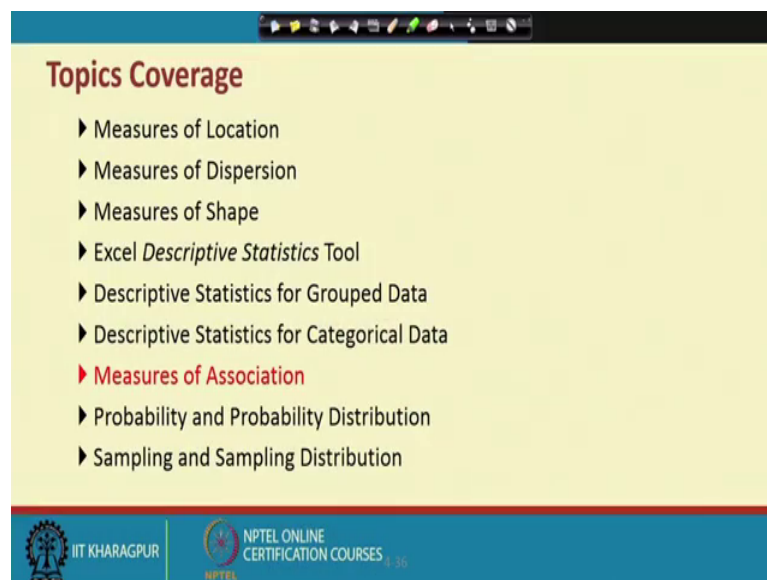


**Business Analytics for Management Decision**  
**Prof. Rudra P Pradhan**  
**Vinod Gupta School of Management**  
**Indian Institute of Technology, Kharagpur**

**Lecture - 12**  
**Descriptive Analytics ( Contd. )**

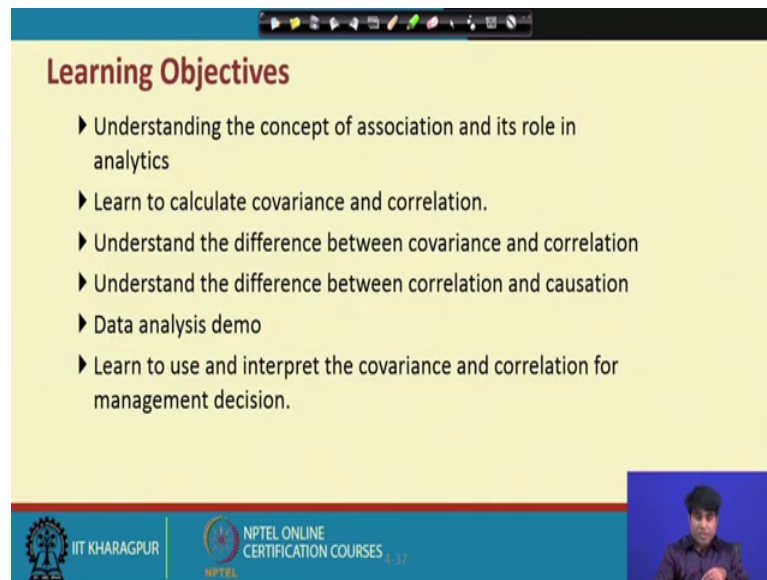
Hello everybody and this is Rudra Pradhan here and welcome you all to BMDA course and we are here to discuss the unit 3 lecture that is on descriptive analytics. In the last lectures we have already discussed something related to descriptive analytics and in fact, I have already highlighted the components that we need to discuss in this particular you know unit.

(Refer Slide Time: 00:38)



So, in the last lectures the discussions like this, measures of locations, measures of dispersions, measures of safe and excel descriptive statistic tool, descriptive statistic for group data and descriptive statistic for categorical data. So, these are the last class discussions and in this lectures we will move to a particular component called as a measures of associations.

(Refer Slide Time: 01:13)



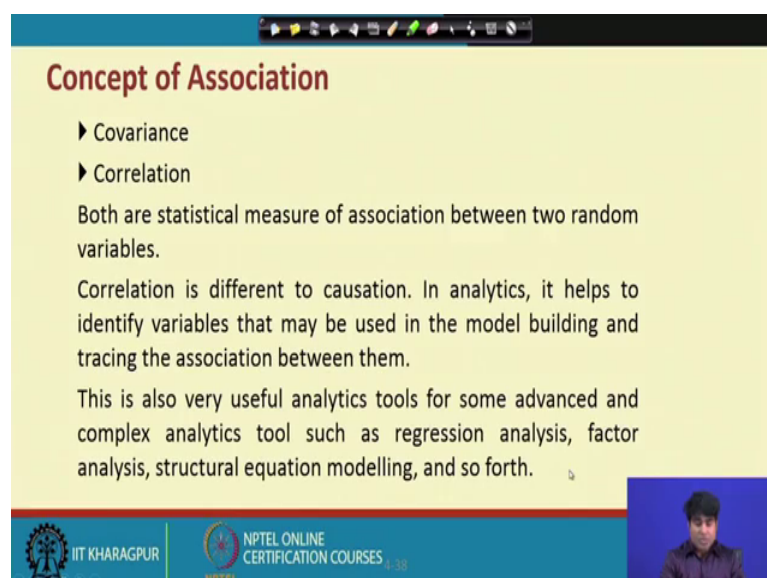
**Learning Objectives**

- ▶ Understanding the concept of association and its role in analytics
- ▶ Learn to calculate covariance and correlation.
- ▶ Understand the difference between covariance and correlation
- ▶ Understand the difference between correlation and causation
- ▶ Data analysis demo
- ▶ Learn to use and interpret the covariance and correlation for management decision.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES 4.37

So, now let me first highlight what are the objective behind this particular unit. So, in this particular lectures we try to understand the concept of association and its role in analytics we like to know how to calculate covariance and correlation and then to understand the difference between covariance and correlation and to understand the difference between correlation and causation. Then we will have some kind of examples and that to data analysis a with respect to covariance and correlations and with some kind of excel demonstration and then finally, learn to use and interpret the covariance and correlation for some kind of management decisions.

(Refer Slide Time: 02:02)



**Concept of Association**

- ▶ Covariance
- ▶ Correlation

Both are statistical measure of association between two random variables.

Correlation is different to causation. In analytics, it helps to identify variables that may be used in the model building and tracing the association between them.

This is also very useful analytics tools for some advanced and complex analytics tool such as regression analysis, factor analysis, structural equation modelling, and so forth.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES 4.38

So, with this objective, we would like to move to the kind of association analytics. So, here so far as association is concerns we must have at least two variables in the particular problem and then we must have more than, at least more than two observations. So, the requirement is, at least two variables and this data you know; that means, the size of the data should be more than two two; so obviously, there is a issue between small samples and large samples, but in the mean times you know the particular requirement is at least there are two variables and the sample size should be more than you know two variables. So, then depending upon the particular situation we will discuss what should be the exact you know sampling rule and sampling structure and then we will discuss in details about the particular structure about the covariance and correlation.

So, here, what do we have already you know mention that two variables and then corresponding to two variables we must have some kind of an adjustment with the data, but for both the variables your sample size should be uniform. If suppose there are two variables X and y, if X is having 30 observations then Y must have 30 observations, if X is having 20 observation then Y must have a 20 observation, if X is having 200 observation then Y must have 200 observation. That is the uniformity is the kind of requirement in the case of you know covariance and correlation.

And this is some kind of something called as you know beautiful technique to know the association between two and more variables and how they are related to each other and what is the nature of their relationship because their nature of relationship will give you some kind of better management decision. So, the relationship may be positive the relationship may be negatives. If the relationship between two variable will be positive then the impact will be or the strategy will be different, and if the relationship will be otherwise then again accordingly the strategy and the decision will be different.

So, now, you know we will go details about this kind of interpretation with respect to a particular problem. In the mean times let me briefing you the tools available in this particular you know associations. So, in the case of you know association these two standard techniques are called as you know correlation and then a sorry variance and correlation and both are statistical measures, association you know between two random variables.

Correlation is different to causation in analytics it helps to identify variables that may be used in the model building and trying to build the association between the two. This is also very useful analytics tools for some advanced or complex analytics a tool, such as you know regression analysis, factor analysis, structural equation modelling and something like that multiplied to try to and decision making model. So, these are the kind of importance about this particular you know association rule or association analytics.

So, now we like to discuss what is the actually the importance or where we can you know the particular techniques or tools can be huge.

(Refer Slide Time: 05:24)

**Measures of Association**

**Standard application**      **business application**

- Data from 49 top liberal arts and research universities can be used to answer questions:
- Is *Top 10% HS* related to *Graduation %*?
- Is *Accept. Rate* related to *Expenditures/Student*?
- Is *Median SAT* related to *Acceptance Rate*?

	A	B	C	D	E	F	G
1	Colleges and Universities						
2							
3	School	Type	Median SAT	Acceptance Rate	Expenditures/Student	Top 10% HS	Graduation
4	Amherst	Lib Arts	1315	22%	\$ 26,636	85	93
5	Barnard	Lib Arts	1220	53%	\$ 17,663	69	80
6	Bates	Lib Arts	1240	36%	\$ 17,554	58	88
7	Berkeley	University	1176	37%	\$ 23,665	95	68
8	Bowdoin	Lib Arts	1300	24%	\$ 25,703	78	90
9	Brown	University	1281	24%	\$ 24,201	80	90
10	Bryn Mawr	Lib Arts	1255	56%	\$ 18,847	70	84

IIT KHARAGPUR      NPTEL ONLINE CERTIFICATION COURSES

So, I will just connect with here some example here. So, this is actually data set up you know 49 different educational institutes and the idea is that, we like to establish the relationship between these variables. So, the data like to address the questions that you know, these are the following variables here. So, let me highlight what are the variables we have taken median SAT score, acceptance rate, then educational expenditure for students and top 10 percent of high schools and then graduation percentage and this is the school indications and this is the type of you know university or educations. So, we like to know. So, altogether we have actually 5 different variables. So, this is let us say X 1, this is X 2, this is X 3, this is X 4 and this is X 5 right. So, we like to know how these variables are you know relating to each others.

So, now, the you know typical questions we would like to address here is this you know each top 10 percent of HS related to graduation percentage or is acceptance rate related to expenditure per student is medians SAT related to acceptance rate like this you know there are many such you know questions can be designs and then we like to and I means through this association technique we may be in a position to answer all these questions. So, these are the typical hypothesis need to be tested and then we will get some kind of inference accordingly. This particular problem is a related to educational industry. So, accordingly a particular school or particular university will try to do some kind of strategy or do some kind of as per their requirement and that is possible or if you actually connect with this particular association techniques right.

So, let me give you the kind of structure. This is how the kind of entry points and with the help of data entry. So, we have to connect with the particular association technique then we will get some kind of output and when the basis of these outputs we will discuss the particular you know management requirement.

(Refer Slide Time: 07:54)

**Measures of Association**

- Covariance is a measure of the linear association between two variables, X and Y

$$\text{cov}(X, Y) = \frac{\sum_{i=1}^N (x_i - \mu_x)(y_i - \mu_y)}{N}$$

- For a population: `=COVARIANCE.P(array1, array2)`

$$\text{cov}(X, Y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

- For a sample: `=COVARIANCE.S(array1, array2)`

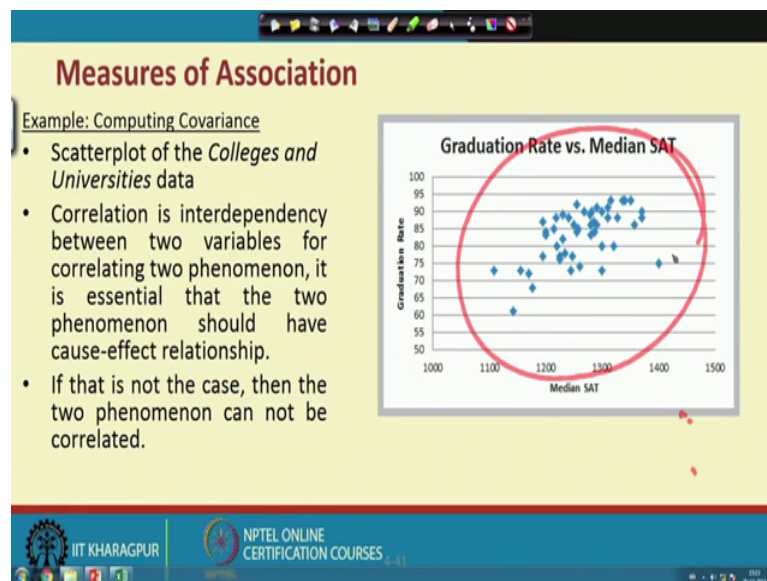
IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

And so as far as a measures of association is concerns we have already discussed there are two tools typically covariance and correlations and it is something the formula is like this. So, this is covariance formula and with the help of you know X information and Y information we can actually calculate this.

So, before you start the particular you know testing or empirical testings, we have already discussed the concept called as a population versus samples. So, most of the times you are testing empirical testing is on the basis of you know sample statistics. But if it is population then the particular you know structure is like this if it is sample specific then the particular structure will be like this, but actually the formula is more or less same, but it is only with respect to the degree of freedom adjustment or that is the that is with respect to the sample observations. And against in the last you know unit we have already discussed the excel spreadsheet and excel spreadsheet has option to report you the variance statistic and also correlation statistics. And we used to follow the particular you know you know a command to get the covariant statistics once you have actually X information and Y information and that must be very consistent right. So, on the basis of that we can actually in a position to calculate the covariance and then we will interpret as per the particular requirement.

So, now, we will go a little bit more.

(Refer Slide Time: 09:38)



So, in order to actually, since we have 5 different variables, we can go actually one by one. So, now, before we start the particular you know as association or the linkage. So, you see you can go for some kind of scatter plot and then you will find median SAT score and the graduation rate. So, there you will find these are all different kind of for

what we called as kind of plotting and the plotting itself will give you some kind of idea that there is a kind of relationships.

So, now the idea is that you know correlation or covariance is a kind of interdependence technique between the two variables, but it may not give you exactly the kind of cause and effect, but it will ensure that since there is a relationship or association. So, some cause and effect relationship is there which will study in the later case or we can examine in a later stage with respect to regression analytics. But here you know we like to check whether there is a kind of association between the two variables. If our objective is not to check which one is the cause and which one is the effect then the association itself will give you some kind of inference and that to get some kind of you may be in a position to take some managerial decision.

So, now, corresponding to these, what will you do.

(Refer Slide Time: 11:11)

**Measures of Association**

Example (continued):  
Computing the Covariance

$$\text{cov}(X, Y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

	Graduation % (X)	Median SAT (Y)	X - Mean(X)	Y - Mean(Y)	(X - Mean(X))(Y - Mean(Y))
1	93	1315	9.755	-41.850	-408.2025
2	80	1220	-3.245	-43.102	139.861743
3	88	1240	4.755	-23.102	-109.8525614
46	78	1234	-5.245	-29.102	152.637245
47	86	1250	2.755	-13.102	-36.09745939
48	91	1290	7.755	26.898	208.5964182
49	93	1336	9.755	72.898	711.1270304
50	93	1350	9.755	86.898	847.638459
51 Mean	83.245	1263.102			12641.77951
52			Sum		49
53			Covariance		263.3783221
54					
55			COVARIANCE.S		263.3783221

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, this is what actually the particular structure in the excel sheet you can calculate and then you can actually you know do some kind of processing to get the correlation coefficient. So that means, see here, this is what actually the graduation in percentage and this is the medians course which this is with respect to the previous plotting and you see here what I have already mentions that you must have uniform observations for the both the variables. So, this is the data points and exactly the same data points are here in



the second column so that means, it is very much consistent this is the first requirement and first check before you use this particular analytics.

So, now the second step of this particular process is to find out the mean and this is your mean of the second variables, then we like to find out the difference you know mean from the original observation or actual observation that is with respect to X and then this is with respect to Y. So, then you will find out some kind of deviations of Y and this is also deviation for you know with respect to Y then you we have to just multiply the difference between the two variables and then we will report here is the you know you know combine the impact. Then this is for first observation and it will continue for the rest of the observation. Then finally, you take this sum, now once you get the sums then you have to count how many numbers are there means observations are there and that will adjusted with the degree of freedom. So, this sum divided by the sample observation that is the degree of freedom n minus 1 and then you will get the kind of covariance statistics and that will give you some kind of inference about the particular problems right.

So, I will take you to the excel sheet and let you know how it is actually happening, all right.

(Refer Slide Time: 13:15)

**Measures of Association**

- ▶ Correlation is a measure of the linear association between two variables, X and Y.
- ▶ Correlation Coefficient formulas:
  - For a population:  $\rho_{xy} = \frac{\text{cov}(X, Y)}{\sigma_x \sigma_y}$
  - For a sample:  $r_{xy} = \frac{\text{cov}(X, Y)}{s_x s_y}$  CORREL=(array1, array2)
- ▶ The Correlation Coefficient does not depend upon units of measurement (unlike covariance).
- ▶ Also known as the:  
Pearson product moment correlation

So, now, this is actually we have already discussed and corresponding to covariance and this is a correlation kind of structures and both are actually same structure or same kind



of association rule, but the thing is that you know correlation is a better technique than covariance because it is the unitless measurement technique and corresponding to covariance. So, the upper part of this, upper part of this correlation is nothing, but called as a covariance which we have already discussed in the last slides.

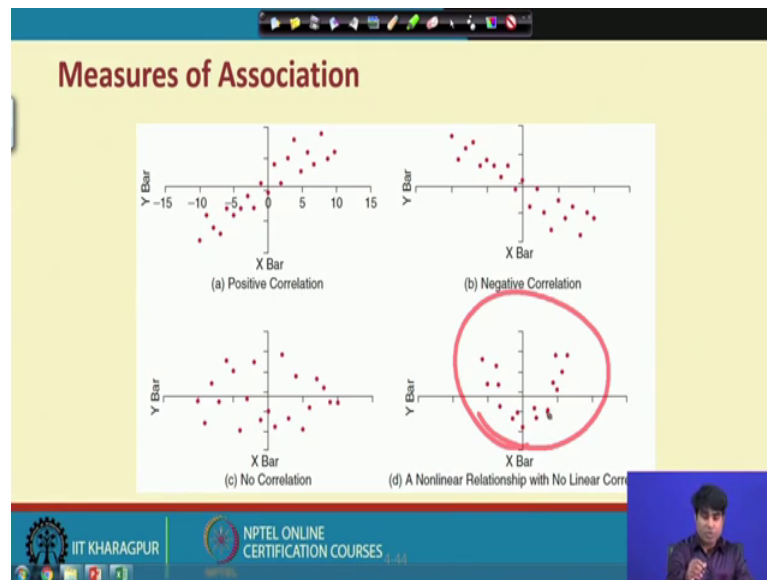
Now, if we will divide with the standard deviation of X first variable and standard deviation of Y that is the second variable then you will get a component called as correlation coefficient. So, now when you calculate correlation coefficient then this time your sample adjustment will be by default you know you know adjusted. So, as a result, you have to just calculate you know covariance of X Y then variance of X and variance of Y or that is nothing but you know square root of variance that is standard deviation of X and the standard deviation of Y. So, this is this is the typical formula for sample and you have to calculate actually covariance first then report the standard deviation of X and standard deviation of Y. So, you will get a kind of component called as a correlation coefficient.

Corresponding to the covariance correlation coefficient is very standardized and the values of the correlation coefficient usually between minus 1 to 1. So, if it is minus 1 that is the indication is the perfectly negatively correlated, if it is 1 perfectly positively correlated, if it is 0 then there is no correlations or there is no association.

So, a higher the movement towards the towards you know for 1 then that is you know high positive correlation and if the correlation value will be close to minus 1, then it is it is you know the indication is that it is having high negative correlation between. If it is close to 0 then it will be having low correlation or if it is exactly 0 then there is no (Refer Time: 15:21). This is how the interpretation is all about. So, because your objective is to know what is the association between the two, so you have to, we have to find out what is the percentage of you know association between the two. So, the correlation coefficient then that can be transferred into actually 0 to 100. If you square it then; obviously, then and multiplied by 100 then this will be you know indicated with respect to 0 to 100 scale right.

So that means, say what is the degree of association between the two variables, if it is say 0.8 then 80 percent if it is 0.9, 90 percent and like that you have to interpret right. So, in this case.

(Refer Slide Time: 16:08)



So, the association structure will be like this sometimes there is a positive sometimes there is a negative and sometimes there may be no relationship that is when that is the case when you will get  $r$  equal to 0 and a positive correlation when  $r$  will become positive value when negative correlation, when  $r$  is having actually a negative value. And there may be some kind of non-linear relationship and the picture will be the non-linear linear shape will be like this. So, it will have some kind of spreading like you know exponential or something like that so that means, it will give you some kind of non-linear relationship between the two variables. But whether it is a linear relationship or non-linear relationship, so there the idea is to just to check whether there is association between these two variables. So, that is how we are here to discuss the covariance and correlation tool.

(Refer Slide Time: 17:10)

## Measures of Association

Example:  
Computing the  
Correlation  
Coefficient  
*(Colleges and Universities data)*

► *Graduation % and Median SAT*

$$r_{xy} = \frac{\text{cov}(X,Y)}{s_x s_y}$$

	A	B	C	D	E	F
		Graduation % (X)	Median SAT (Y)	X - Mean(X)	Y - Mean(Y)	(X - Mean(X))(Y - Mean(Y))
1						
2		93	1315	9.755	51.890	506.269875
3		80	1220	-3.245	-43.102	139.8617243
4		88	1240	4.755	-23.102	-109.8525614
47		86	1250	2.755	-13.102	-36.09745939
48		91	1290	7.755	26.898	208.5964182
49		93	1336	9.755	72.898	711.1270304
50		93	1350	9.755	86.898	847.698459
51	Mean	83.245	1263.102	Sum		12641.77551
52	Standard Deviation	7.449	62.676	Count		49
53				Covariance		263.3703231
54				Correlation		0.564146827
55						
56				CORREL Function		0.564146827

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES 4-45

So, this is how the particular structures which we have already in discussed and then I will take you to the particular you know, let me take you to the particular examples.

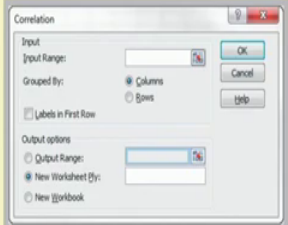
(Refer Slide Time: 17:24)

## Measures of Association

Excel Correlation Tool

► *Data*  
 ► *Data Analysis*  
 ► *Correlation*

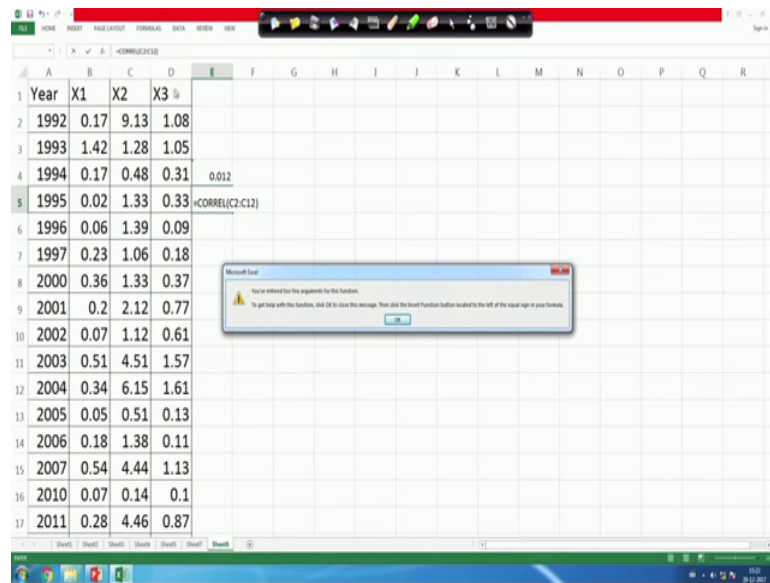
Excel computes the correlation coefficient between all pairs of variables in the *Input Range*.  
*Input Range* Data must be in contiguous columns.



IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES 4-46

So, what will we do? So, let us let us go to this particular you know spreadsheet.

(Refer Slide Time: 17:36)



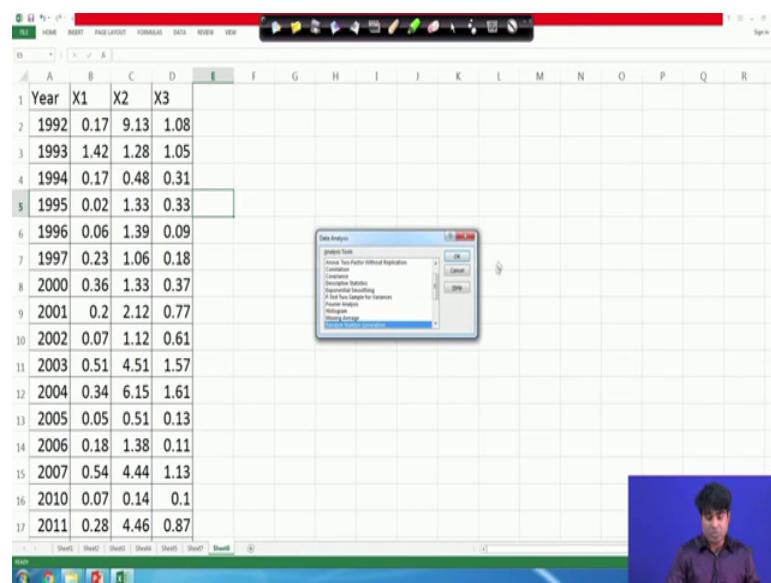
Year	X1	X2	X3
1992	0.17	9.13	1.08
1993	1.42	1.28	1.05
1994	0.17	0.48	0.31
1995	0.02	1.33	0.33
1996	0.06	1.39	0.09
1997	0.23	1.06	0.18
2000	0.36	1.33	0.37
2001	0.2	2.12	0.77
2002	0.07	1.12	0.61
2003	0.51	4.51	1.57
2004	0.34	6.15	1.61
2005	0.05	0.51	0.13
2006	0.18	1.38	0.11
2007	0.54	4.44	1.13
2010	0.07	0.14	0.1
2011	0.28	4.46	0.87

So, this is actually the kind of structure what I will just you know a little bit moderate. So, what will you do, what will you do here. So, the there are 3 variables informations and that is the you know with respect to tie and you will be find there is there are lots of you know missing observations. So, what will you do first, we will adjust the missing observation by simply deleting this particular you know row and then we will do some kind of adjustment and then finally, you must have a consistent data set and, so we have no early business here just we are you know interested to know whether there is a association between X 1, X 2 and X 3 right.

So, what will you do actually? So, in the first instance you need to know what is the relationship between X 1 and X 2. So, we start with the two variables and it can be connected with the more than two variables. So, you simply start with here you know putting equal to sign and then you know give the command here is covariance. So, if you will give the command about covariance then it will help you to find out to you know covariance sample here is. So, just click it and then you just indicate the sample range and against followed by indication of the second sample range and then you close the loop you will get something you know correlation covariance like this right. Similarly you can find out the correlations by you put simple equal to then you know put actually correlation command and then you will get some kind of correlation matrix here. Then against you indicate some range here is up to this and then again you indicate this same range here up to this then you just put you know close the loop and then enter.

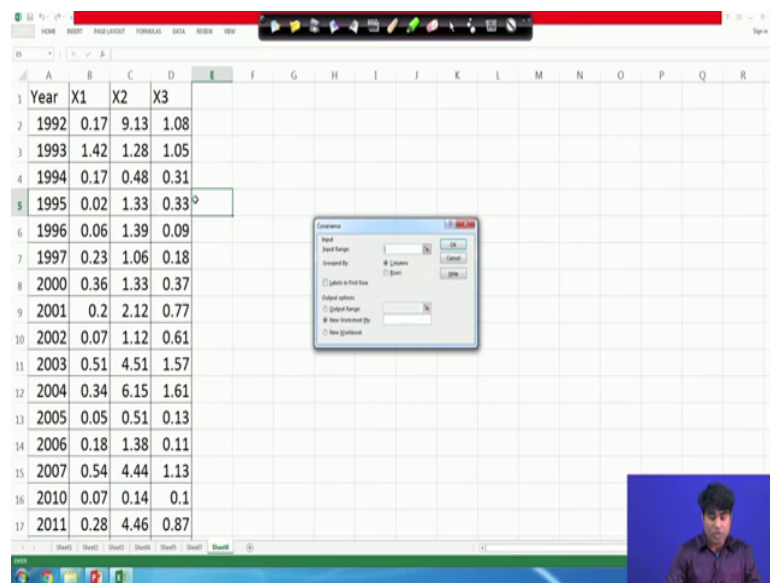
So, see here, it will not work because if you check there then it is not actually consistent. So, what will you do again? So, you just go to this particular you know row against then you indicate the a first row up to this 2014, then you know put comma and then you start with the second row and against I think. So, this will give you some kind of correlation coefficient and that is with respect to X 1 and X 2. So that means, you know just putting equal to signs and then you can obtain the covariance and correlation and in fact, what will you do. So, you can actually just you know simply delete and then you go to the data and then data analysis packets.

(Refer Slide Time: 20:31)



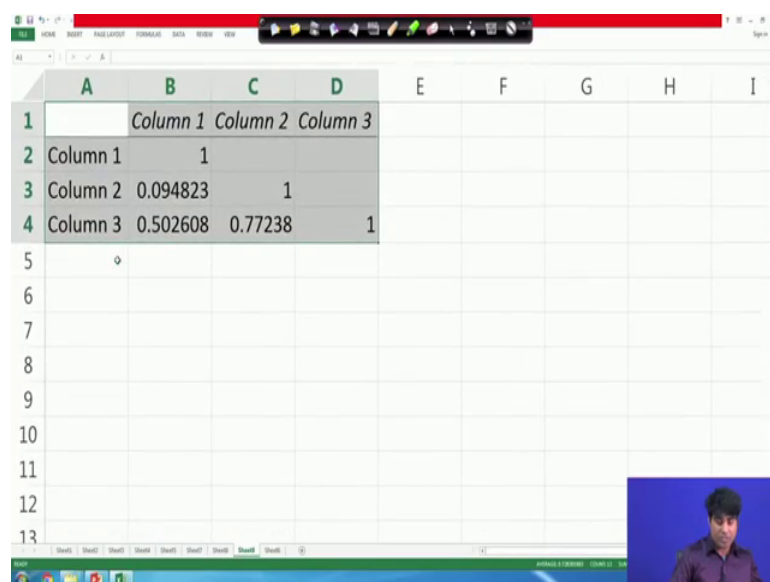
So, if you click here data analysis package then you will find plenty of you know statistics are here. So, what will you do here? Just you first connect with the covariance and put ok.

(Refer Slide Time: 20:40)



So, this will ask you to give you know a range of these two variables. So, you just specify the range of two variables same sample observation. Then you will get actually simply a covariance matrix right.

(Refer Slide Time: 20:48)



So, covariance of you know X 1 upon X 1 covariance of X 1 and X 2 then against covariance between X 2 and X 2. So, this is how the typical structure; that means, if you have actually a more than two variables then you know you defined a structure called as you know variance and covariance matrix. Similarly if you have more than two variables

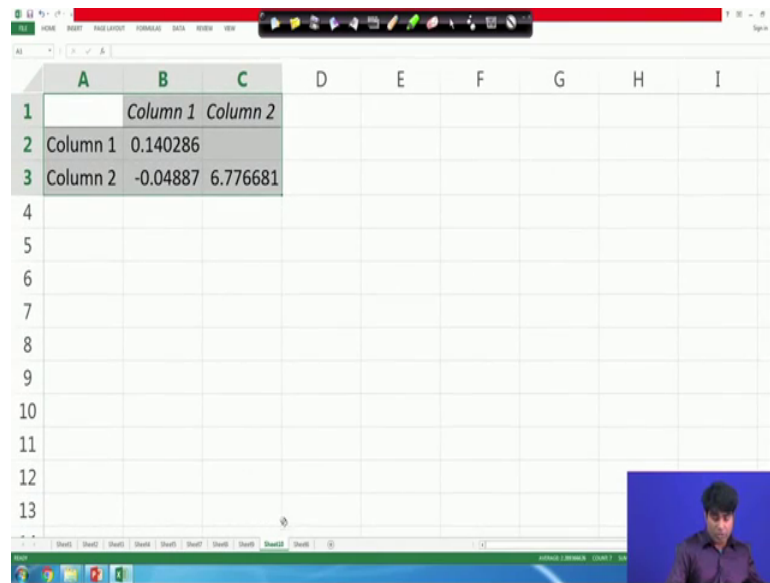
you can find a structure called as you know correlation matrix right. So, what will you do here? So, the same sequence you just again go to the data and then again you choose the data analysis packets then you apply correlations then when you click correlation it will ask you to indicate the particular requirement. So, now, there are 3 variables we like to know what is the association between these 3 variables, and then you just you know indicate with all these 3 variables together with all samples, then you put ok, just this will give you something like this.

So, you see here is the difference you see here the difference. So, this is actually correlation matrix and you see here. So, the correlation between same variable will always equal to 1, then this is first variable second variable and third variables, and this is also first variable second variable and third variable. In this particular you know indication is nothing, but you know correlation upon X 1 upon X 1, then this is nothing but correlation upon X 1 upon X 2 this is nothing but correlation upon X 1 upon X 3 and similarly this is X 2 upon X 3.

So, these are the 3 correlation coefficient we will get, but this one, this one is in all these diagonal elements are 1 because you know when you correlate one upon another you know same variables then the correlation coefficient will be always equal to 1. So, this is actually standardized you will see here the correlation coefficients are lying between actually 0 to 1 only. So, now, similarly if you go to this spreadsheet against then you know you just again go to the data analysis package and then click here is again you put covariance compared to you know correlations. Now, you just you know already indication is there. So, just you put to ok.



(Refer Slide Time: 23:04)

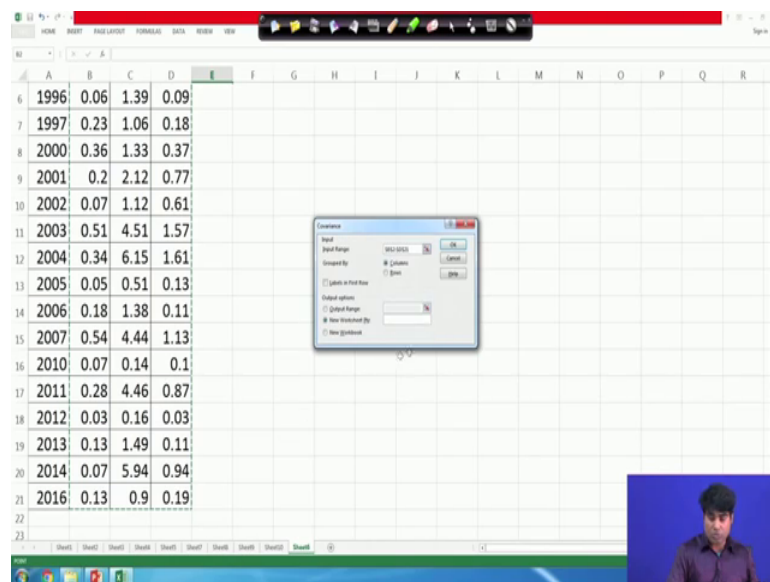


The image shows an Excel spreadsheet with the following data:

	A	B	C	D	E	F	G	H	I
1		Column 1	Column 2						
2	Column 1	0.140286							
3	Column 2	-0.04887	6.776681						
4									
5									
6									
7									
8									
9									
10									
11									
12									
13									

So, this is actually with respect to two variables right this is with respect to two variables and we can actually have a with respect to with a 3 variables also. So, then in this case against you go to the data package then covariance then you put ok.

(Refer Slide Time: 23:28)



The image shows an Excel spreadsheet with the following data:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
6	1996	0.06	1.39	0.09														
7	1997	0.23	1.06	0.18														
8	2000	0.36	1.33	0.37														
9	2001	0.2	2.12	0.77														
10	2002	0.07	1.12	0.61														
11	2003	0.51	4.51	1.57														
12	2004	0.34	6.15	1.61														
13	2005	0.05	0.51	0.13														
14	2006	0.18	1.38	0.11														
15	2007	0.54	4.44	1.13														
16	2010	0.07	0.14	0.1														
17	2011	0.28	4.46	0.87														
18	2012	0.03	0.16	0.03														
19	2013	0.13	1.49	0.11														
20	2014	0.07	5.94	0.94														
21	2016	0.13	0.9	0.19														
22																		
23																		

A dialog box titled "Covariance" is open, showing the "Input Range" as "=\$A\$6:\$D\$21" and "Output Range" as "=\$E\$6:\$H\$21". The "Output options" section is checked, and the "New Worksheet By" option is set to "New Worksheet".

Now, you know reset this particular process you indicate with respect to all the 3 variables simultaneously and with the all samples you see here all samples are you know uniform then you put ok. So, you will be find the correlation covariance matrix is like this.

(Refer Slide Time: 23:38)

	A	B	C	D	E	F	G	H	I	J
1		Column 1	Column 2	Column 3						
2	Column 1	0.09379								
3	Column 2	0.06943	5.71634							
4	Column 3	0.07691	0.92276	0.24969						
5										
6										
7										
8										
9										
10										
11										
12										
13										
14										
15										

This is what the covariance matrix see here. So, the covariance matrix will be like this it will not be in kind of standardized format compared to the correlation. So, the correlation matrix is a correlation matrix is here sorry this is. So, yes this is the correlation matrix and this is what the covariance matrix.

So, now, both are having actually similar kind of structure, but the thing is that you know. So, the idea is that you know we like to know what is the association between the two variables or between these 3 variables and in the case of covariance it will give you some signal and correlation also give the same signal, but correlation is the kind of standardized technique. But covariance is not actually standardized technique. So, when there is a kind of comparative kind of situation. So, it is better to use correlations rather than you know covariance because correlation is a standardized tool. So, that you know you can compare you know very easily. For instance there are 3 different projects and all projects revenues are reported in some currency, country currency. Let us assume that there are 3 projects and 3 different countries and the revenues are reported in 3 different local currency.

Let us say Indian currency, USA currency and UK currency then obviously, if you apply covariance then the covariance between X and Y will be in Indian currency again USA currency and UK currency. So, now so far as a comparison is concerned, you have to go to the currency calculator then you will convert and after that you can compares, but

correlation is the you know I have means it is the beautiful technique. So, you need not require actually currency calculator. So, you just you know connect with all these 3 variables it will be automatically make the particular component unitless. So, then you can in a position to compare and conclude as per the particular requirement.

So, this is how the correlation structures will help you to take decisions and now come back to our original discussion here.

(Refer Slide Time: 25:49)

**Measures of Association**

**Example 4.22 Using the Correlation Tool**  
(Colleges and Universities data)

	A	B	C	D	E	F
	Median SAT	Acceptance Rate	Expenditures/Student	Top 10% HS	Graduation %	
1	Median SAT	1				
2	Acceptance Rate	-0.601901959	1			
3	Expenditures/Student	0.572741729	-0.284254415	1		
4	Top 10% HS	0.503467995	-0.609720972	0.505782049	1	
5	Graduation %	0.564146827	-0.55037751	0.042503514	0.138612667	1

- ▶ Lower acceptance rate, higher median SAT
- ▶ Lower acceptance rate, higher % top 10 HS students
- ▶ Lower acceptance rate, higher graduation rate
- ▶ Higher median SAT, higher graduation rate

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, this is how our actually the way I have discussed the particular you know problem in the excel sheet. So, now, here is we have our original problem is like this. So, we have 5 different variables median SAT score, acceptance rate, expenditure per students and top 10 percent of HS, and graduation rate. Now, all these variables if you connect with actually you know software, software will give you results like this, which we have already discussed and this is actually correlation matrix among these 5 variables.

So, now, you will be find the way we have already discussed. So, all the diagonal elements are one. So, that indicates that you know the variable upon variable is always equal to 1 that is perfectly correlated and other cross correlations are in between 0 to minus sorry minus 1 to 1 and some where we will find negative correlation somewhere we will find positive correlation.

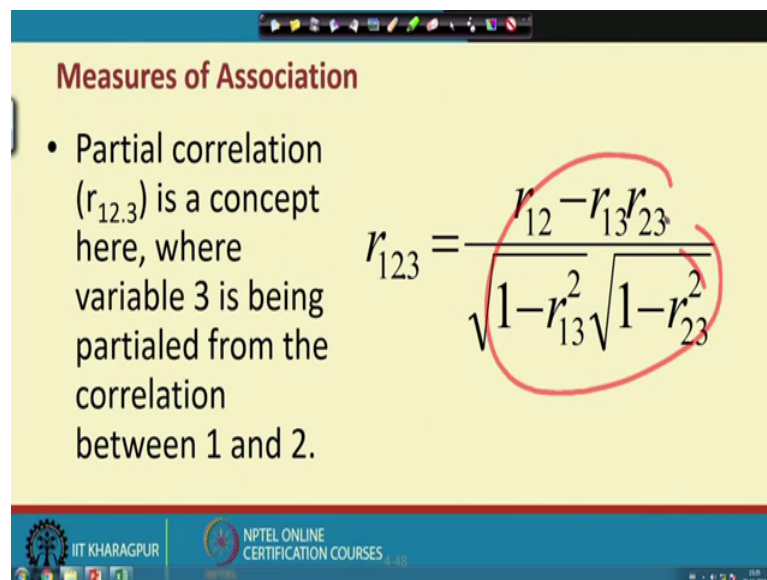
For instance, median SAT score with acceptance rate is coming negatives that is minus 0.60; that means, there is a 64 percent association between the two, but their relationship is completely negative. Similarly median SAT square with the expenditure per student it is coming 0.57 and; that means, 57 percent and the relationship is positive similarly median SAT score with the top 10 percent of high school. So, this is coming 0.50 and that means, 50 percent associations and that is positively related to each other again. Then graduation percentage and median SAT score is also positive and they the association, degree of association is a 56 percent.

Then again with respect to acceptance rate then you have expenditure per students it is coming negative and then again top 10 percent of HS with acceptance that it is also coming 6 minus 61 percent. So, negatively related then a graduation percentage and acceptance rate is also negative and the percentage association is 55 percent and then expenditure with top 10 percent of HS it is a positive and graduation with expenditure per student it is also positive, but it is a very low correlation that is 0.04 that means, 4 percent and then top 10 percent of HS with graduation it is coming actually 14 percent. So, now, this is how the particular you know standardized you know inference or the kind of output which you have obtained through the technique called as you know correlation and the same data you can connect with the covariance, but the covariance results will be coming like this, but it will not be in a kind of standardized format.

But when you will be transfer into correlation then this is in a standardized format and you are in a position to explain in a better way. So, now, so obviously, our idea is how to interpret this result, so far as you know management decision is concerned. So, far as you know acceptance rate and SAT score is concerned. So, the first correlation coefficient that is minus 0.60. So, now, the interpretation will be like this lower the acceptance rate higher is the median SAT score. Then second one is the acceptance rate with a higher you know this ones a higher percentage of top 10 HS students. So, lower the acceptance rate then top 10 percent of HS score will be higher. So, then lower the acceptance rate higher graduation rate because this is again negatively correlated to each other. Now, coming to coming to median SAT score with the graduation rate. So, median SAT score with the graduation rate this is coming actually positives. So, as a result, so the interpretation will be higher the median SAT score higher is the graduation rate.

So, likewise you can compare with the other you know and like you know other situation like expenditure with a graduation then tip to top 10 percent of HS with the graduation and so on. So, the whole idea is you know you have to check the results then you see whether the relationship is positive or the relationship is negative then accordingly you can apply the strategy or you have to take a management decision. If course, this is not enough, after getting this correlation that need to be actually statistically tested which we will discuss in the later stage in the predictive analytics and prescriptive analytic case, but here is in the meantime with the available information or output you may be in a position to address the problems and you may be in a position to take some kind of management decision right.

(Refer Slide Time: 30:35)



**Measures of Association**

- Partial correlation ( $r_{12.3}$ ) is a concept here, where variable 3 is being partialled from the correlation between 1 and 2.

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1 - r_{13}^2} \sqrt{1 - r_{23}^2}}$$

The slide includes a red circle around the denominator of the formula. At the bottom, there are logos for IIT KHARAGPUR and NPTEL ONLINE CERTIFICATION COURSES.

So, now corresponding the examples or the kind of correlation which you have discussed is called as you know simple correlation structures. But you know when you have more than you know two variables then and the correlation coefficient can be divided into two parts against that one part is called as a partial correlation coefficient another part is called as a multiple correlation coefficient. So, partial correlation coefficient is nothing, but actually if there are 3 variables. Then the relationship between two variable keep in third one remain constant. For examples in the business type of you know situations. So, you like to trace the link between price and demand subject to advertising expenditure you know given right. So, with a particular level of advertising expenditures you like to trust the impact or the association between price and demand.

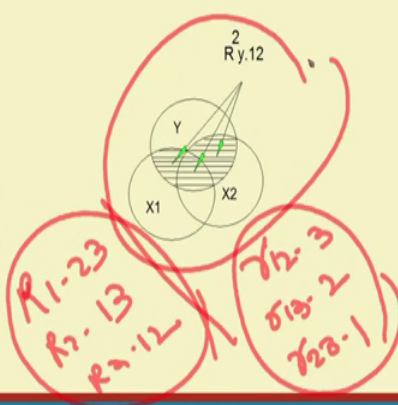
So, likewise you will find plenty of such examples where you know partial correlation coefficient can be applied and this is the simple formula through which you can apply the partial correlation coefficient, so  $r_{12.3}$ , the standard formula is like this standard formula is like this. So, this is what actually. So, it can be it can be  $r_{13.2}$  it can be  $r_{23.1}$ . So, like that you know you will find actually different kind of association structures ultimately. So, you will get a correlation coefficient and again it will be in between minus 1 to 1, but whether it is a partial correlation coefficient or multiple correlation coefficient. So, the inputs to these partial and multiple will be the simple correlation. Since there are 3 variables, you have actually a 3 different kind of correlation coefficient one will be  $r_{12}$ ,  $r_{13}$  and  $r_{23}$  these are all 3 simple correlation coefficient.

Then on the basis of you know reporting 3 simple correlation coefficient you can calculate partial correlation coefficient that will be again 3 types and then you will get some kind of multiple correlation coefficient. But whether it is a partial correlation or multiple correlation it will be also in between the minus 1 to 1, but simple correlation coefficient interpretation will be different partial correlation and multiple correlation coefficient you know structure will be different. And with the help of you know simple correlation your managerial decision will be something different again with respect to partial and multiple the management decision will be different right. So, this is multivariate kind of scenario, but in the case of you know simple this is a bivariate kind of scenario.

(Refer Slide Time: 33:14)

**Measures of Association**

Multiple correlation coefficient:  
With variables  $x$ ,  $y$  and  $z$ , multiple correlation coefficient can be obtained by

$$R_{z.xy} = \sqrt{\frac{r_{xz}^2 + r_{yz}^2 - 2r_{xz}r_{yz}r_{xy}}{1 - r_{xy}^2}}$$


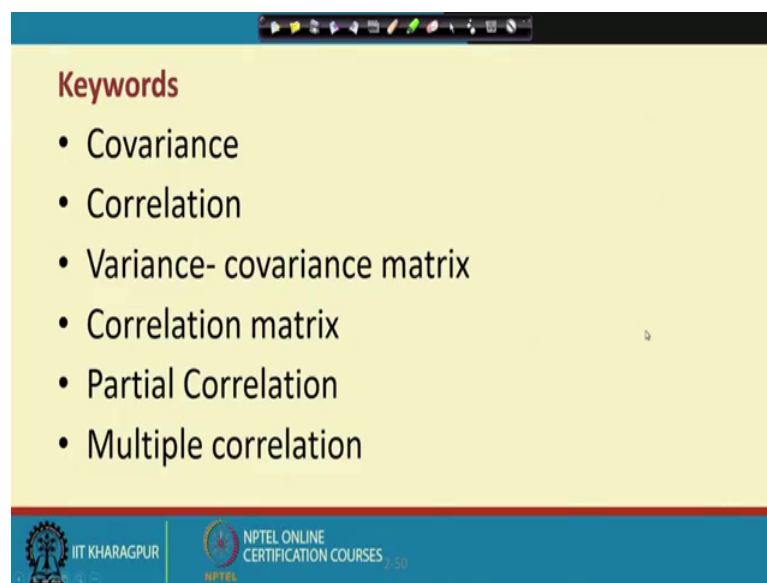
$R_{1-23}$   
 $R_{2-13}$   
 $R_{2-12}$   
 $r_{12-3}$   
 $r_{13-2}$   
 $r_{23-1}$

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, now like the partial correlation multiple correlation can be calculated like this. So, here the multiple correlation coefficient structure will be a between you know 3 variables then  $R_{1.23}$ ,  $R_{2.13}$ ,  $R_{3.12}$ . Say corresponding to partial correlation point  $r_{12.3}$ ,  $r_{13.2}$  and  $r_{23.1}$ . So, these are all 3 partial correlation coefficient and these are the multiple correlation coefficient. And this is actually simple structure of you know multiple correlation coefficients, so that means, in the we will it detail discuss about this multiple correlation coefficient in the case of regression analysis because in the case of multiple regression coefficient here you are connecting with a particular you know variables let us say dependent variables corresponding to two independent variables.

But whether it is actually partial and multiple, our idea is it to know what is the kind of association with respect to multiple kind of involvement or multiple variables involvement. So, in any case, these are all association kind of rules through which you can actually analyze a particular business problem and then accordingly you can take a management decisions.

(Refer Slide Time: 34:39)



**Keywords**

- Covariance
- Correlation
- Variance- covariance matrix
- Correlation matrix
- Partial Correlation
- Multiple correlation

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, now with this we like to conclude here. So, the topics of you know discussion is here with respect to covariance, correlations, variance covariance matrix, then correlation matrix, the structure of partial correlation and the structure of multiple correlations. So that means, technically in a business kind, any kind of business environment we may have a multiple kind of variables involvement to address the problem and to take a



management decision and then association techniques like covariance and correlation can be applied to solve the problem and to take some kind of management decisions.

For instance actually a price and demand. So, there is a link if you find there is association then you like to know how much price you can change. So, that you know demand can be increased or how much you know price you can decrease. So, that demand can be like that you know you are supposed to know something you know with the help of association rules. So, if it is actually coming positive correlation then your strategy will be different, if it is coming negative correlation then the strategy will be different. So, that is why, you must be very careful how to get all these you know outputs and on the basis of these outputs and I am very sure you may be in a position to take some kind of strategy to you know solve your you know management problems and accordingly your decision will be very effective and efficient which we actually need with the help of business analytics right.

So, with this we will conclude here.

Thank you very much.