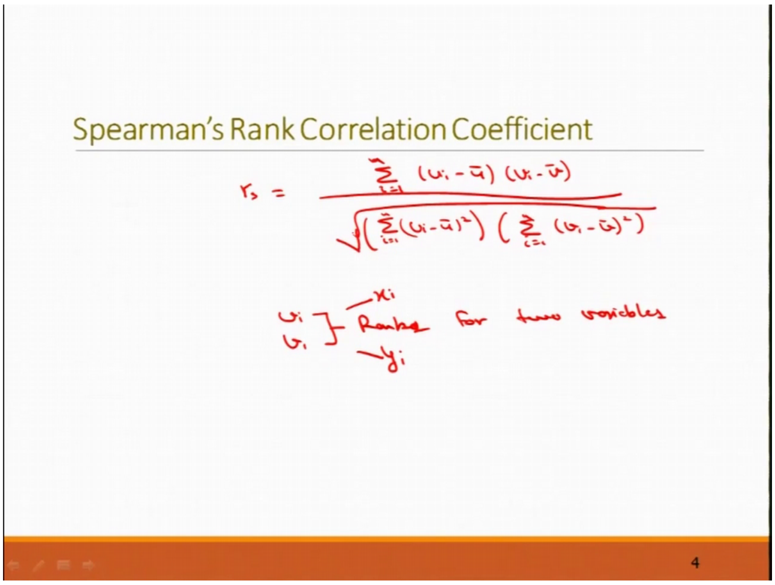
(Distribution Free)

So, first we had Pearson's product moment correlation that was based on interval data; the question comes what if the data is ordinal can we have the correlation? In that case we can compare the orders or I can say we can take the help of Spearman's rank coefficient to convert it is used to convert the interval to ordinal scale. So, when we are talking about ordinal scale; well talking about the ranks are obviously, in ordinal scale we have ranks rank 1, rank 2, rank 3 this is the order associated with this.

So, Pearson's product moment correlation this measure only the degree of linear association; does get this correlation between x or y it is based on the assumption of bivariate normality of two variables. Spearman's rank correlation takes into account the ranks and it measure the degree; here also it measure, here it measure the degree of monotone association to clarify this I will take an example or solve it before you.

So, infancies on the rank correlation are distribution free because we are just talking about the ranks; once we have given the ranks to our data, we have rank data we are put our data in ordinal scale then we do not think about the distribution and so, on.

(Refer Slide Time: 03:18)



Spearman's Rank Correlation Coefficient

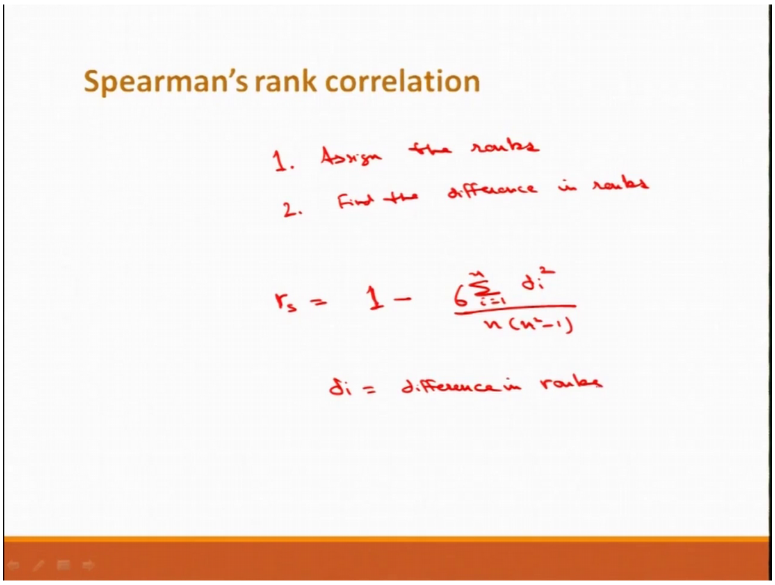
$$r_s = \frac{\sum_{i=1}^{n} (u_i - \bar{u})(v_i - \bar{v})}{\sqrt{\left(\sum_{i=1}^{n} (u_i - \bar{u})^2\right)\left(\sum_{i=1}^{n} (v_i - \bar{v})^2\right)}}$$

$\left.\begin{matrix}u_i \\ v_i\end{matrix}\right\} \underset{y_i}{\overset{x_i}{\text{Ranks}}}$ for two variables

So, how do we calculate the Spearman’s rank correlation coefficient is noted by r S; r S this is equal to summation i is equal to 1 to n U i minus U bar V i minus V bar over summation i is equal to 1 to n u i minus U bar whole square; summation i is equal to 1 to n V i minus V bar whole square.

So, what are u i and V i here; u i and V i are the ranks for two variables may be I call it a for X i this is for y i and U bar is the average rank here, V bar is the average rank for variable y and this relation tells us the Spearman’s rank correlation; so, this certain steps in this.

(Refer Slide Time: 04:54)



Spearman's rank correlation

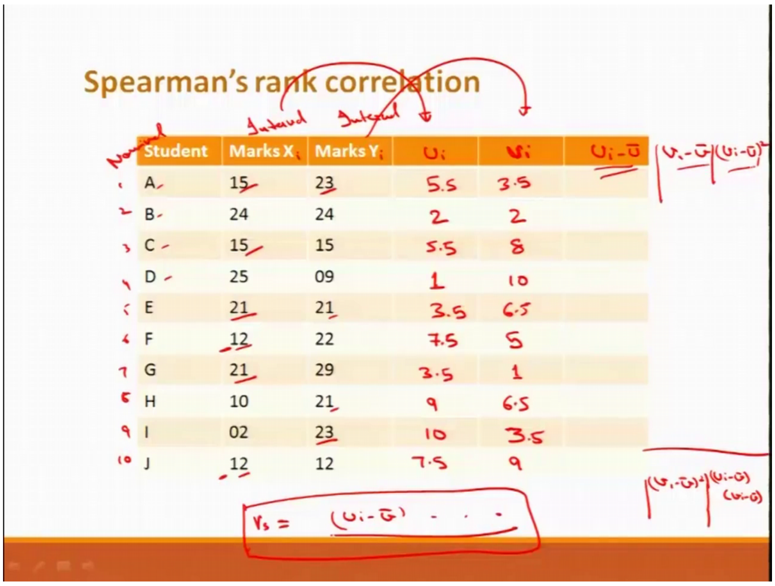
1. Assign the ranks
2. Find the difference in ranks

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

d_i = difference in ranks

First we do number 1 we assign the rank assign the rank for data then we find the difference in ranking in ranks. Assume the relation that we mentioned above we find the Spearman's rank correlation and also U i and V i are integers Spearman rank correlation can be given by this relation where V i is actually the difference in ranks.

(Refer Slide Time: 05:52)



Spearman's rank correlation

Nominal Interval Interval

	Student	Marks X_i	Marks Y_i	U_i	V_i	$U_i - \bar{U}$
1	A	15	23	5.5	3.5	
2	B	24	24	2	2	
3	C	15	15	5.5	8	
4	D	25	09	1	10	
5	E	21	21	3.5	6.5	
6	F	12	22	7.5	5	
7	G	21	29	3.5	1	
8	H	10	21	9	6.5	
9	I	02	23	10	3.5	
10	J	12	12	7.5	9	

$(V_i - \bar{V})$ $(U_i - \bar{U})^2$

$(V_i - \bar{V})^2$ $(U_i - \bar{U})$ $(V_i - \bar{V})$

$$\boxed{r_s = (U_i - \bar{U}) \ldots}$$

So, to elaborate this further I like to take an example here. So, this is the students in the class there are like we say the name of student A, B, C, D this is just a nominal scale and we have marks for the subject X and marks for subject Y these are both interval scales.

So, let me say this is my X i and this is my y i and I have the rank u i and V i which are for X i and y i respectively here. So, if I rank them based on the marks the student having maximum marks would get rank 1.

So, I think this gets rank 1 and similarly rank 2; 21 is rank 3 we have two 21. So, if you there is a tie we give the mid ranks in this case we have two 21 values 1 and 2 and ranks were 3 and 4 we will give 3.5 here and 3.5 here. And if there are three or four variables which have tie in them; then similarly we can also give the mid ranks. So, let us suppose we have 4; 21 3, 4, 5, 6 the average would be 5.5 all the 4 these 21 values would be given rank 5.5, 5.5, 5.5 and 5.5.

In this case we have this tie. So, this rank is given 3.5 and 3.5 after 21 we have next smaller number is 15. So, 3 4 goes out and we have rank 5 then we have 15 and 15 two numbers again we have a tie, then we will put 5.5 and 5.5. So, this gives rank 5 and 6 those go and we have left with ranks 7. Again we have ties 12 and 12; this 12 and this 12 then 7.5 and 7.5; 7 and 8 goes then we are left with 9 and 10 this is the rank for X i similarly would will give the ranks for Y i.

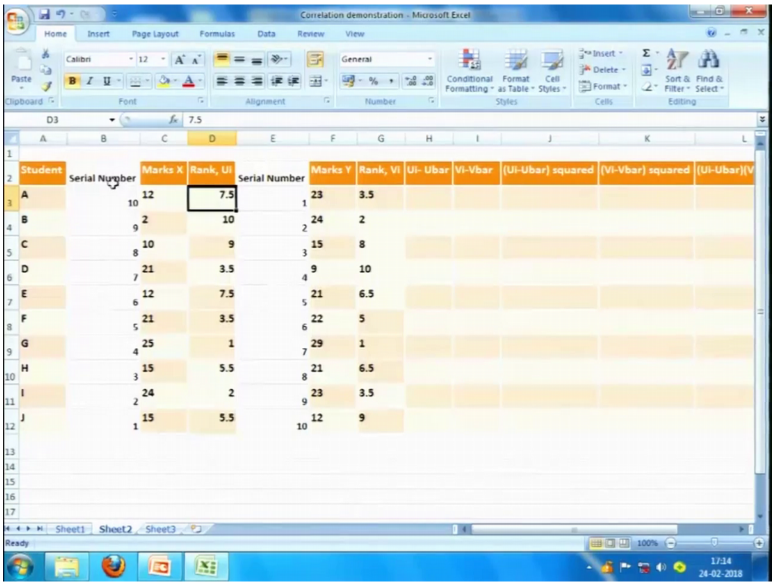
So, in this case 29 is the maximum after 29; 2 is 24, 3 is 23; 1, 2, 3 actually there is a tie 23 and 23 here. So, I will give 3.5 and 3.5 here; so, after 23 we have 22; 3 and 4 the rank is 5. Then 21 is again tie we give 6.5 and 6.5 after 21 it is 15; 7 and 8 then 9 and 10.

So, these are u i and V i then we put U i minus U bar V i minus V bar then we put u i minus U bar whole square then in the next column I will put it here; the next column would be V i minus V bar whole square and the last column would be u i minus U bar into V i minus V bar. So, I can have total sums here then I can use the relation to find the rank to find the Spearman's rank correlation to see the association here.

So, then I can use the relation that is the r S relation which we have U i minus U bar and so, on the relation is given the relation give is given in the previous slide here. So, we can use the r S relation to find the degree of monotone association. So, this is these are only 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 I just created an hypothetical data here. So, we had only 10 entries here; so, we just saw and assign the rank here.

What if we have 100 entries what if we have 1000 or more entries we can do it manually very easily. So, I will try to take this data in Microsoft Excel software and we will try to solve this we will try to do all this calculations using the software using Microsoft Excel.

(Refer Slide Time: 11:02)



Student	Serial Number	Marks X	Rank, Ui	Serial Number	Marks Y	Rank, Vi	Ui- Ubar	Vi-Vbar	(Ui-Ubar) squared	(Vi-Vbar) squared	(Ui-Ubar)(V
A	10	12	7.5	1	23	3.5					
B	9	2	10	2	24	2					
C	8	10	9	3	15	8					
D	7	21	3.5	4	9	10					
E	6	12	7.5	5	21	6.5					
F	5	21	3.5	6	22	5					
G	4	25	1	7	29	1					
H	3	15	5.5	8	21	6.5					
I	2	24	2	9	23	3.5					
J	1	15	5.5	10	12	9					

So, I have copied my data in Microsoft Excel we had student name here and we have X i and Y i

Now, I will calculate rank U i that is for X i rank that is U i then we will calculate rank for y i Y i then I will see U i minus U bar then V i minus V bar; actually this is average value U bar and V bar then I will put this U i minus U bar square; similarly I can put V i minus V bar squared then I will take the product of the differences U i minus U bar into V i minus V bar.

So, first thing is assigning ranks; so, word or excel help us to do that. So, first I will put some number here some number. So, let me cut this and paste it here I will put some number here or I can even call it a serial number. Serial number 1, 2 since the number is 1 and 2 in excel this is one kind of plus the plus the curser of you see looking at here this is selection tool. Also we have moved tool the second class this is the moved tool we conclude move this one.

The third kind of small plus is the copy tool and it will copy the format here. So, if I drag it here it will give this number 1 to 10 here; similarly I will insert a row here and copy this row and paste it here as well; now I assign the ranks. So, let me bring this rank here all it will be work.

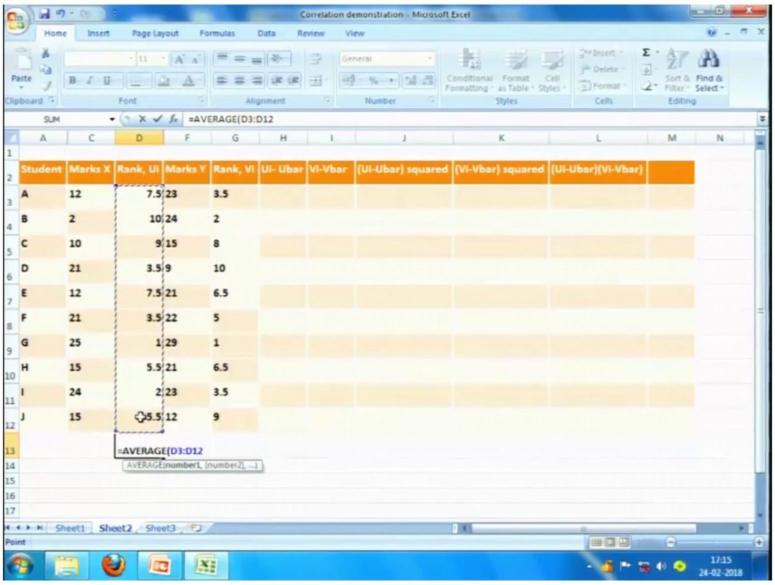
So, first let me work on the variable X what I will do I select these two columns and go to sort and filter. So, I can custom sort I can sort this based on the serial number actually I sort it based on the marks x for which I need to assign the rank. So, largest numbers is given the rank 1. So, I sorted from largest to smallest. So, it has sorted it down. So, I can assign the ranks here 1, 2, 3.5, 3.5, 5.5, 5.5, 7.5, 7.5, 9 and 10.

Now, I can bring them back to the previous order this was the purpose for which I induced this serial number here the serial number was induced bring it back to the previous order. So, I will sort it again sort custom sort based upon the serial number now we have the same marks it is in the same order the student A; 12 marks student bB 2 marks and it has this rank pretty sorted accordingly.

Similarly, we can assign the ranks to my variable Y. So, this is actually rank V i. So, let me do it again I will sort it from largest to smallest; custom sort sorted based upon the marks Y from largest to smallest ok. Then we have got largest number here number 1 rank 2, 3.5, 3.5, 3.5 oh and 5, 6.5, 6.5, 8, 9 and 10. Then I will bring it back to the original order; so as for each student for the specific student A; his marks comes correspondingly. So, let me sort based upon the serial number from smallest to largest. So, this was the use of the sort to which I did here to keep the ranks.

So, now we can hide these numbers even you continue the column if you want. So, we have got the ranks U i, V i now I can have the average here I put here is equal to when we put is equal to, it gives the formula.

(Refer Slide Time: 17:43)



SUM =AVERAGE(D3:D12

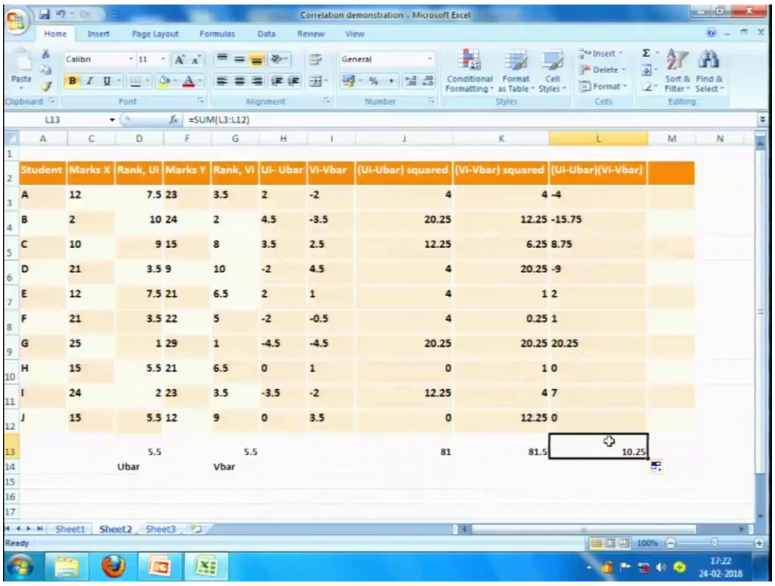
Student	Marks X	Rank, Ui	Marks Y	Rank, Vi	Ui- Ubar	Vi-Vbar	(Ui-Ubar) squared	(Vi-Vbar) squared	(Ui-Ubar)(Vi-Vbar)
A	12	7.5	23	3.5					
B	2	10	24	2					
C	10	9	15	8					
D	21	3.5	9	10					
E	12	7.5	21	6.5					
F	21	3.5	22	5					
G	25	1	29	1					
H	15	5.5	21	6.5					
I	24	2	23	3.5					
J	15	5.5	12	9					
		=AVERAGE(D3:D12							

AVERAGE(number1, [number2], ...)

So, it become the formula cell now in this formula cell I need to take the average of this value I will write a v e r a g e average I found the value a double click here now it is asking average from which to which number?

So, I will select this whole column here and press enter. So, it has averaged the ranks here U i bar similarly I will take the average for V i.

(Refer Slide Time: 18:19)



L13 =SUM(L3:L12)

Student	Marks X	Rank, Ui	Marks Y	Rank, Vi	Ui- Ubar	Vi-Vbar	(Ui-Ubar) squared	(Vi-Vbar) squared	(Ui-Ubar)(Vi-Vbar)
A	12	7.5	23	3.5	2	-2	4	4	-4
B	2	10	24	2	4.5	-3.5	20.25	12.25	-15.75
C	10	9	15	8	3.5	2.5	12.25	6.25	8.75
D	21	3.5	9	10	-2	4.5	4	20.25	-9
E	12	7.5	21	6.5	2	1	4	1	2
F	21	3.5	22	5	-2	-0.5	4	0.25	1
G	25	1	29	1	-4.5	-4.5	20.25	20.25	20.25
H	15	5.5	21	6.5	0	1	0	1	0
I	24	2	23	3.5	-3.5	-2	12.25	4	7
J	15	5.5	12	9	0	3.5	0	12.25	0
		5.5		5.5			81	81.5	10.25
		Ubar		Vbar					

So, this is equal to average V i; this is actually now U bar and V bar I can write it here, this is U bar, this is V bar. So, now, U i minus this U i minus U bar we will take the difference I will put the formula here is equal to this value U i; it is the cell d 3 minus this U bar enter; so, 7.5 minus 5.5 is equal to 2.

Now, I will copy the formula before doing that I will fix this cell U bar because this is the cell which is being subtracted from the each value here. So, to fix this in this formula I will put a dollar sign before D and 1 3. Now column D of excel sheet and cell number 13 of this excel sheet is locked column D and cell 13. So, this is this one 5.5 enter.

Now, if I copy the formula it will subtract U bar from each of the U i 10 minus 5.5 you can see 10 minus D 13; 9 minus D 5 9 minus D 13. So, you have got U i minus U bar similarly we can have V i minus V bar I can copy the formula here and I will just change the V i; V i is column G; column G, column G. So, I feel now the our formula which true here; so, this is 3.5 minus 5.5 that is minus 2.

So, in this case also this is locked V bar is locked now; now I can copy the formula. So, I have got U i minus U bar V i minus V bar, now I need to take the square of this value. Now I will put here this value raised to the power 2 enter. So, this is 2 square 4; similarly we will have if I copy the formula again to repeat to copy the formula the small plus sign would copy the formula here. So, this is 4.5 square, 2 square, this is 2 square, this is 4.5 square, this is 3.5 square and so, on.

Similarly, I can calculate V bar I will just copy the formula like this; now in this case he has it was cell H for which was it was taken square when I copy the formula I have moved it one cell forward. So, it has taken one cell forward here from it was in cell J it was H 3 square; if I move forward after H we have I it has turn to I 3 square. So, in the similar way I will copy the formula see this is I 3 square, this is I 4 square, this I 5 square and so, on.

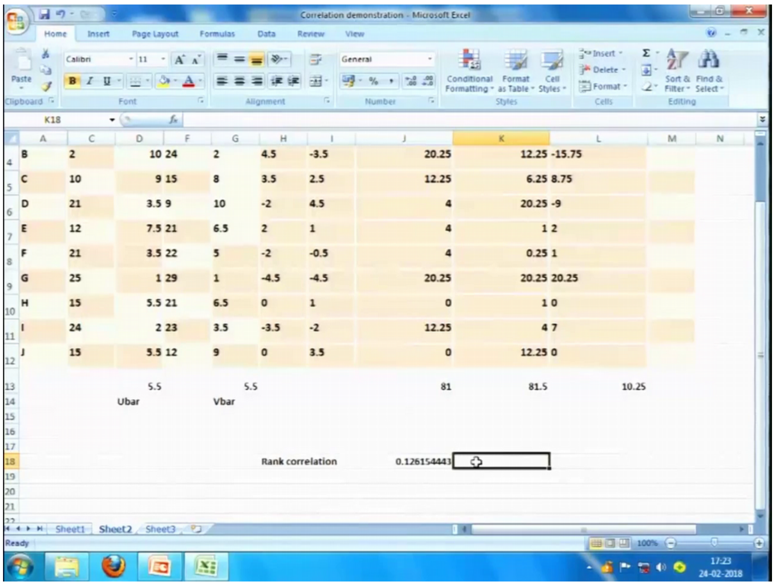
So, we have got U i minus U bar V i minus V bar squared then we need to take the product of U i minus U bar and V i minus V bar. So this would be equal to U i minus U bar multiplied by this star is multiply multiplied by V i minus V bar enter. So, we have got the product 2 into minus 2 is equal to minus 4; similarly 4.5 into minus 3.5 this value would come.

Now, in the formula if you remember we had U i minus U bar summation in the numerator. And in denominator we have the square root of the product of U i minus U bar squared and V i minus V bar squared; the product of these two taking square root is in the new denominator in the numerator we have this product.

Now this squared values always a positive value; now, the direction of the correlation would depend upon these values. Because this is the product of the negative and positive values these two values are being multiplied here. So, let me take this sum this is equal to s u m; sum formula double click sum of these enter copy the formula here it has taken the sum of K; K 3 to K 12 K 3 to K 12 12 sum is taken.

So, similarly I can copy the formula here it has taken the sum of U i minus u bar. So, since this value is positive we will have a positive correlation. Now if the correlations strong, intermediate or week that we will see this is for sure that the correlation will be positive.

(Refer Slide Time: 24:21)



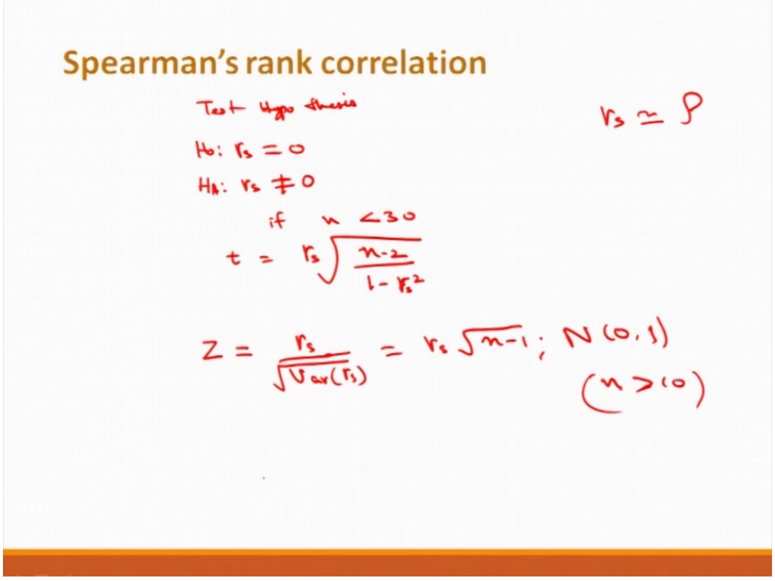
	A	C	D	F	G	H	I	J	K	L	M	N
4	B	2	10	24	2	4.5	-3.5	20.25	12.25	-15.75		
5	C	10	9	15	8	3.5	2.5	12.25	6.25	8.75		
6	D	21	3.5	9	10	-2	4.5	4	20.25	-9		
7	E	12	7.5	21	6.5	2	1	4	1	2		
8	F	21	3.5	22	5	-2	-0.5	4	0.25	1		
9	G	25	1	29	1	-4.5	-4.5	20.25	20.25	20.25		
10	H	15	5.5	21	6.5	0	1	0	1	0		
11	I	24	2	23	3.5	-3.5	-2	12.25	4	7		
12	J	15	5.5	12	9	0	3.5	0	12.25	0		
13			5.5		5.5			81	81.5	10.25		
14			Ubar		Vbar							
18						Rank correlation		0.126154443				

Now, let us see the direction of the correlation. So, this r S or rank correlation I will put it in this cell this is equal to this summation divided by the product of square root of first square root of the product of the differences this cell into this cell; so, enter.

So, this is 0.126 is my rank correlation so; that means, if you remember 0.12 value is a weak correlation, but it is a positive correlation that is the subject marks the rank the subject student is getting marks in subject 1 good marks in subject 1 is also getting good marks in subject 2, but the correlation is very weak here it is 0.12. So, this tells the monotone association between the two variables that is marks X and marks Y; we will attach this excel sheet for your help in notes save.

So, we have calculate this Spearman's rank correlation coefficient that is denoted by r S.

(Refer Slide Time: 26:16)



Spearman's rank correlation

Test Hypothesis

$r_s \simeq \rho$

$H_0: r_s = 0$

$H_A: r_s \neq 0$

if $n < 30$

$$t = r_s \sqrt{\frac{n-2}{1-r_s^2}}$$

$$Z = \frac{r_s}{\sqrt{Var(r_s)}} = r_s \sqrt{n-1}; \; N(0,1)$$

$(n > 10)$

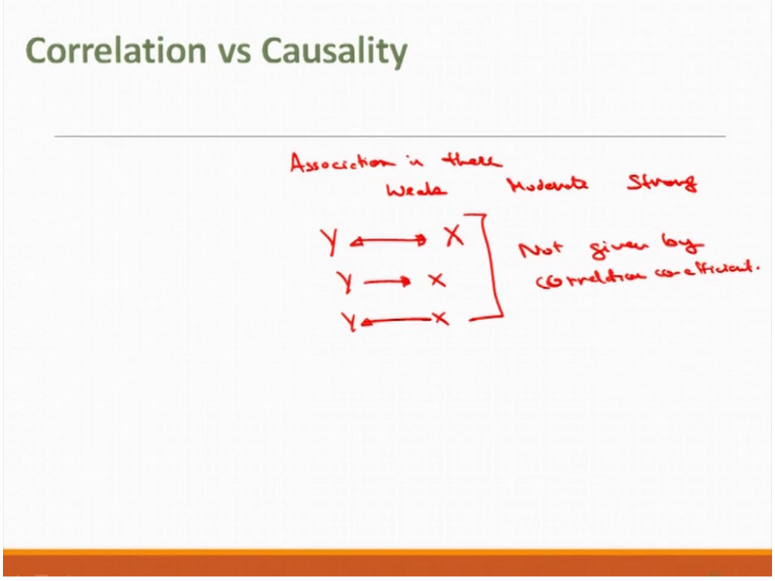
In some books we will find is denoted by rho; so, now, we will test the hypothesis. So, first the hypothesis null hypothesis that r S is equal to 0; there is no correlation between two variables that is no monotonous association.

Now, alternative has hypotheses here would be that is r S is not equal to 0. So, we calculate the value of r S and if n is less than 30; we will use a t test in which the relation for t is t is equal to r S n minus 2; 1 minus r S square. So, whole sample size is large we can also use the normal approximation that is normal deviate is equal to r S by variants of r S under root which is equal to r S into simply n minus 1; that is the normal distribution with 0 mean and then deviation 1. So, this is true if n is greater than 10 at least; so, this can be used to test the hypothesis.

Next comes a very important aspect in correlation very important topic here correlation and causality. Correlation tells that there is an association between two variables; so, it does not tell the cause. So, if I take an example instance if I say when it is raining my friend was saying when it is raining; he has an urge to have samosas with tea.

So, his consumption of samosas if I take the at that data and rainfall, if you plot the data that is I that would have a positive correlation, but does that tell that rainfall caused that urge or could I say that whenever he eat samosas, there is rainfall this look stupid. So, correlation does not tell any correlation. So, what is the independent variable what is the dependent variable that is not considered in correlation?

(Refer Slide Time: 29:03)



Correlation vs Causality

Association is there

Weak Moderate Strong

$Y \leftrightarrow X$

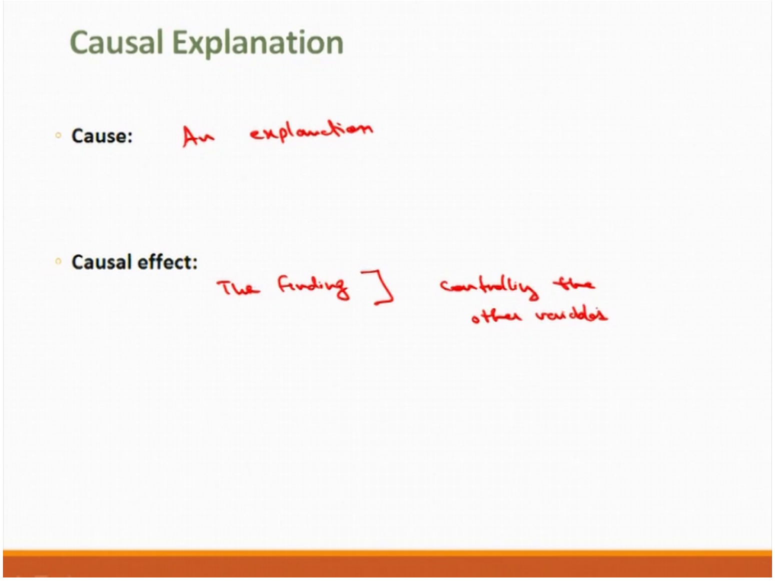
$Y \rightarrow X$

$Y \leftarrow X$

Not given by correlation co-efficient.

So, correlation tells that association is there whether weak moderate or strong, but if Y and X have association it does not tell the direction does not tell that X is being influenced by Y or Y is being influenced by X this is not given by correlation coefficient.

(Refer Slide Time: 29:51)

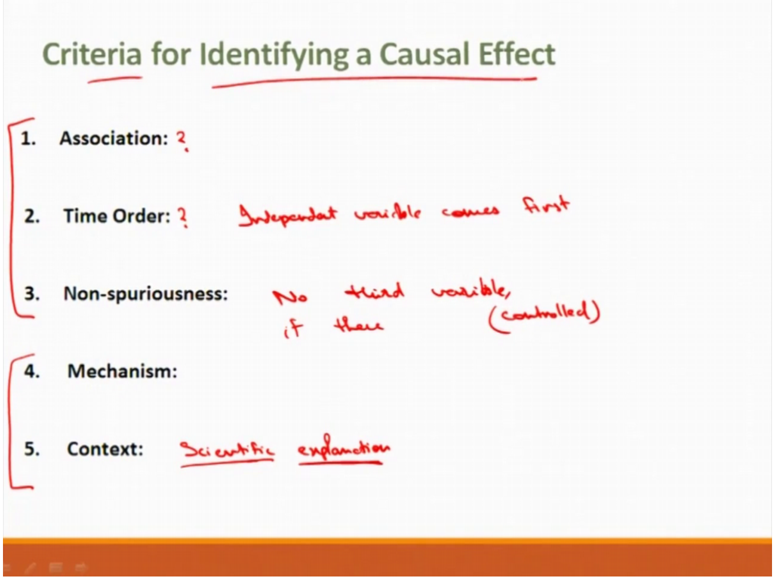


Causal Explanation

- Cause: An explanation
- Causal effect: The finding] controlling the other variables

So, what do we have? We have a cause a cause is an explanation; explanation of some characteristics, some attribute, some behaviour of groups or individuals other entities and we have a causal effect causal effect is the finding that change in one variable leads to change in another variable all other things being equal the finding controlling the other variables.

(Refer Slide Time: 30:34)



Criteria for Identifying a Causal Effect

1. Association: ?
2. Time Order: ? Independent variable comes first
3. Non-spuriousness: No third variable, if there (controlled)
4. Mechanism:
5. Context: Scientific explanation

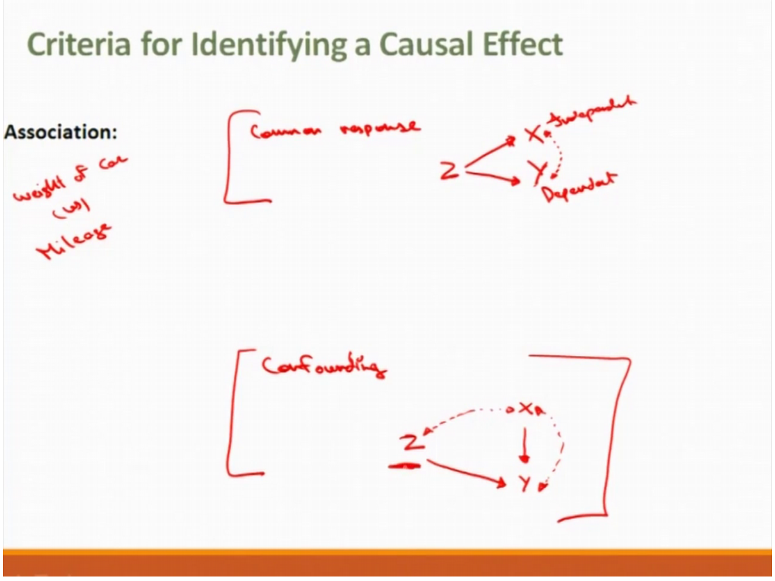
So, criteria for identifying the casual effect first thing is do we have association? So, is there any empirically observed correlation coefficient between the dependent and independent variables? Then do they have time order; time order is which variable come first? As I took example of rainfall and I used to eat samosa which is a dependent variable what do you support.

Dependent variable here would be urge to eat this spicy food. So, rainfall is my independent variable. So, please keep in mind the independent variable comes first then non spuriousness that there is no third variable or if there are third variable or the hidden variable call that the looking variables, those should be controlled; if they are those should be controlled actually these three are the required criteria.

Now, next I also have the mechanism and the context; context is the scientific explanation or if we have some theory behind the correlation coefficient which we are getting. For instance the example which I have been taking the time spent on study and the marks at the grades these are positively correlated. So, that is a theory behind that more the student read more he would grasp the knowledge and more value would be able to perform in the exams.

So, there is a context there is a scientific explanation in scientific research we have the scientific explanation. For instance in manufacturing we have some physics behind each process if there is change in roughness with change in cutting speed of the tool if tool is cutting the workpiece the there is change in roughness with the change in cutting speed there might be some physics behind that that context can support our correlation coefficient whatever we are getting. So, these are the criteria to identify the casual effect.

(Refer Slide Time: 33:09)



Criteria for Identifying a Causal Effect

Association:

So, they are two kinds of relationship one is common response and second is confounding. The common response is possibility that a change in the hidden variable is causing change in both X and Y that is Z is causing change in both independent and dependent variables. Confounding is possibility that if that the change in the explanatory variable that is the independent variable is causing changes in the responsible variable or might be that the hidden variable causing change in the response variable; in this case my Z is the hidden variable. So, actually in common response is there is the correlation between X and Y and Z is effecting both of them.

But in this case confounding my X is effecting Y; also there might be some variables Z that is effecting a Y; that is the dependent variable. So, the dash line or dash double line arrows shows an association that there is an association between two variables. This solid arrows show a cause and effect link the variable x is my explanatory or independent variable and this is my dependent variable.

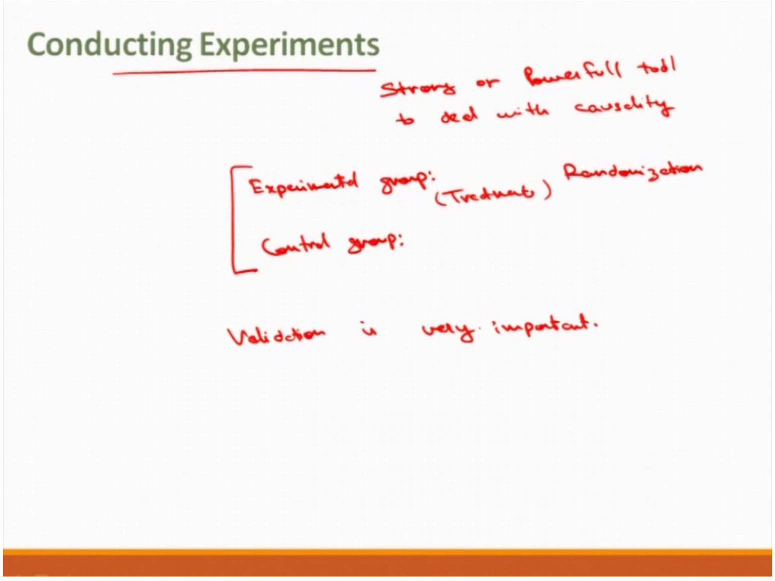
So, this correlation they could be certain examples for this ah; for instance in winter the sale of the one cloth is high I took the example before. So, that is the independent variable is my the temperature the chilling cold and dependent variable there is Y is my sale of the winter warm cloths here. So, also you can take the example vice versa here the temperature rise would increase the electricity consumption in some other electricity consumption is high because of the use of air conditions, fans and some other cooling equipment.

So, this can be an example here for common response. So, for confounding I can say there is a another variable Z that is say purchasing power that is effecting Y the sales, but it is not effecting X here. So, there is a causal relationship another example I can take the weight of the car; weight of car versus mileage. Simple causal effect would be higher the weight lower would be the mileage the more weight would be there more energy would be used to pull the car ok.

So, that is the simple causal relationship; the common response here would be that the type of the car the Z variable here would be the type of the car it can be a van, it can be a SUV, it can be a luxury car, it can be sports car, the type of the car influences the weight plus other factors that effect the gas mileage of the car ok.

Now, the confounding factor here would be the Z there could be actually the while the weight has a casual effect it actually effect cannot be accurately ascertained; since the weight is confounded with a number of factors here these number of factors can be the type of engine the horsepower and so, on.

(Refer Slide Time: 37:20)



Conducting Experiments

Strong or Powerfull tool to deal with causality

Experimental group: (Treatment) Randomization

Control group:

Validation is very important.

So, next is conducting experiments it is the most powerful way to deal with causality here when we conduct the experiments we see the dependent variable and independent variable we conduct experiments and control the other liking variables here the hidden variables or other effecting variables; we conduct experiments between them and this (Refer Time: 37:46) are result that one variable causes the changes another variable or not.

So, conducting experiment is a strong or powerful tool to deal with causality. So, to conduct the experiments first thing is we divide the, we divide our groups into two parts one is experimental group and the control group. We have the treatments here, that is the experimental manipulation or all the sets here and we do not have any treatment in the control group.

So, we should make it sure that there is random assignment of the experiments here. So, while we do the experiments; it is very important to validate what we do we actually divide our data into two parts. If we have suppose hundred observations we first do the experiments which has first analyse the 90 observations the or 80 observations keep the 20 observations for validation.

So, validation of what we do is very important. So, with this I would like to conclude my lecture on correlation we did the Spearman's rank correlation we did the demonstration on the excel sheet here, then we saw the correlation and causality and for now.

Thank you.