Marketing Research and Analysis-II
Prof. Jogendra Kumar Nayak
Department of management studies
Indian institute of Technology Roorkee

Lecture - 41
Covariance and Correlation

Hello Friends I welcome you all to the course marketing research and analysis, the second part. In the last class we have been discussing about the various kinds of experimental statistical designs. For example we covered tests like ANOVA, MANOVA, MANCOVA, ANCOVA right and we tried to understand from one way and N way point of view right. So when there is only one factor more than one factor.

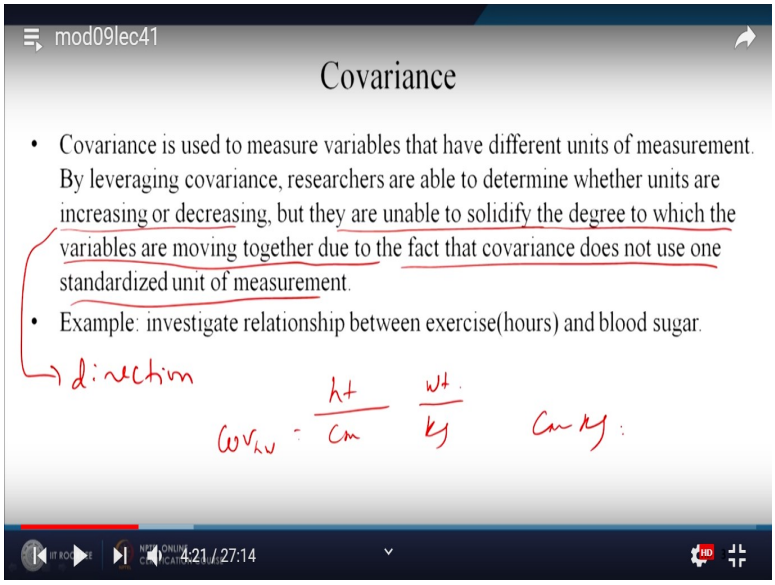
But after that today we will learn something about one of the most important aspects in real life that you face is relationship or association between 2 or more variables. Okay. What is that? What is the importance of understanding this? Now let us think of a suggestion when the temperature increases right. We tend to buy air coolers and fridge or water coolers and all right so that we can satisfy ourselves.

So why what is this relationship. So this relationship is that with the increase in temperature the need for cool or coolness increases. Similarly when you exercise you tend to drink more water right. Similarly when somebody is let say afraid he will speak more mistakes or he will make more mistakes while speaking. So he will tell one lie after another.

So all these things are sometimes directly visible sometimes not directly visible but there are correlations there are relationships or association between 2 or more variables. And by understanding these associations that is present in nature and around you we try to unearth several new you know knowledge. Okay. So today one of these that we are talking about is the one of the associations is called covariance and correlation.

Now this is something which looks very simple and it is simple by the way. But then there is some doubt in peoples mind. Whenever I talk to students I have seen. There is a difference. Is they you know in the; in their thought they are not very clear what exactly covariance and correlation means and how they are different and how they are similar.

(Refer Slide Time: 02:45)



So, today my job is to do that to make it clear to you. So let us see what does this covariance first. Let us start with covariance. So what it says is covariance is used to measure variables that have different units of measurement. Now let us understand we have 2 variables which are different. You have different units. Ok. So, one maybe for example your height which is measured in centimeter and other may be your weight which may be measured in kg.

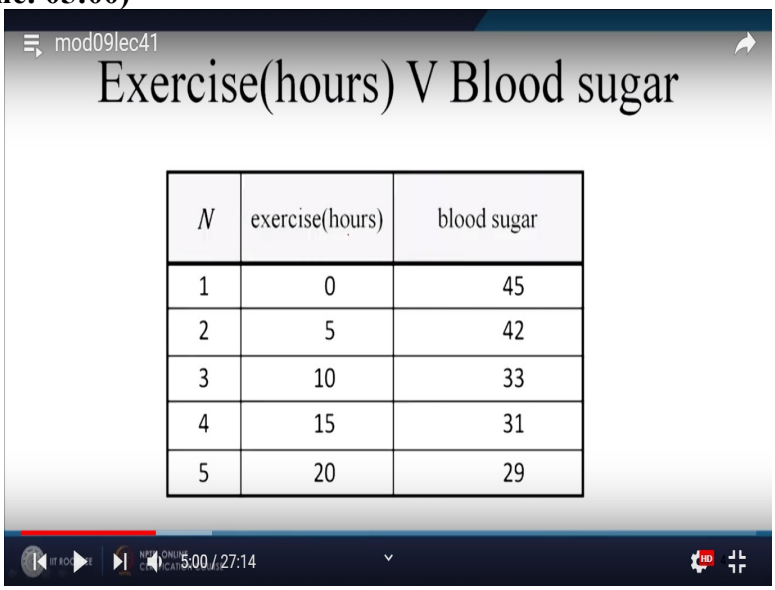
So if I am using 2 different set of variables 2 different sets of units then in that case we want to see how these 2 variables are to be affecting each other or they are moving with each other. So by leveraging covariance researchers are able to determine whether it is units that are increasing or decreasing. That means in simple terms the direction we are talking about the direction. Ok so how are they moving in which direction are they moving? Okay but what is the problem.

They are unable to solidify the degree to which the variables are moving together. Why? Because covariance does not use one standardized unit of measurement, so the problem is that since covariance does not use you use one particular unit of measurement and there are 2 different for example in case of height and let us say weight. So, this is let us say a centimeter and this is kilogram so, something when I do a covariance that means when I do covariance between height and weight.

So my finally; my unit that come is centimeter Kg so which is quite absurd to understand or redundant it looks redundant. So that is why it becomes very difficult to solidify the exact degree to which the standard units that the variables are moving together. But it only helps to

tell the direction whether they are moving positively or negatively. Okay. Example investigates the relationship between exercise hours and blood sugar so a researcher wants to know is there any relationship between exercise hours and blood sugar. Now such kinds of relationships do exist in real life in nature everywhere. So, let us see.

(Refer Slide Time: 05:00)

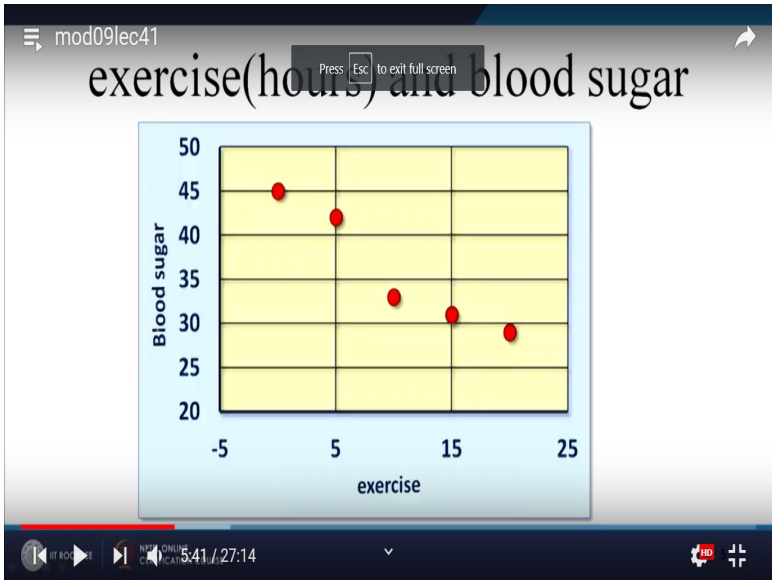


N	exercise(hours)	blood sugar
1	0	45
2	5	42
3	10	33
4	15	31
5	20	29

So this is some 5 respondents have been taking the number of hours they are exercising is given to you and their blood sugar level so you might just take it as an arbitrary number because blood sugar might be measured in some different numbers I do not know more than a 100 or something so just forget it take it is not just a number. Okay. So as it is 0 it is 45 for somebody who does not do any exercise the blood sugar level is 45.

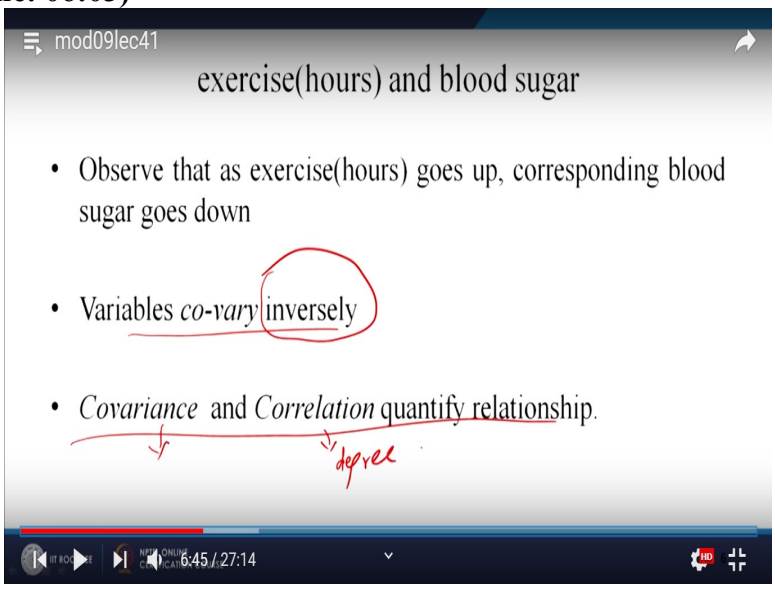
Then it is 5 it is 42, 10 – 33 here we are assuming that exercise is only the reason behind the blood sugar but there might be some several other extraneous variables also which we have not taken into account here. Okay now how do we; How does it look if we plot it?

(Refer Slide Time: 05:41)



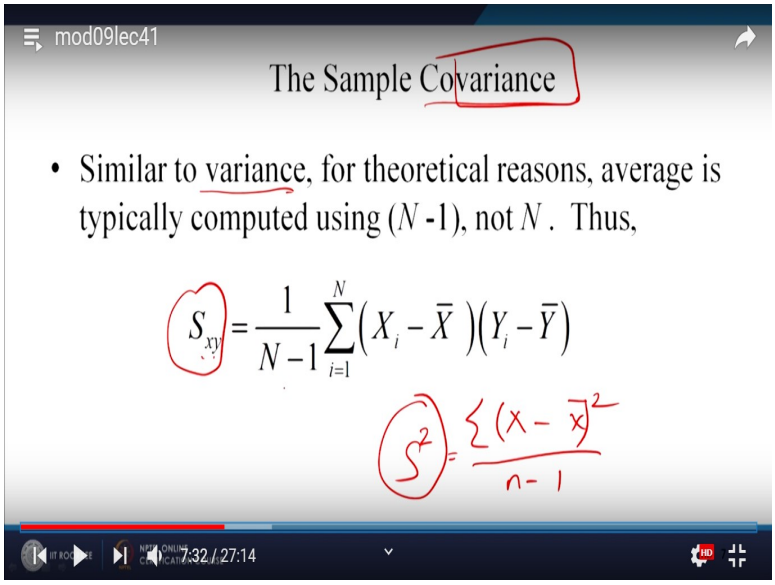
Now let us say this is how it will look. So if I do this. So, 25, 30 that so as my blood sugar is let us says at 45. What is my exercise? Well it is at 0. Let us say this is 0 point, so this 5, 10, 15, 20 so you can see that there is a continuous fall.

(Refer Slide Time: 06:03)



Okay observe that as exercise goes up corresponding blood sugar goes down. Okay. The variables co-vary. But how do they co-vary inversely. That means when one goes up the other goes down. Now, what it is then the relationship between covariance and correlation? Actually covariance and correlation are same thing. You can understand that the same thing. The slight difference exists in between them but covariance does not; is not able to give you a comparison between let us say 2 different variables, because it only tells you the direction but correlation not only tells you the direction but also tells you the degree of the relationship.-

(Refer Slide Time: 06:48)



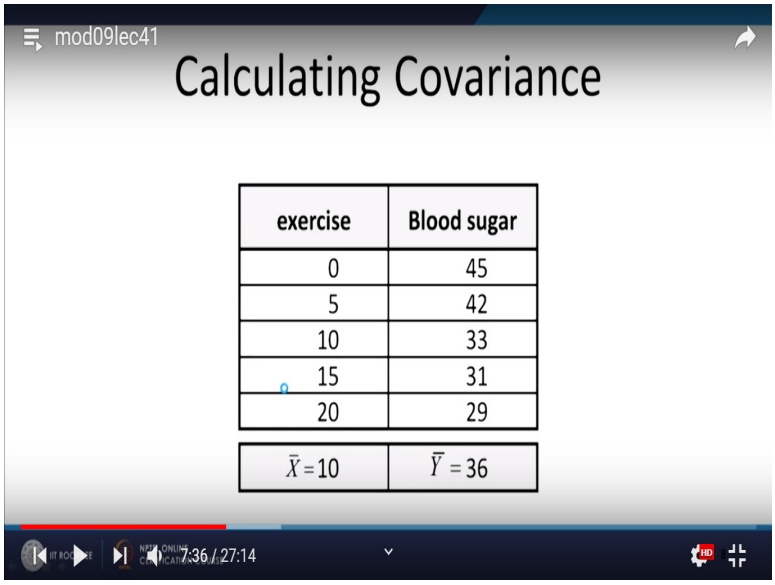
So let me show you how it is done. So first let us understand the covariance little more similar to variance. Now this word has come from the word variance. I hope you understand what variance is. So how do you measure variance is nothing as square of standard deviation

which is given this $$s^2 = \frac{\sum (X - \bar{X})^2}{N-1}$$. \

So this is the formula for variance. Similarly covariance means when 2 variables are covariances which are 2 variables the X and the Y ok. The formula is

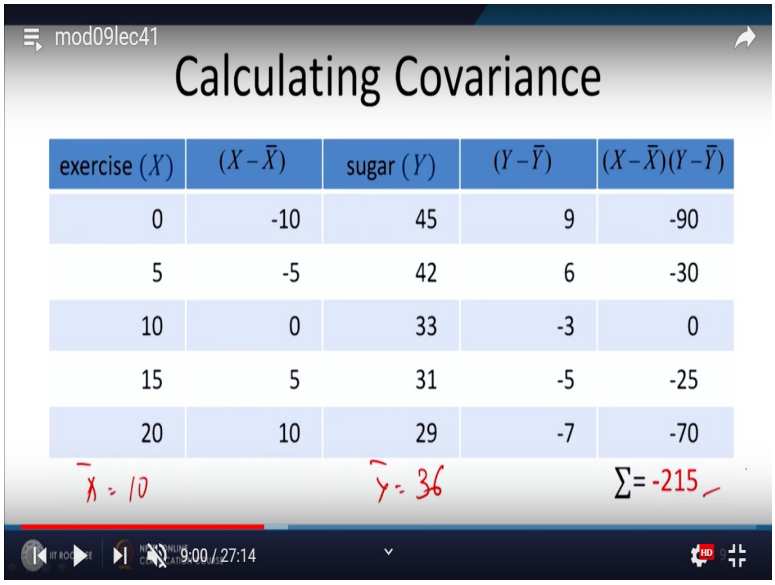
$$S_{xy} = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})$$

(Refer Slide Time: 07:35)



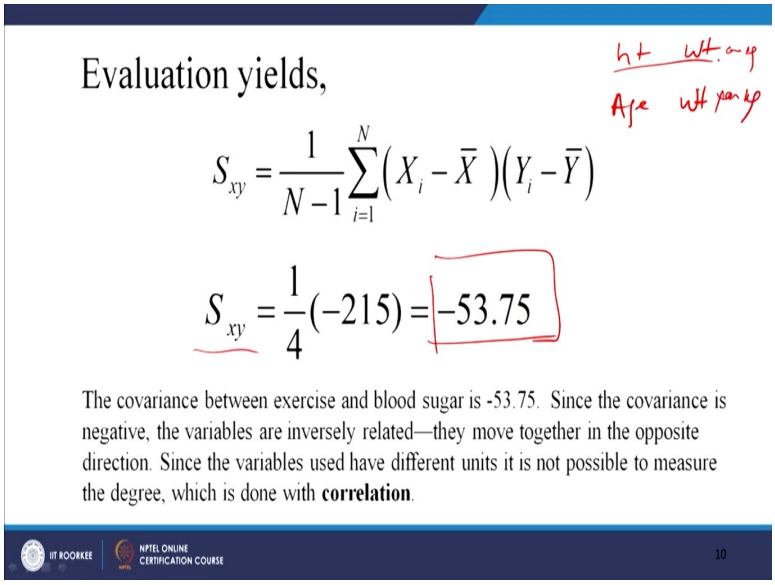
Now let us take the same case so 0, 5, 10, 15, 20, blood sugar is this much. How do we calculate?

(Refer Slide Time: 07:40)



So $(X - \bar{x})$ So what is my. In this case what is my $\bar{x}$ is 0+5+10+15+20/ 5 so my $\bar{x}$ is 10. Similarly my $\bar{y}$ is equal to how much it is 36 right. Now we can calculate $(X - \bar{x})$ so first is 0 – 10,-10, 5 -10, -5, 10 -10, 0, 15 – 10, 5, 20 -10, 10, similarly for sugar Y, 45 – 36, 9, 42 – 36, 6, 43 – 46, -3, 31 – 36, -5, 29 – 36, -7 okay. Now we made we multiplied together $(X - \bar{x})^* (Y - \bar{y})$ but so this is giving us how much - 10 * 9, - 90, - 5 * 6 ,- 30, 0 * 3, 0, 5 * -5, -25 summation of this together gives us -215 ok.

(Refer Slide Time: 09:02)



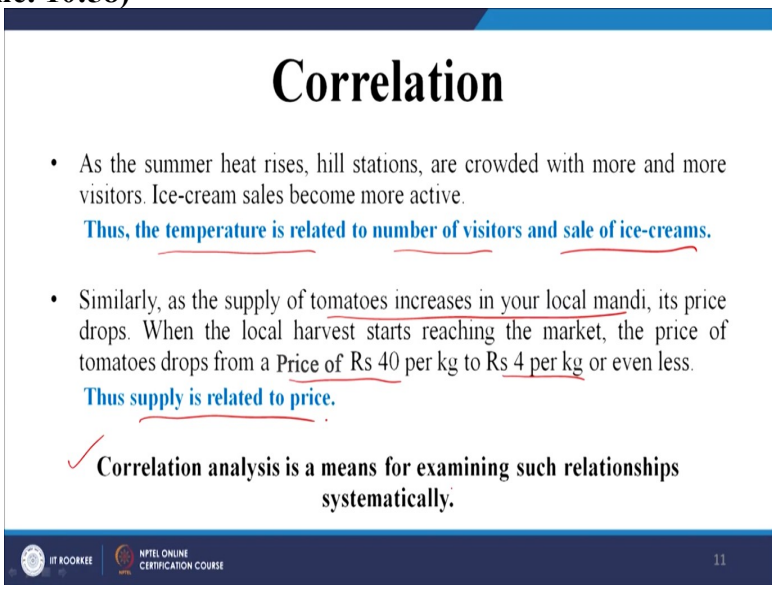
Now if I take this $(X - \bar{X})^* (Y - \bar{y})$ so that is -215 / N – 1, N was 5, so if 5 respondents so 5 – 1 = 4 right. So my covariance value now is equal to how much -53.75. So the covariance between exercise and blood sugar is -53.75 since the covariance is negative the variables are inversely related they move together in the opposite direction that means one goes up the other goes down.

Since the variables has used have different units it is not possible to measure the degree which is done with correlation. So this is not possible here. Let me show you one more example. Let us say you want to measure; compare the relationship between you know that when age increases people's weight increases. Similarly when you know height increases also the weight increases. So suppose you want to take a relationship between height and weight and let us say age and weight right.

So, if I do this let us say if I do this and I find out well this is the relationship between height and weight is something some value we have got let us say when we take height as X let us say weight as my Y similarly then I get some value this covariance value. Similarly when I take age my X and Y my weight then I am getting another value. So you see now so here are my units would be like what height is centimeter Kg and this is year suppose ages in year and kg.

Now just look can I compare these 2? Well no, I cannot. Why? Because they are completely of course different units, right. So that is what it is saying. So how do we go ahead? Here then comes the use of correlation.

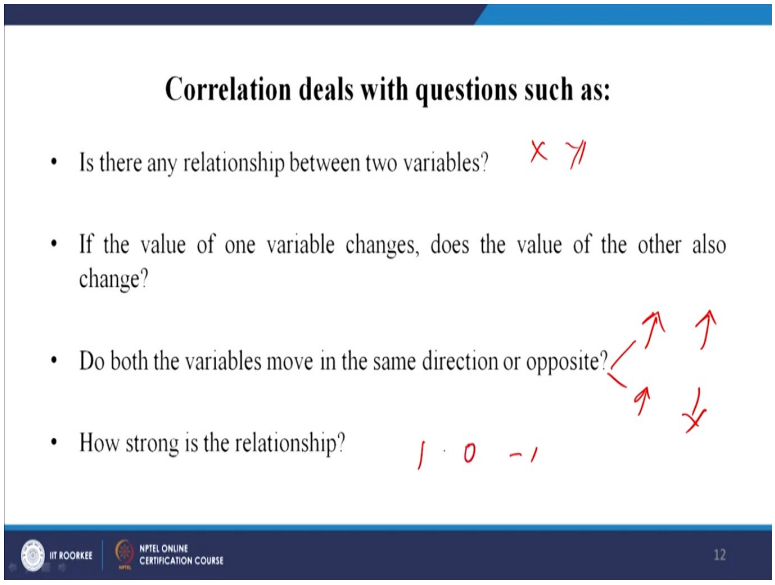
(Refer Slide Time: 10:58)



Now let us see correlation as the summer heat rises. Hill Stations are crowded with more and more visitors. Okay. Ice cream sales as become more active. So the temperature that means we can say the temperature is related to the number of visitors visiting the place and also the sale of ice creams. So if the temperature is increasing outside so people try to visit the hill stations which is more, cooler and the consumption of ice creams also increases because it is hot outside.

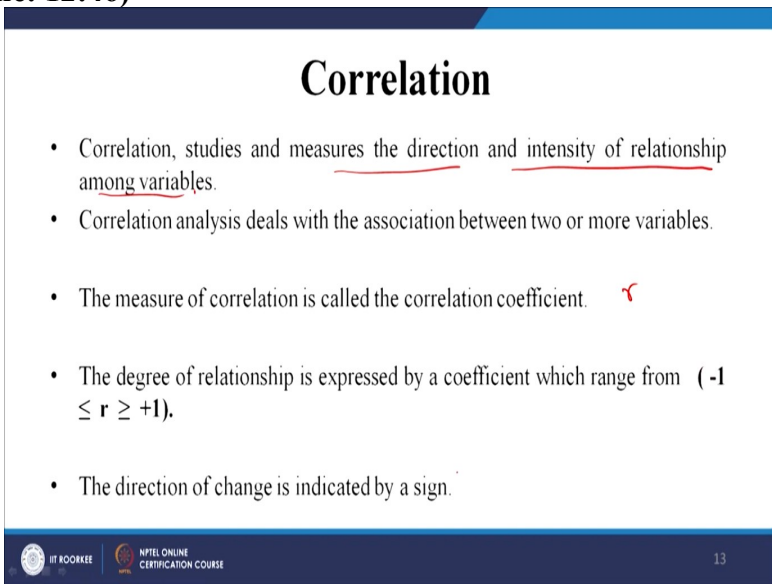
Similarly as the supply of tomatoes increases in your local market its price drops. When the local harvest starts reaching the market the price of tomatoes drops from a price of rupees 40 per kg for maybe 2 rupees 4 per kg for or even less. So the supply is related to price so that means these are all things that we know that they are associated. So these associations are very important for us to understand in real life. Okay so what is this correlation saying correlation analysis is a means for examining the relationships in systematic manner ok.

(Refer Slide Time: 12:04)



Deals with several questions what are the questions? Is there any relationship between 2 variables X and Y. Okay if the value of one variable change is suppose X does the value of the other changes that means for X changes does Y you also change. Do both the variables move in the same direction or opposite? That means if X goes up does Y also go up or if X goes up that is Y go down. So what is the; which relationship comes, how strong? What is the strength of the relationship? Is it let say very strong relationship? Then we will say it is +1 if it is very weak relationship. It is -1 if it is no relationship it is 0.

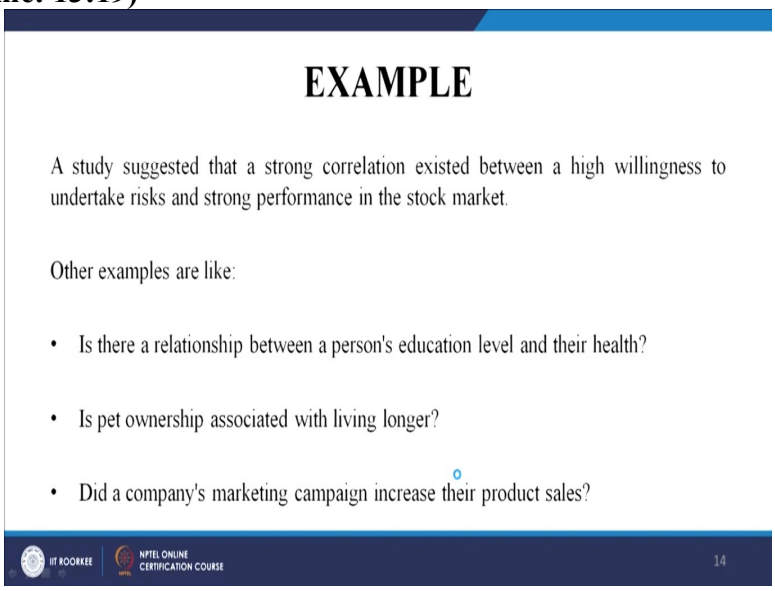
(Refer Slide Time: 12:46)



So, correlation studies and measures the direction which covariance was also doing and intensity of the relationship among the variables. So, it is helping you to not only understand the direction but also the intensity and how is it doing that. So we will see that correlation analysis deals with the association between 2 or more variables. Okay. The measure is called

the correlation coefficient given by "r" and this ranges from -1 to +1 the direction of change indicated by sign. Okay.

(Refer Slide Time: 13:19)



Now this example but before that let us understand when you talked about covariance. So in covariance when I; what is how correlation is measured. Correlation takes covariance of X and Y divided by standard deviation of X by standard deviation of Y. So when we are doing this in one way the correlation helps you to standardize the values and bring it to range of -1 to +1.

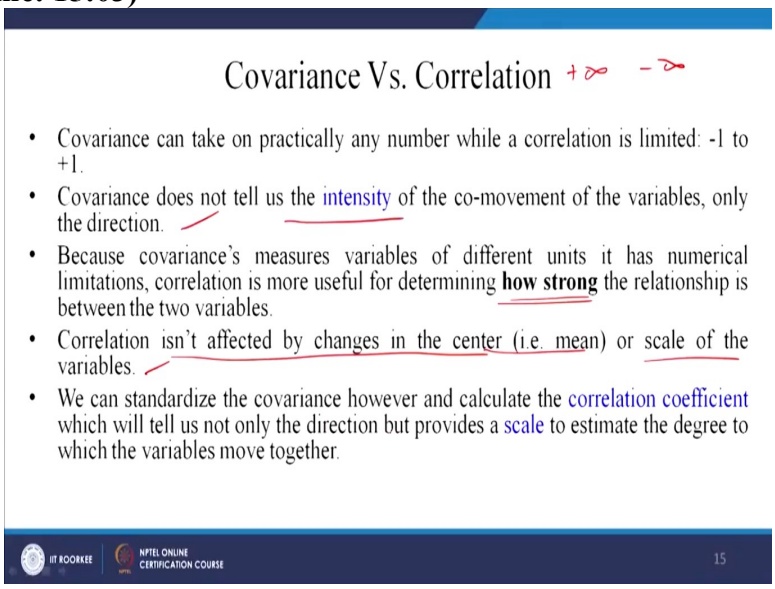
So anything that is now anything that you can you are doing with any kind of variables they all are in a range between -1 to +1. So now since they are now it has become unit less. Okay so since it has become a unit less you can easily compare it. Okay a study suggested that a strong correlation existed between high willingness to undertake risk and strong performance in the stock. That Means people are more risk taking. They do well in the stock market.

So some other examples are like is there a relationship between a person's education and their health. That means if people are more educated is that is it does it have a relationship with their health. Could it be that some people might argue that people were educated they are more conscious about their health. Owning a pet does it associated with living longer like in some countries they say if you have owned a dog or a cat it helps you to live longer because you are more compassionate. You are more kind.

So, these are positive traits which help you to live longer, so there are less heart attacks and all. Did a company's marketing campaign increase their product sales? So we want to see

such associations.

(Refer Slide Time: 15:03)

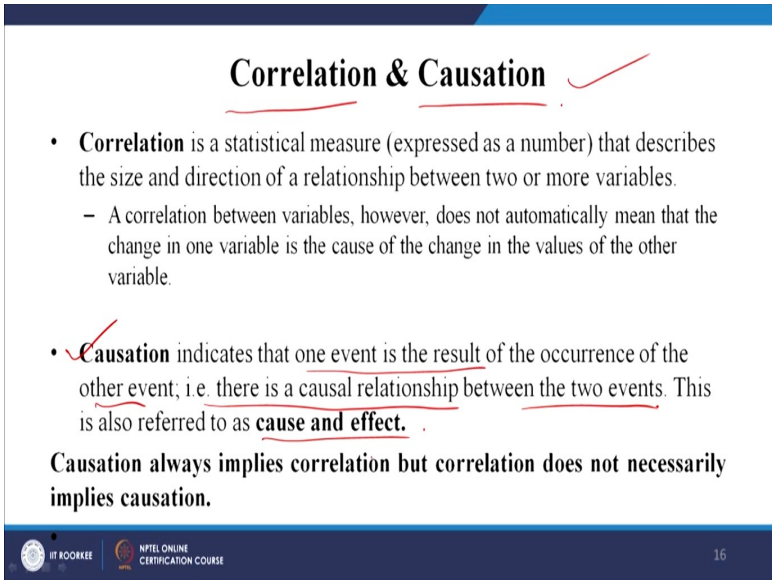


So what is the difference, let us say comeback, covariance can take on practically any number while a correlation is limited to -1 to +1 therefore it is a covariance lies between + infinity to - infinity and correlation lies between -1 to +1 . Covariance does not tell us the intensity only the direction because covariance measures variables of different units. It has numerical limitations obviously different units are there centimeter and kg. Correlation is more useful for determining how strong the relationship is between the 2 variables.

It is not affected by changes in the mean or the scale of the variables because it is standardizing them. We can standardize the covariance however and calculate the correlation coefficient. So the similarity is that both are doing the same thing but the difference is that covariance is an unstandardized value and correlation is the same value which half of the covariance which has been standardized now which will tell us, not only the direction but provide a scale to estimate the degree to which the variables moved together.

So that means in real life this becomes a more powerful tool out of the covariance and correlation, the correlation becomes a more powerful tool for us. Now what is a correlation and causation? So, 2 variables are correlated so that we understand as there is a correlation.

(Refer Slide Time: 16:31)

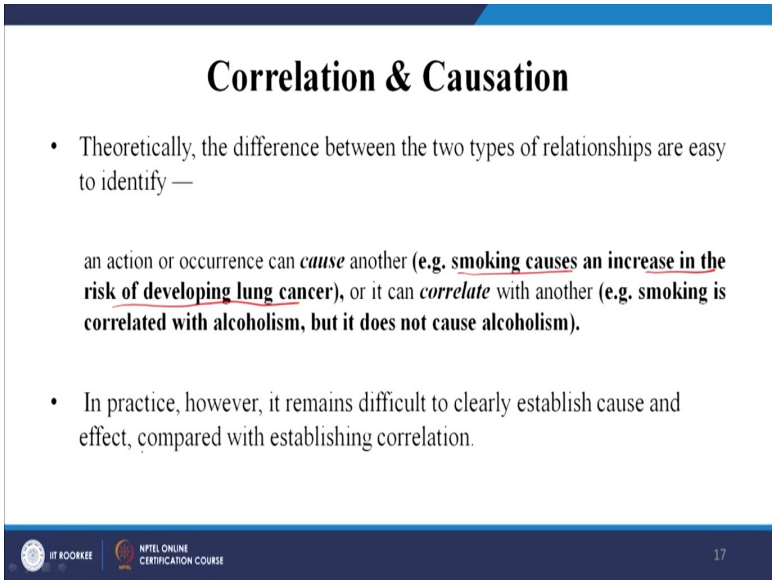


But what is causation? Let us understand, we have understood correlation it says a statistical measure that describes the size and direction of a relationship. A correlation between variables however does not automatically mean that a change in one variable is the cause of the change in the values of the other variable. That means if one variable changes automatically it effects the other, so that it does not say. But whereas, what is causation saying, it indicates that one event is the result of the occurrence of the other event. If this event would not have been there the other would not have happened that is what it says.

That is there is a causal relationship between the 2 events, as you have studied, if I remember concomitant variation when I was talking about concomitant variation when I was teaching ANOVA at that time we said how 2 variables are effected that we cannot say after it has happened that you know any variable the role of any variable at the same time 2 variables A and B this at the same time how their varying.

So this is also referred to as cause and effect causation always implies correlation, causation always implies if there is a cause there is a correlation but correlation does not necessarily imply causation, that means what, if there is a 2 correlation between 2 variables that does not necessarily mean that one effects the other or one is the cause of the other. But if there is a cause and effect relationship that means that is surely a correlation between them. I hope you understand, this causation leads to means there is a correlation but correlation only does not necessarily mean that there is a cause and effect.

(Refer Slide Time: 18:19)

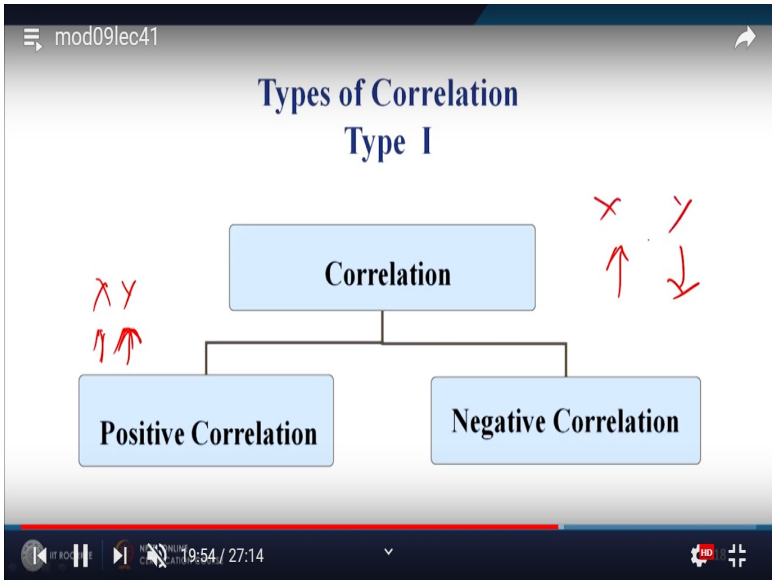


So theoretically the difference between the 2 types of relations are easy to identify an occurrence action or occurrence can cause another see smoking causes an increase in the risk of developing lung cancer or it can correlate with another example smoking is correlated with alcoholism but it does not cause alcoholism. We can say smoking an alcoholism are related. We have seen people who smoke they are more you know the chances of their consuming alcohol is more than a person who is not smoking.

Or we can connect with alcoholism and non-vegetarian food. People who are fond of non-vegetarian food maybe I am not trying to put forth an argument but maybe the chances are that in India atleast we have seen that these people are more prone to have a taste for alcohol. So but that does not mean if somebody is eating non veg he will that is a cause for becoming consuming alcohol.

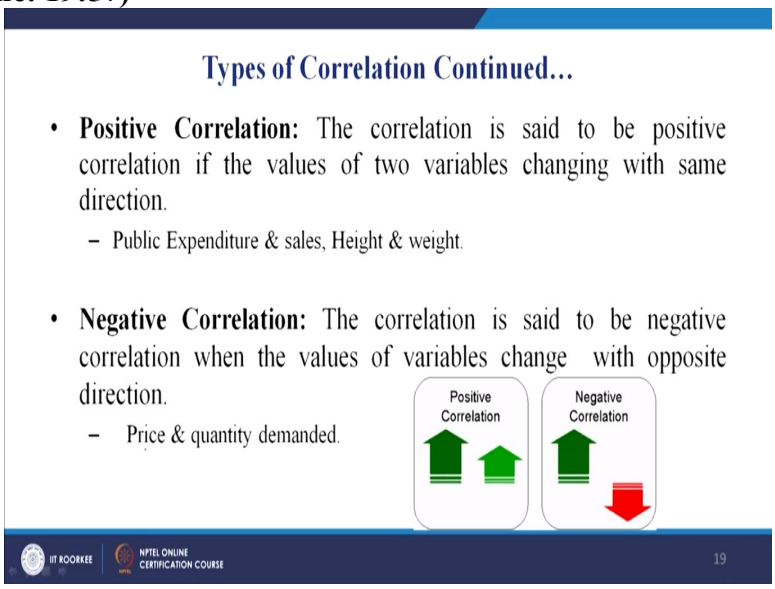
So that is what we were trying to discuss in the last slide. In practice however it remains difficult to clearly establish cause and effect compared with establishing correlation cause and effect is very difficult to establish but correlation we can find out the types of correlation.

(Refer Slide Time: 19:40)



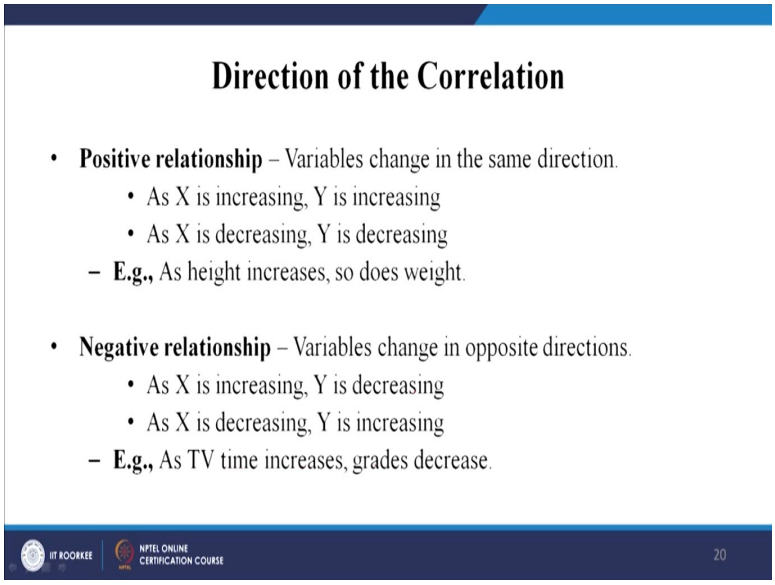
So basically there are 2 types of correlation Positive and negative so positive you understood, if the 2 variables X Y are positively correlated, If X increases, Y increases, negative correlation if X increases, Y decreases.

(Refer Slide Time: 19:57)



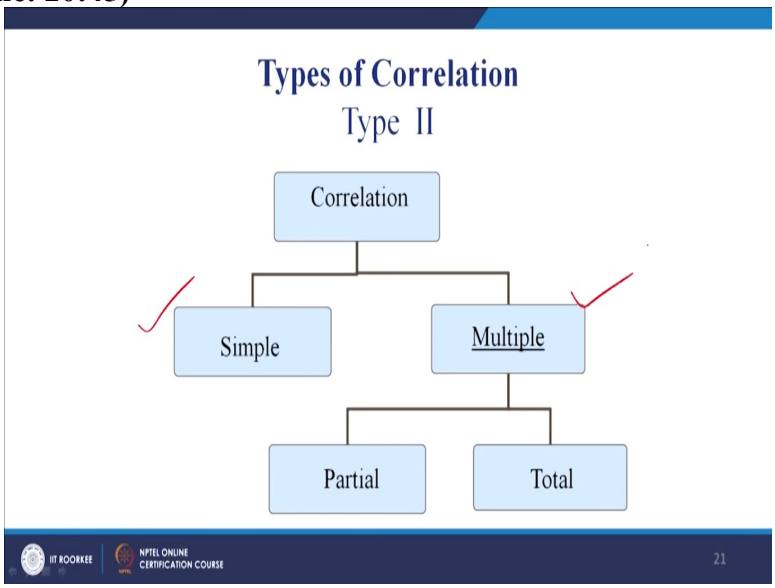
Now the correlation said to be positive if the 2 values of the values of the 2 variables changing in the same direction public expenditure and sales. So if public expenditure is increasing, sales is increasing, height is increasing, weight is increasing. Negative, the correlation is said to be negative when the values change with opposite direction, prices increasing quantity demanded is decreasing this is the case. And the first one this is the case.

(Refer Slide Time: 20:24)



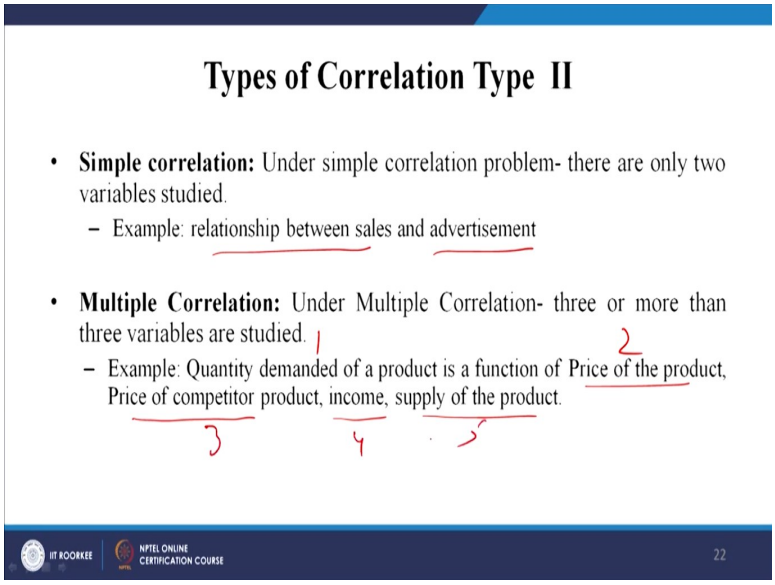
As X is increasing Y is increasing as X is decreasing Y is decreasing this is also a positive relationship. That means what they are moving in the same direction. But negative you see as X is increasing Y is decreasing X is decreasing Y is increasing so opposite. As TV time increases grades decrease. I do not know whether students will agree to this or not but this is a fact.

(Refer Slide Time: 20:45)



Now how do you define correlation in other ways? Let us see simple correlation and multiple correlations. So simple means the relationship correlation between 2 variables and multiple means whether you have more than 2 variables and we are trying to establish a relationship among the 3 or 4 or 5 whatever it is.

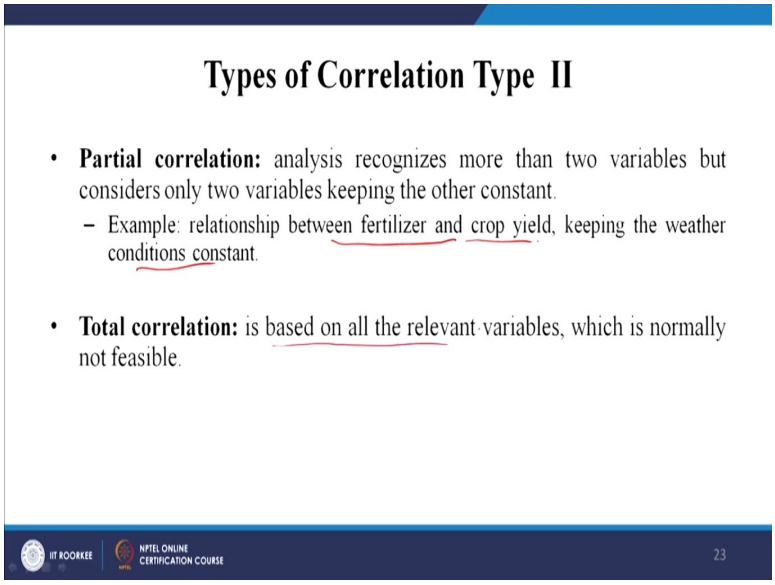
(Refer Slide Time: 21:08)



So let us see, this simple correlation under a simple correlation problem there are only 2 variables studied relationship between sales and advertisement. Multiple correlation, 3 or more than 3 variables are studied at least 3 or more than 3 quantity demanded of a product is a function of price of the product, price of the competitors product, income supply of the product.

Now you see how many variables are 3 quantity demanded of a product is saying is a function of price or quantity demanded is associated with price of the product, price of the competitors product, income of the people and the supply of the product available in the market. So 1, 2, 3, 4 and 5 so there are 5 variables. And suppose you are measuring a correlation and by measuring a correlation here we can say which 2 are most highly correlated we say quantity demanded has the highest correlation with price of the product or price of the competitors product or income of the person or the supply in the market.

(Refer Slide Time: 22:11)

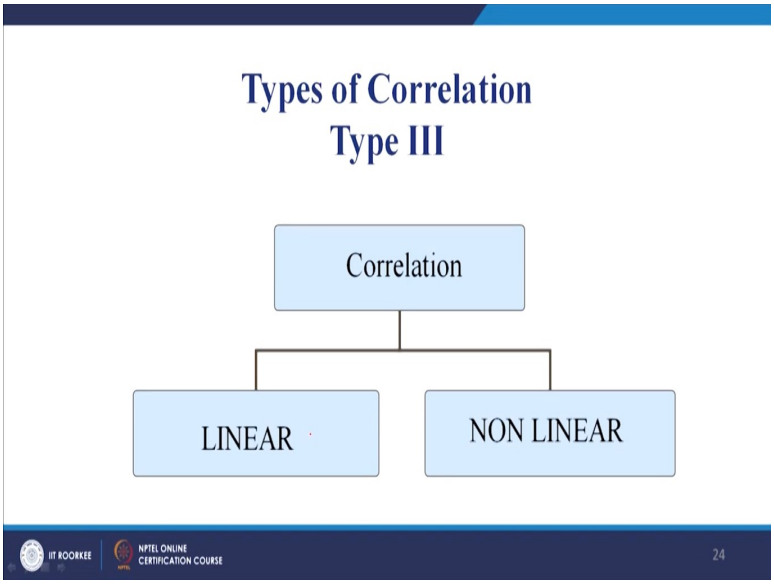


Third is a partial correlation is also equally very, very important to understand what is this partial correlation? Many of times students ignore and are they just missed it out but that should not happen. What it says is, it recognizes more than 2 variables but considers only 2 variables keeping the other constant that means supposing there are in this case there are 4 variables.

Let us say, this is the one with the main and the rest are affecting it that we want to see the correlation. Now suppose I want to see the correlation between quantity demanded and the price of the product let us say, but and I will keep these things as I suppress them in that case I am doing a partial correlation. Example you see, relationship between fertilizer and crop yield.

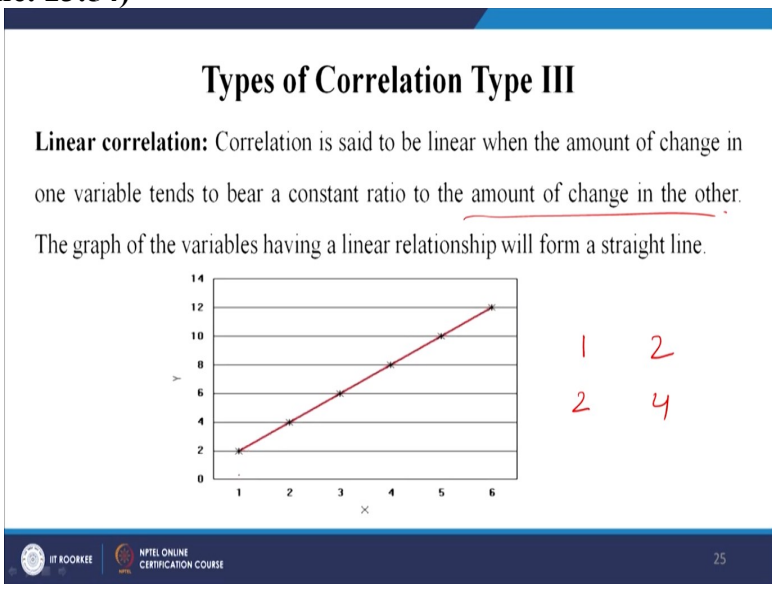
Keeping the weather conditions constant or you can just say you can you suppressing the weather condition. So here similarly we are suppressing the price of the competitor's product income and supply. So when I do this it is the case of a partial coalition so you can do ones for less the quantity demanded and price of the product. Then you can do 1 for 1 and 3 keeping 2, 4 and 5 as my suppressing the 3 variables or you can do 1 and 4 keeping 2 for 2, 3 and 5 constant similarly 1 and 5 keeping 2, 3 and 4 constant. What does this total correlation? Is based on all the relevant variables which is normally not feasible.

(Refer Slide Time: 23:40)



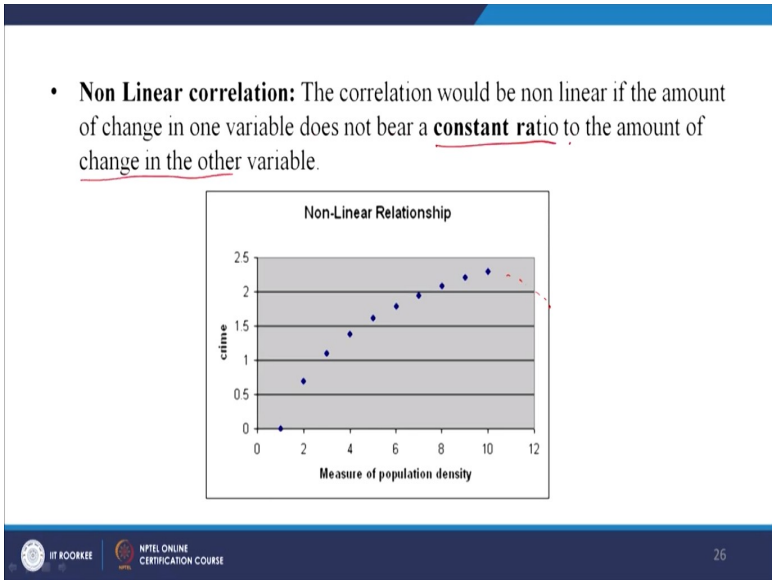
So let us see what is exactly meaning of that, so types of correlation, correlation linear and nonlinear another way of understanding, linear and non-linear. So you understood simple, multiple, partial. Now understand it from a linear and non-linear point of view.

(Refer Slide Time: 23:54)



Linear is very simple we have understood when the amount of change in one variable tends to bear a constant change or the constant ratio to the amount of change in the other, if there is a suppose 1 unit change here leads to 2 units change there, so in the second case 2 units changes here leads to 4 units changes there. So when I am doing this it is a constant ratio. So the graph of the variables having a linear relationship will form a straight line you see. So 1 2, 2 4, 3 6, 4 8, 5 10, 6 12 goes on.

(Refer Slide Time: 24:26)



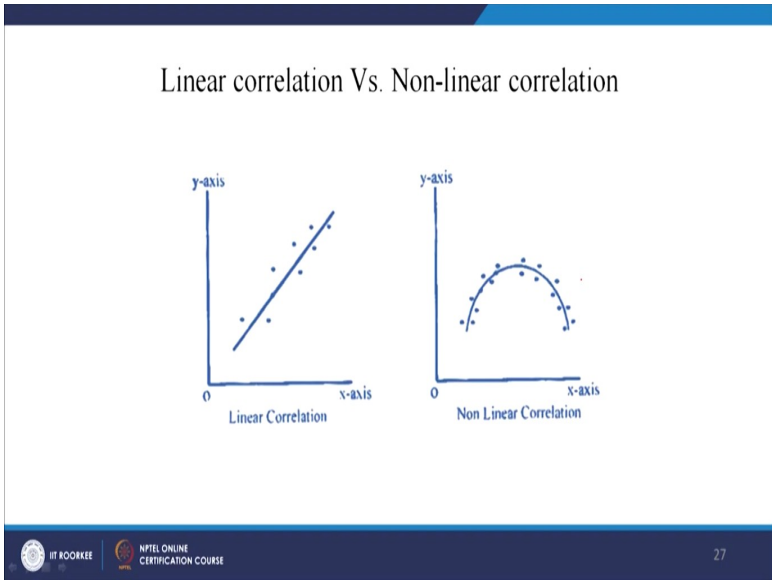
But what is this non-linear relationship? The correlation would be non-linear if the amount of change in one variable does not bear a constant ratio to the amount of change in the other variable in fact in life most of the relationships are not linear. Although we assume them to be linear for a statistical you know to make a statistically simplicity but they are not linear.

For example you see crime and population density so as it is you see the measure of population density is increasing the number of crime is increasing but the increasing rate is not linear. So it is increasing at let us say in the initial period at a very, very high pace. But then after that the pace is slowing down. And it could be also possible that after a certain time it may even come it may fall down.

So this non-linear correlation says that the correlation would be non-linear if the amount of change in one variable does not bear a constant ratio to the amount of change in the other variable. So that means the relationship could be just you know a nonlinear that means, so here it could be for one it is 1:2 in the second case it could be 1:5, it could be 1:3, 1:4 or it could be an exponential growth.

So then we are having such kind of a relationship we say it is a non-linear or curvilinear relationship. So when there is a curvilinear relationship when I move from a correlation to regression we will also be solving problems on that and that is very essential to understand that most of our relationships in life might not be linear. So this is how you see this.

(Refer Slide Time: 26:08)



So this is a linear correlation and this is a non-linear or curvilinear relationship, well today, I will just stop it here in the next lecture will continue further with the correlation and we will try to understand how to solve a problem for a correlation problem and understand and the concept of regression for that the dynamics of regression and use in real life. Well thank you so much. Thank you for the day.