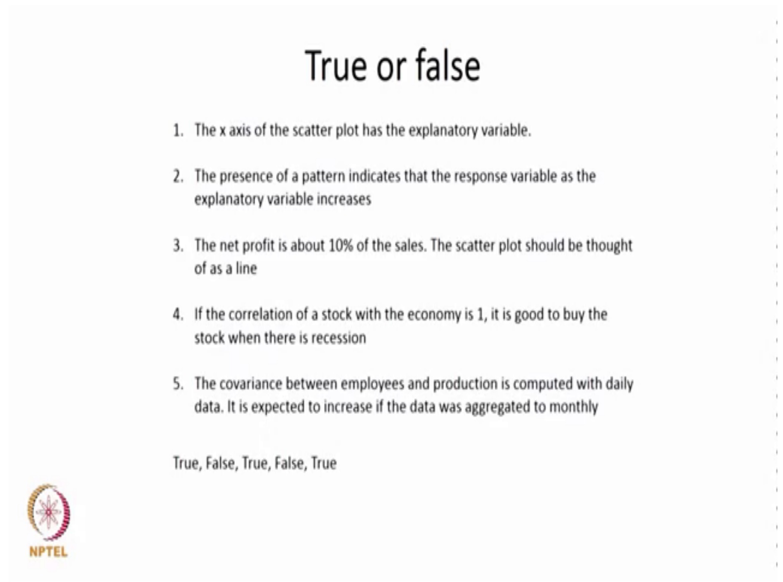
Introduction to Probability and Statistics
Prof. G. Srinivasan
Department of Management Studies
Indian Institute of Technology, Madras

Lecture – 11
Association between Numerical Variables (Continued)

In this lecture we continue to discuss association between numerical variables. So, we look at a few exercise questions, try to understand the concepts further and then we will summarize what we have learnt in the 11 lectures and try to wind up the discussion on statistics in this course and then go on to do probability and start models for probability in the next lecture. So, we have looked at association between numerical variables.

So, we first looked at covariance as a association measure of association, we also said that covariance can be negative and then from the covariance we moved on to describing the correlation coefficient, which looks like a more compact measure because it takes on values between minus 1 and plus 1. And because covariance is negative and individual standard deviations are positive, correlation coefficient can also take negative values it takes values between minus 1 and plus 1.

(Refer Slide Time: 01:32)



True or false

1. The x axis of the scatter plot has the explanatory variable.
2. The presence of a pattern indicates that the response variable as the explanatory variable increases
3. The net profit is about 10% of the sales. The scatter plot should be thought of as a line
4. If the correlation of a stock with the economy is 1, it is good to buy the stock when there is recession
5. The covariance between employees and production is computed with daily data. It is expected to increase if the data was aggregated to monthly

True, False, True, False, True

So, with this let us move on to some questions some true or false questions, the x axis of the scatter plot has the explanatory variable the answer is also given so, the answer is true. So, the x axis is the independent variable or the variable that tries to explain something happening and the y axis now has the variable on which the effect of the explanatory variable is felt therefore, x axis has the explanatory variable is true.

Question number 2, the presence of a pattern indicates that the response variable increases as the explanatory variable increases. So, the answer is not necessarily true because we may have a pattern where as the x variable increases the y variable can decrease so, that happens when there is a negative correlation. So, it is not entirely true though one might be tempted to think it is true because our mind normally makes us believe that there is a positive correlation. So, if there is a positive correlation then as x increases y also would increase, if there is a negative correlation as x increases y would decrease and therefore, this statement that the presence of a pattern indicates other that the response variable would increase as the explanatory variable increases is not entirely true.

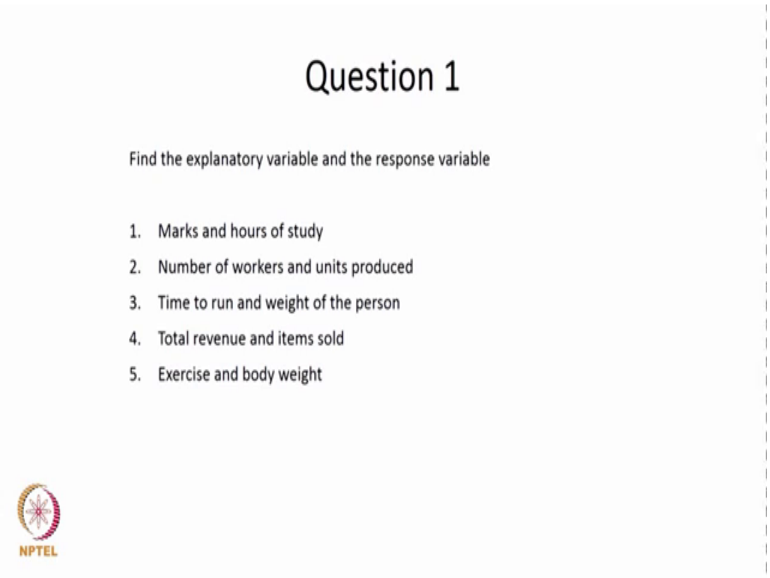
Third question it serve a situation where the net profit is about 10 percent of the sales. So, the scatter plot should be thought of as a line. So, the question is now does this look like a line or would it be non-linear and so on. Now the net profit is about 10 percent of the sales gives us an indication that we have a line of the form y is equal to a plus 0.1 x and so on. Roughly the slope can be thought of as 0.1 and therefore, one can believe that when we start plotting this data, such a data would approximate to a line and therefore, the answer could be true for this statement.

Statement number 4, if the correlation of a stock with the economy is 1, it is good to buy the stock when there is recession. Now, the answer is given here is false because as the stock is entirely dependent on economy and entirely correlated with it with the correlation of one. So, when the economy is down the stock will also be down and therefore, it depends on what we want to do with the stock if you want to trade it very regularly buy and sell the next day and so on, then it is not a very good thing.

But therefore, the answer is false, but if we have a person who simply buys the stock keeps it for a very long time waits for the economy to recover. So, that the stock prices also go up and then the person wants to sell it then the answer could be true, but in general the answer is false because if economy is down the stock price will also be down.

Question number 5, the covariance between employees and the production quantity is computed with daily data it is expected to increase if the data was aggregated to monthly yes the as we aggregate data we realize that the covariance increases. So, these questions have helped us understand what is correlation, what is covariance, and how we model a linear relationship, it also helps us understand what is an explanatory variable and what is the dependent variable and so on.

(Refer Slide Time: 05:23)



Question 1

Find the explanatory variable and the response variable

1. Marks and hours of study
2. Number of workers and units produced
3. Time to run and weight of the person
4. Total revenue and items sold
5. Exercise and body weight

So, let us move to the next, a simple question find the explanatory variable and the response variable. So, the explanatory variable is the x variable response variable is the y variable. So, we have to look at these situations and try to find out which one has an effect on the other or which one can be explained by some other variable.

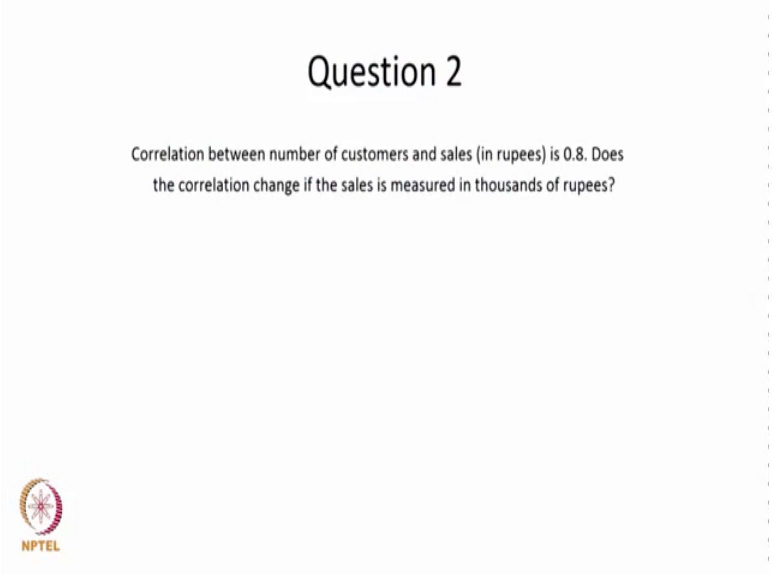
So, marks obtained in an exam with hours of study. So, as the student puts in more effort in terms of more hours of study the mark is expected to increase. So, hours of study is the x variable or the explanatory variable, while the marks obtained is the y variable or the response variable.

Number of workers and quantity produced or units produced. So, here as we put in more workers the we end up producing more quantity. Therefore number of workers is the x variable or the explanatory variable and units produced or quantity produced is the y variable or the response variable.

Third question time taken to run a particular distance and the weight of a person so, there is a general assumption that the as the person is heavy and has more weight the person would take more time to run. And therefore, in this case weight of the person can be the x variable or the explanatory variable and the time taken to run is the response variable or the y variable.

Total revenue and items sold so, again the assumption is as we sell more items or the revenue increases or the revenue is comes because of sale of items. So, items sold is the x variable or the explanatory variable, while total revenue is the y variable or the response variable. The exercise done the amount of time spent on doing exercises and the body weight. So, again there is a general assumption here in this statement that as we spend more time on exercising the body weight reduces and the body weight has an effect on the amount of time spent on exercise. Therefore, the time spent on exercise would be the x variable or the explanatory variable and the weight of the person would be the y variable or the response variable.

(Refer Slide Time: 07:55)



Question 2

Correlation between number of customers and sales (in rupees) is 0.8. Does the correlation change if the sales is measured in thousands of rupees?

Move to the next question, correlation between number of customers and sales in rupees is 0.8 does the correlation change if the sale is measured in 1000s of rupees. The answer is the correlation does not change when it is measured in 1000s of rupees or when it is measured in equivalent denominations could be even for example, you could have a set where the sale is given in rupees and then we multiply by a constant to make it into dollars or some other form of currency and as long as we multiply by the same constant the correlation does not change. So, if the sale is measured in 1000s of rupees is equivalent of dividing it by 1000. So, it does not change.

(Refer Slide Time: 08:49)

Question 3

Would correlation change if we add a constant to a variable? If we multiply by a constant?

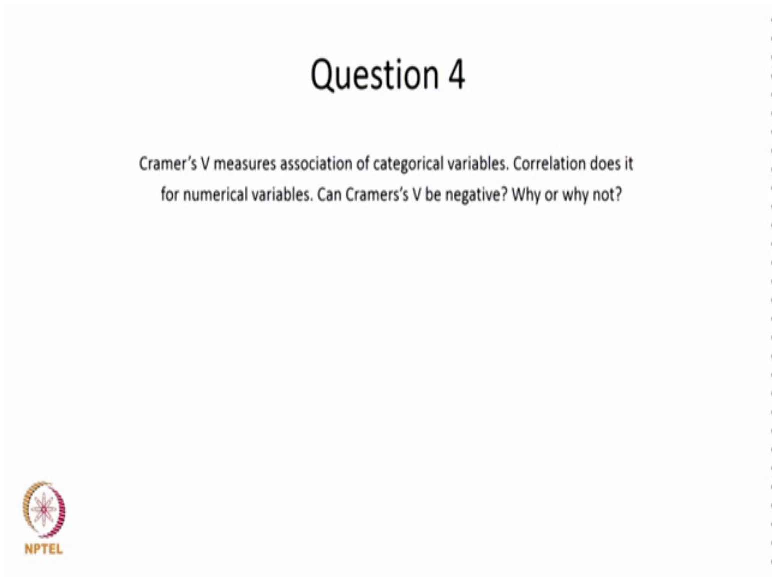
Question number 3, would correlation change if we add a constant to a variable or if we multiplied it by a constant we will answer the first part first and then the second, again the correlation would not change if we add a constant to a variable. Let us assume we are adding a constant to the y variable. So, as we add the same constant to each of the y values the assuming that the constant is positive. So, y bar would increase by the same constant and therefore, y minus y bar would remain the same in all these cases.

So, when y minus y bar remains the same in all these cases, the variance of y remains the same and the standard deviation of y remains the same, covariance would also remain the same because y minus y bar does not change and the covariance remains the same, the standard deviation remains the same and therefore, the correlation coefficient would also remain the same.

What happens if we multiply by a constant, this was the question given in the earlier question 2 when we said if it is measured in 1000s of rupees. So, when we multiply 1 by a constant let us say we multiply the x variable by a constant. So, the x bar gets multiplied by at the same constant, since x bar gets multiplied by the same constant, individual x minus x bars get multiplied by the same constant and therefore, the standard deviation gets multiplied by the same constant. And then the covariance same since x minus x bar gets multiplied by the same constant, the covariance also gets multiplied by the same constant.

Now, with respect to the standard deviation since x minus x bar gets multiplied by the same constant, when we compute the variance we square it therefore, it becomes square of the constant and then to get the standard deviation we take the square root and therefore, the standard deviation gets multiplied by the same constant, covariance gets multiplied by the constant and therefore, the correlation coefficient would remain the same because both the numerator and the denominator are multiplied by the same constant. In the case of addition the numerator and denominator remain the same therefore, the ratio is the same in case of multiplication both the numerator and the denominator get multiplied by the same constant and therefore, the ratio remains the same.

(Refer Slide Time: 11:25)



Question 4

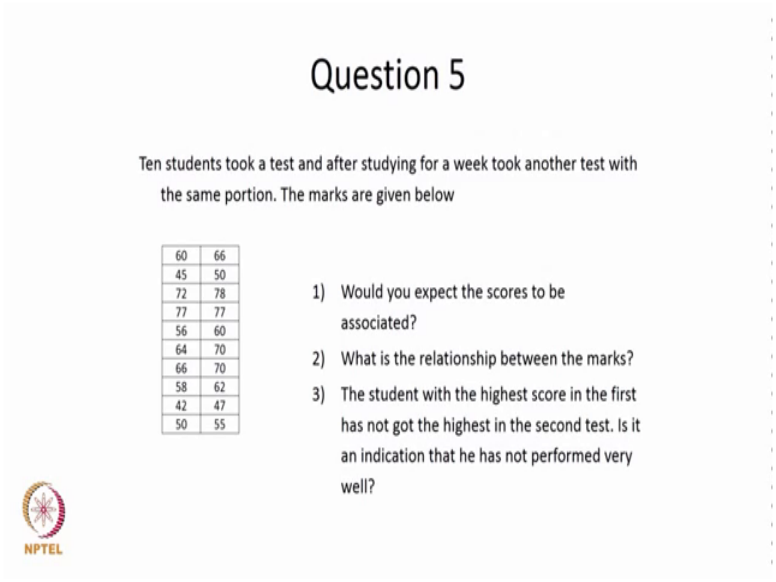
Cramer's V measures association of categorical variables. Correlation does it for numerical variables. Can Cramers's V be negative? Why or why not?

The question number 4, Cramer's V measures association among or between categorical variables correlation is used as a measure for numerical variables now correlation can be between minus 1 and plus 1, now can Cramer's V be negative why or why not. So, whatever we saw in the earlier lectures Cramer's V is the value of chi square divided by minimum of the number of rows minus 1 number of columns minus 1.

So, in the Cramer's V the denominator is a positive quantity while the numerator which is the value of chi square is also a positive quantity because or 0 because it squares numbers therefore, the way we computed Cramer's V, Cramer's V cannot take a negative value whereas, correlation coefficient also has a numerator and a denominator. The denominator part which is the standard deviations is either 0 or positive whereas, the numerator part which is the covariance can be negative and then we said correlation is between minus 1 and plus 1, plus 1 indicates some kind of a positive association and minus 1 kind of indicates a association in the opposite direction.

Now, since we look at categorical variables in Cramer's V we only check whether there is an association and we do not further qualify the association to be positive positively associated or not positively associated. Also because in categorical variables there is no question of difference between the values there is only a category and therefore, we do not further qualify the association as positive or not positive therefore, it is only fair that Cramer's V shows whether there is an association or not, but does not try to say whether there is a positive association. So, Cramer's V will take a positive value whereas, correlation can also show some kind of a negative association where as x increases y can decrease.

(Refer Slide Time: 13:39)



Question 5

Ten students took a test and after studying for a week took another test with the same portion. The marks are given below

60	66
45	50
72	78
77	77
56	60
64	70
66	70
58	62
42	47
50	55

1) Would you expect the scores to be associated?
2) What is the relationship between the marks?
3) The student with the highest score in the first has not got the highest in the second test. Is it an indication that he has not performed very well?

Ten students took a test and after studying for a week took another test with the same portions let us say the marks are given. So, would you expect this course to be associated most probably yes because we assume that when they took the first test they were still good enough and then the extra study would help them to get a slightly higher mark than what they would have got in the first test so, we would expect an association.

Now, what is the relationship between the marks? We can calculate the correlation coefficient in this case and we can also expect the marks to increase and if we actually compute the correlation coefficient which you can do as an exercise it would be very close to 1, I think in this case we get some 0.98 or something as the correlation coefficient.

The student with the highest score in the first test has not got the highest in the second is it an indication that he has not performed very well, in some ways the answer lies in the correlation. If we look at the second column the highest mark is 78 which is got by a person who got 72 in the first test whereas, the person who got 77 in the first also got 77.

If the correlation had been a plus 1 then it is quite likely that there will be an increase in each one of them since, it is not plus 1 very close to 1. So, these things can happen, but certainly that is not an indication that the person who got highest in the first has not performed well in the second. So, with this we come to the end of our discussion on association among numerical variables we will just spend a minute to summarize what we have seen in these 11 lectures and with this 11th lecture we complete the course content on statistics or introduction to statistics and then from the next lecture we move on to probability.

So, we began with defining statistics and trying to understand why we study this subject and then at some point we started understanding data and we also understood the data need not be numbers, data can also be text and information and then we learned to categorize data into 4 types of data and 2 broad types of data. And then we looked at each of these classifications categorical data and numerical data and then try to identify measures of central tendency and said for the categorical data mode and if the data is ordinal then median and if the data is numerical interval and ratio then we could have mean median and mode and then we also defined standard deviation and variance. So, they could have measures of dispersion as well with standard deviation and variance.

We also looked at for the categorical data we then looked at association and before that we also looked at the inter quartile range if the data can be sorted and ordered and then we did the inter quartile range and we also did that for the numerical data did inter quartile range and then we moved on to define measure of association between categorical data and defined chi square and Cramer’s V.

And then we moved on to define measures of association for numerical data where we looked at covariance we also looked at coefficient of variation in summarizing the data and as regards measures of association we looked at covariance and then we looked at correlation coefficient. So, with this we kind of come to the end of the course content for the statistics portion of this course and then in the next lecture we will start probability.