Introduction to R Software

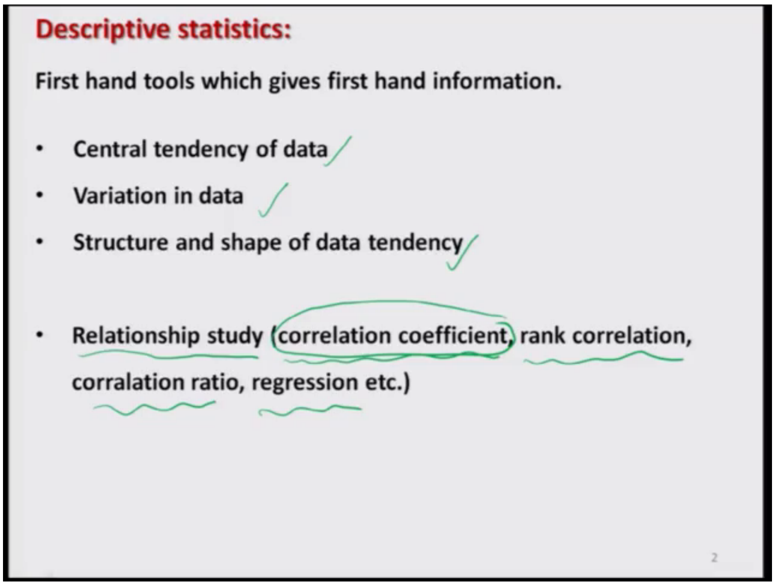
Prof. Shalabh
Department of Mathematics and Statistics
Indian Institute of Technology, Kanpur

Lecture - 40
Correlation

Welcome to the next lecture on the course Introduction to R Software. You may recall that in the earlier lecture we started our discussion on the bivariate data. Bivariate data means there are 2 variables and in order to study their relationship we have 2 options: one graphical procedure and some quantitative analytical procedures. So, in the last lecture we had discussed different types of plots and from there we had tried to see how one can analyze a relationship between the 2 variables and how one can take some inference whether there is a relationship or not. And in case if relationship is there whether it is linear, non-linear or anything else.

So, now in this lecture we will continue on the same lines and we are going to discuss some quantitative procedures to measure the degree of linear relationship between the 2 variables.

(Refer Slide Time: 01:21)



Descriptive statistics:

First hand tools which gives first hand information.

- Central tendency of data
- Variation in data
- Structure and shape of data tendency
- Relationship study (correlation coefficient, rank correlation, corralation ratio, regression etc.)

2

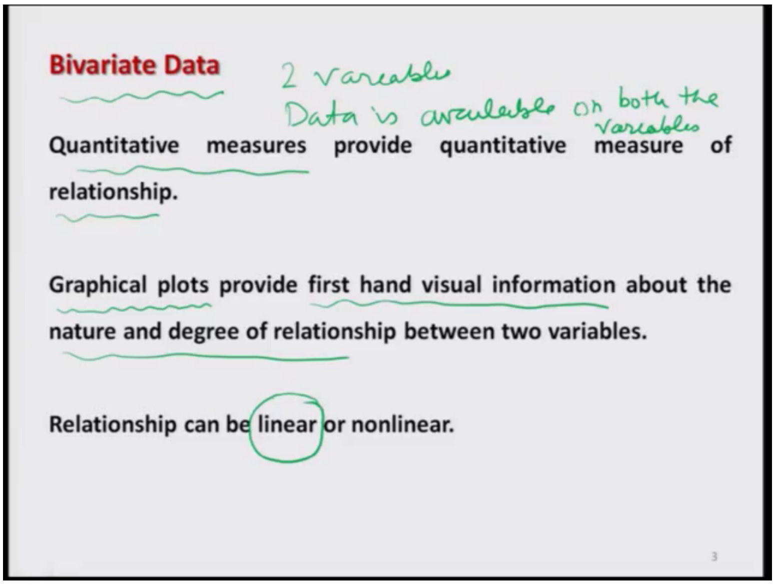
So, we start our lecture and you may recall that when we started the discussion on descriptive statistics first we took the aspect of central tendency of the data and we discussed different measures like as mean, median, harmonic mean, geometric mean and so on. And the next day was to study the variation in the data and for that we had taken different tools like as variance, mean deviation, quartile deviation etcetera. Then the next aspect what we studied was a structure and shape of the data. So, there we studied the concept of Skewness and Kurtosis and we also learnt how to compute all this things in the R software.

Now, we are going to the last aspect that we are going to consider that is to study the relationship. What do you mean by relationship? Suppose there are more than one variable, for the sake of understanding we are going to consider here only 2 variables. And suppose these 2 variables are not independent, they are related. For example, you can see whenever the height of a child increases, the weight also increases. And this process happens up to certain age. Similarly we had taken an example where we saw that whenever the temperature increases the consumption of water also increases. Somewhere there can be another variable say whenever the level of humidity also increases then also the water consumption increases.

So, these are some phenomena in which there are 2 variables and they are inter-related, they are inter-related in the sense that the happening or non-happening of one variable is causing the happening or nonhappening or say change in the other variable. So, we would like to study that how to quantify this degree of relationship and here in this lecture we are going to study how to quantify the degree of linear relationship. And in order to study the relationship there are different types of tools available in statistics, one is correlation coefficient, rank correlation, correlation ratio, regression and so on.

But here our objective is not to study the statistics. We are here just to show you that those statistical tools can be computed and can be used through the R software. So, here in this lecture we are going to talk about the correlation coefficient and we will try to show you first what is correlation coefficient, a brief introduction; what is its interpretation and how we are going to compute it.

(Refer Slide Time: 04:40)



Bivariate Data

2 variables

Data is available on both the variables

Quantitative measures provide quantitative measure of relationship.

Graphical plots provide first hand visual information about the nature and degree of relationship between two variables.

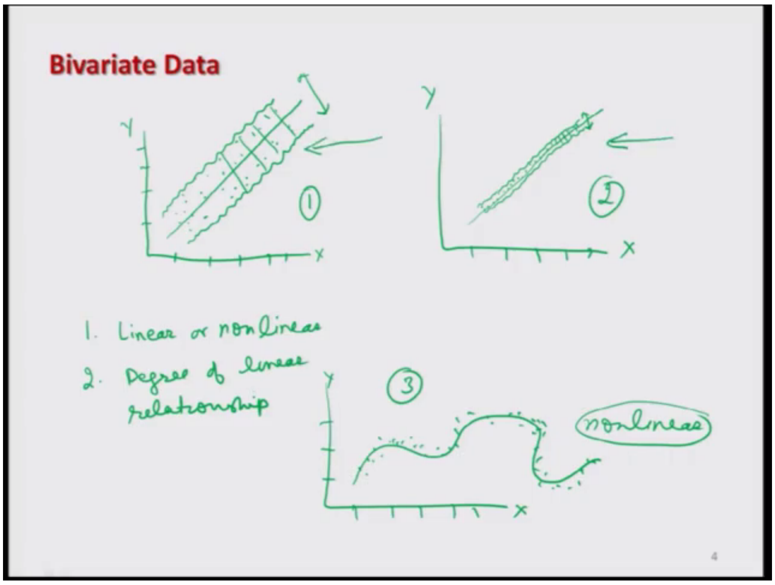
Relationship can be linear or nonlinear.

3

So, this lecture is going to have a statistical flavor once again. So, now, we consider a bivariate data. Bivariate data means there are 2 variables and we have collected the data on both the variables. So, data is available on both the variables, right and we want to quantify this relationship. So, there are certain quantitative measures and they provide the quantitative measurement of the relationship.

Whenever the data comes first we try to use the graphical plots and they give us the first hand visual information about the nature and degree of the relationship between the 2 variables. And most over this relationship can be linear as well as non-linear also.

(Refer Slide Time: 05:51)



Bivariate Data

1. Linear or nonlinear
2. Degree of linear relationship

So, now here we are going to consider only the linear relationship. And for that we have got a measure what is called as correlation coefficient, but before that let us try to see this picture. Suppose I try to make here 2 pictures or rather here 3 pictures and suppose here is my x, y, x, y axis, x and y axis and here I am trying to plot a data what is available to us and you can see here that this is trying to show a trend which is non-linear.

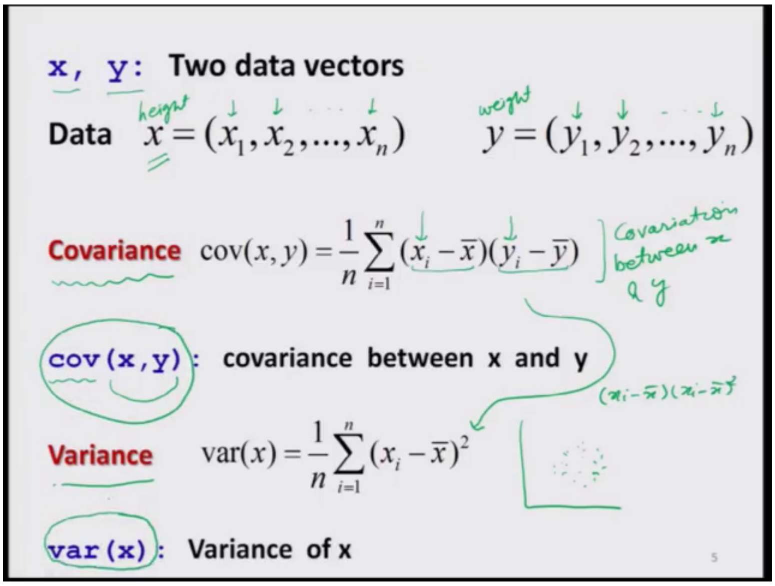
So, that is clear to us, no issues. Now here I am going to consider here linear trend data, it is something like this. And then there is another data set which is something like this and we believe that the scaling on all the figures that is the same. Let me call this as figure number 1, this is figure number 2 and this is my figure number 3. So, now what do you really observe? Between the figures 1, 2 and 3 it is clear that three is non-linear; that means, the trend in the data is showing that there can be a non-linear relationship.

Whereas in case if you try to concentrate on figure number 1 and 2 yes, I can say that the relationship between x and y is quite linear and one possible line may be drawn something like this over here. But now you can see here what is the difference between the 2 figures that is 1 and 2. In figure number 1 you can see that all the lines are lying close to the line and they can possibly be enclosed in a band like this one. This is a zigzag band. And in the figure 2 these points are also not exactly lying on the line and they can be enclosed in a band like a zigzag line which I have drawn.

But definitely you can see here the width of this zigzag band and width of this zigzag band in figure 1 and 2 are different. The width of this band here is more than the width of the zigzag band in figure number 2; that means, I can conclude that in figure number 1 and 2 the data is lying near the line. But in figure number 1 the data is more scattered around the line and in figure number 2 the data is more close to the line. So, now, if you try to see here we are talking of 2 features first feature is whether the relationship is linear or non-linear.

And second thing is this we are talking of degree of linear relationship, what do you mean by degree of linear relationship? For example, if you try to observe the figure in say 1 and 2 means I can say that the degree of linear relationship in figure 2 is more than the degree of linear relationship in figure number 1.

(Refer Slide Time: 09:59)



x, y: Two data vectors

Data $x=(x_1,x_2,\ldots,x_n)$ $y=(y_1,y_2,\ldots,y_n)$

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

cov(x,y) : covariance between x and y

Variance $\operatorname{var}(x)=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$

var(x) : Variance of x

height
weight
Covariation between x & y
$(x_i-\bar{x})(x_i-\bar{x})$
5

So, now the question is how to measure it. So, for that we have a a concept of correlation coefficient and this correlation coefficient depends on a quantity which is called as covariance. So, first we try to understand this thing and then the correlation coefficient. Suppose I have got here 2 data vectors x and y.

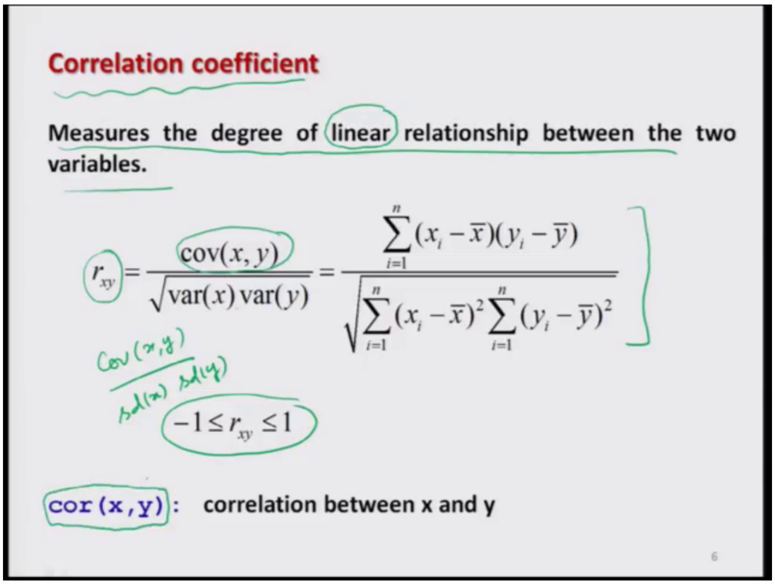
So, as we had discussed earlier x is a data vector in the sense that for example, x can be my variable say height and y can be my variable here as a weight. And now I am observing say smaller number of observations on say height; first value, second value they are denoted at x 1 and x 2 and similarly we have say n values. And similarly, for the weight also we have got the first value y 1, second value y 2 and so on we have got y n nth value. So, all x 1, x 2, x n, y 1, y 2, y n, they are some numerical values of height and weight respectively.

Now, in case if you try to remember earlier we had discussed the concept of variance. The variance was trying to measure the degree of a scatteredness around some point, right. Now try to extend this concept, now suppose there are 2 variables. Then we can compute a quantity what is called as covariance. Covariance is defined as say the arithmetic mean of the product of divisions of observations on x and y from their respective arithmetic mean, what is this mean? We try to understand. If you try to see this is my here x i this is the observation on first variable ith observation on first variable and similarly this is my y i which is the ith observation on the second variable.

Then I am trying to see the difference between the 2, that is x i minus x bar that is the deviation from the arithmetic mean for x i and similarly we observe the deviation of y i from the arithmetic mean and then I am trying to take their product and then I am trying to find out the average of their product. So, I can believe that this quantity will give us an idea about the covariation, covariation mean joint variation between the 2 variables x and y.

And in case if you try to see between covariance and variance, in case if you try to say take only 1 variable for both, 1 variable for x i and the same variable for y I, that is x i. So, this becomes a simply here variance, right. So, anyway our objective is how to compute this covariance in R. For that we have the syntax as cov and then inside the argument we have to give that 2 data vectors x and y and this is how we can compute the covariance between 2 data vectors. And similarly we already have done that in order to compute the variance the command is var and the inside the argument we have the data vector.

(Refer Slide Time: 13:36)



Correlation coefficient

Measures the degree of linear relationship between the two variables.

$$r_{xy} = \frac{\text{cov}(x,y)}{\sqrt{\text{var}(x)\,\text{var}(y)}} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

$$\frac{Cov(x,y)}{sd(x)\,sd(y)}$$

$$-1 \le r_{xy} \le 1$$

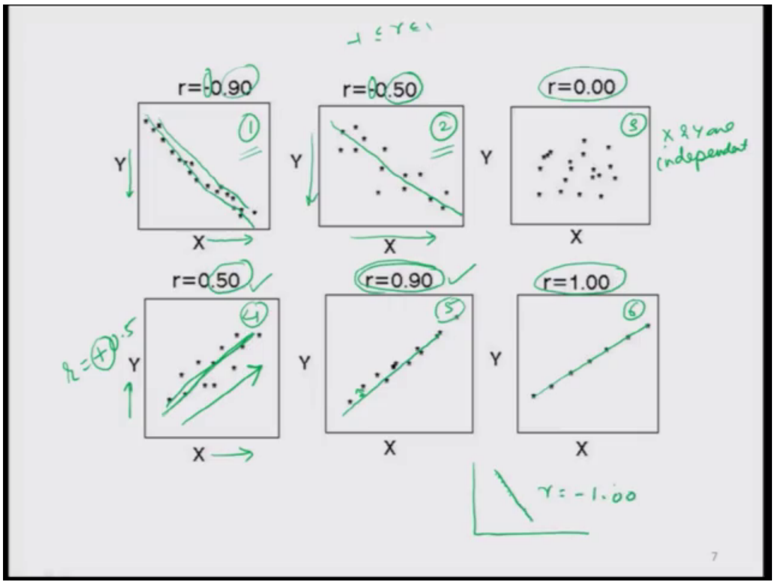
`cor(x,y)`: correlation between x and y

6

So, now based on that we define the correlation coefficient, right; his correlation coefficient is essentially a measure of the degree of linear relationship between the 2 variables. A very important point is linear relationship, the relationship can be linear or say non-linear, but correlation coefficient measures only the degree of linear relationship. And this coefficient is defined as here r xy which is the ratio of covariance of xy and the square root of variance of x into variance of y.

So, essentially this is actually covariance between x and y upon standard deviation of x into standard deviation of y which is given by here this expression. And this correlation coefficient lies between minus 1 and plus 1. This correlation coefficient can be computed in r using the syntax cor and inside the argument we have to give the data vector, right. So, the next question comes what is the interpretation of this coefficient.

(Refer Slide Time: 15:00)



So, we try to look at this figure. Now in this figure, let us call them as figure number 1, 2, 3, 4, 5 and here 6.

Now, we try to see what is really happening. In figure number 1 all the points they have a decreasing trend like this one, but if you try to make a line passing through most of the points then points are very very close to the line. So, this decreasing trend is indicating that as the values of x are increasing, the values of ys are decreasing. So, this decrement is indicated by this negative sign. And since the points are very very close to the line and since we know that r lies between minus 1 and plus 1. So, this value is 0.90 which is very very close to 1.

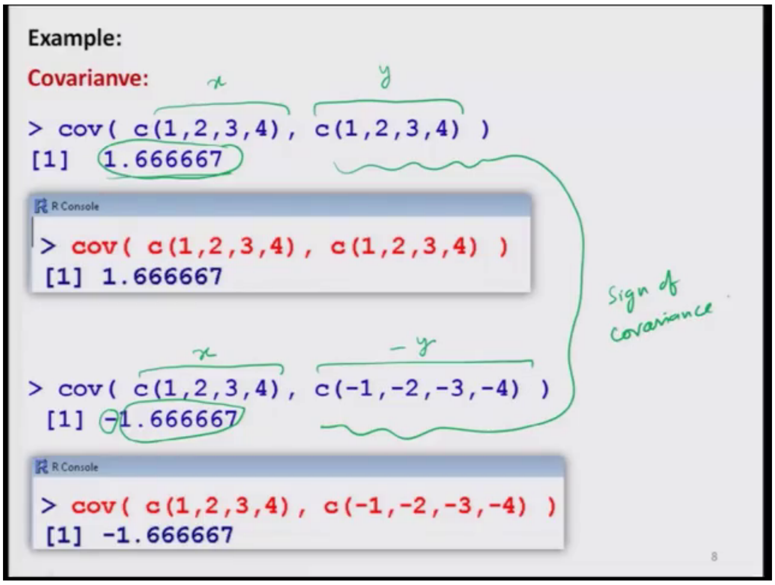
Similarly, when you come to figure number 2, I can see here a downward trend, but now the points are not so close as in figure number 1. So obviously, the value of the correlation coefficient in second case is going to be smaller than the value in the figure number 1 and since this is decrement the relationship is decreasing in the sense as the value of x are increasing the values of y are decreasing. So, this is indicated by this sign and this is the value 0.50 for the correlation coefficient. So, the value of correlation coefficient in the first figure is 0.9 and in the second figure this is 0.5. So, that is indicating that the points are not so close to the line.

Now, if you come to figure number 3 here I cannot see any pattern of relationship between x and y. So, in this case the value of correlation coefficient is nearly 0 and we say that x and y are independent, they have no relationship. Now we come to figure number 4, here you can see that the trend in the values with respect to the relationship between x and y is increasing; that means, as the values of x are increasing the values of y are also increasing.

And this is indicated by the fact that the value of r here is reported as say plus 0.5. So, this increasing trend is indicated by the positive sign and since the values are quite close to the line here that is indicated by the value of the correlation coefficient as 0.5. And similarly, when you come to the figure number here 5, you can see here the trend is again increasing. And the value of the correlation coefficient here is 0.9 which is indicating that since the points are more close to the line here than in the figure number 4 that is why this value r equal to 0.9 is higher than the value of r in figure number 4 as 0.5.

And finally, in the figure number 6, you can see here that all the points are lying exactly on the line, there is no deviation. And this is indicated by the fact that r takes the maximum value plus 1. And similarly, in case if all the points have a decreasing trend, but they are lying exactly on the same line then in this case the value of correlation coefficient will be minus 1.00 and so on. So, that is the interpretation of the value of correlation coefficient and that by looking as the value of correlation coefficient as well as the sign you can decide whether the relationship is increasing or decreasing and what is the magnitude of degree of linear relationship, right.

(Refer Slide Time: 19:28)



Example:

Covarianve:

```
> cov( c(1,2,3,4), c(1,2,3,4) )
[1]  1.666667
```

```
> cov( c(1,2,3,4), c(1,2,3,4) )
[1] 1.666667
```

```
> cov( c(1,2,3,4), c(-1,-2,-3,-4) )
 [1] -1.666667
```

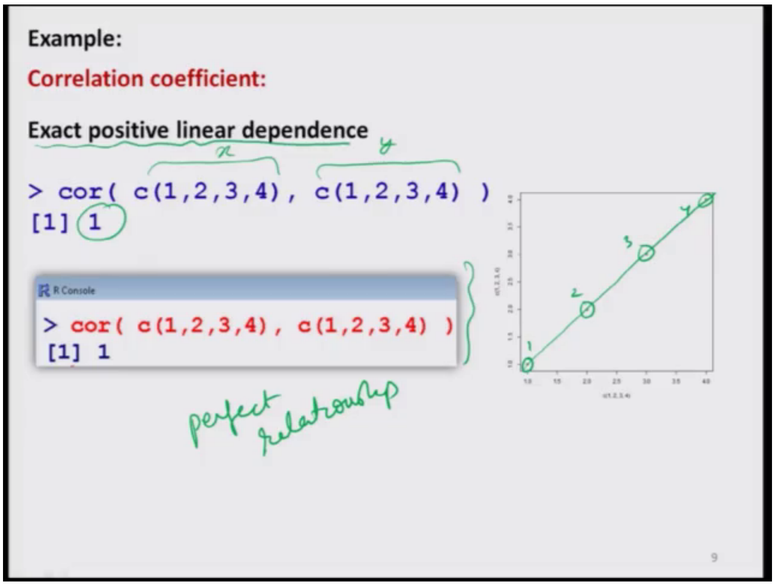
```
> cov( c(1,2,3,4), c(-1,-2,-3,-4) )
[1] -1.666667
```

Now, we try to implement the covariance and correlation function on the R software. So, if you try to see here, I am trying to take here first 2 vectors, and both the vectors are identical; 1, 2, 3, 4, 4 values combined in a vector. So, this is my here something like x vector and this is my here y vector. So, you can see here the value comes out to be a 1.66 and so on.

And similarly, when I try to take the same here x in the second case this is here same as x, but in the second vector I try to compute this covariance with minus of y. So, all the values in this vector and in this vector, they are the same but only the sign is changing. And you can see here both the values of covariance are 1.66, the difference is coming only through the negative sign. So, that is really indicating that in case if the relationship is positive or negative that is determined by the sign of covariance.

The covariance is responsible for informing us the sign that is the relationship between x and y is increasing or decreasing, right.

(Refer Slide Time: 20:52)



Example:

Correlation coefficient:

Exact positive linear dependence

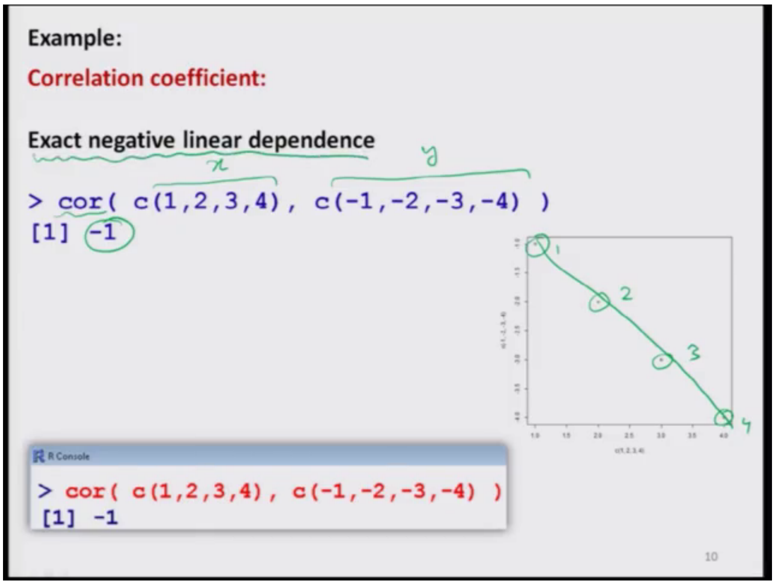
```
> cor( c(1,2,3,4), c(1,2,3,4) )
[1] 1
```

R Console

```
> cor( c(1,2,3,4), c(1,2,3,4) )
[1] 1
```

And then I try to find out here is the correlation between the 2 vectors; 1, 2, 3, 4 and 1, 2, 3, 4, what do you expect? There is a perfect relationship and you can see here this is point number 1, point number 2, point number 3 and point number here 4 on the graphic. And these point are exactly lying on the straight line and this is indicated by the correlation value between x and here y which is here 1 and this shows that there is an exact positive linear dependence between x and y.

(Refer Slide Time: 21:38)



Example:

Correlation coefficient:

Exact negative linear dependence

```
> cor( c(1,2,3,4), c(-1,-2,-3,-4) )
[1] -1
```

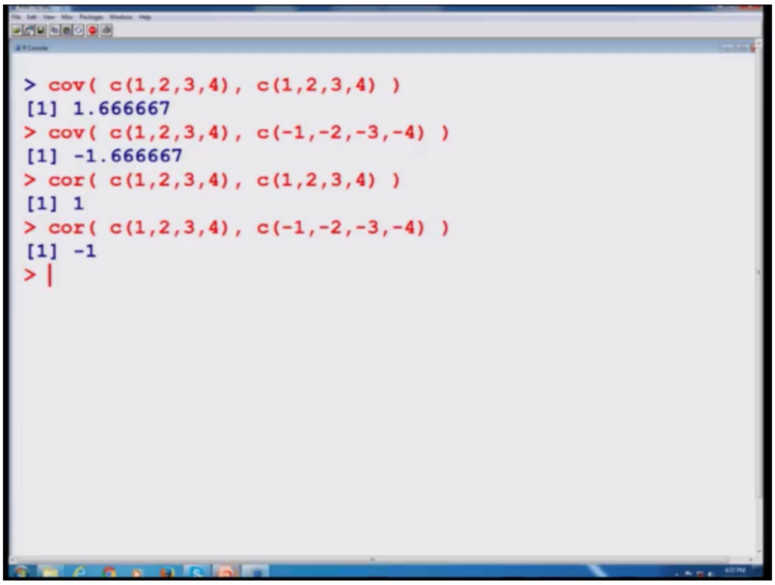
R Console

```
> cor( c(1,2,3,4), c(-1,-2,-3,-4) )
[1] -1
```

And now we take another example in which I try to take the same vectors, x is taken here as say x vector of 1, 2, 3, 4 values; but I try to take the y vector here as say minus 1, minus 2, minus 3, minus 4 and I try to find out the correlation between them by cor and you can see here this is minus 1.

Now why it is minus 1? You can see here in the graphic this is my point number 1, this is my point number 2, this is my point number 3 and this is my point number 4 and there are exactly lying on the straight line and the relationship is decreasing. So, in case if I try to find outs the correlation coefficient in such a situation, this indicates the exact negative linear dependence. And incase if you try to take another value that will show you the same thing. And we try to do this thing on the R and try to see whether these things are working or not.

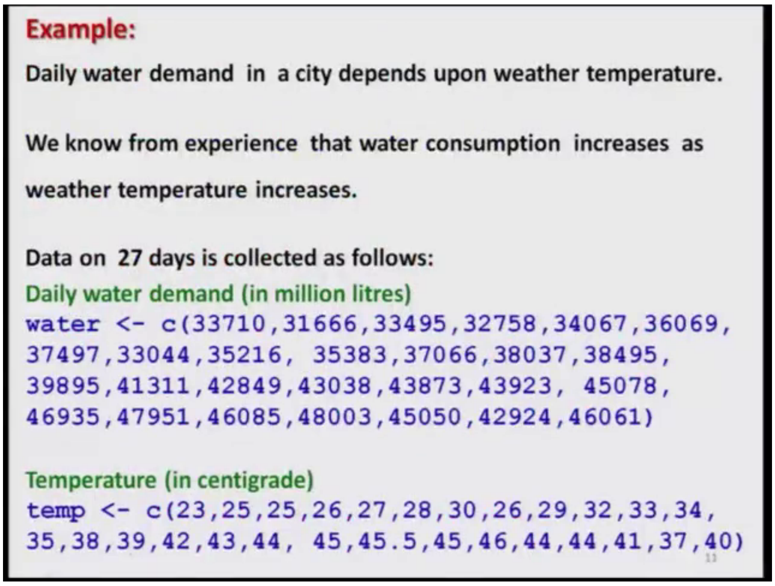
(Refer Slide Time: 22:38)



```
> cov( c(1,2,3,4), c(1,2,3,4) )
[1] 1.666667
> cov( c(1,2,3,4), c(-1,-2,-3,-4) )
[1] -1.666667
> cor( c(1,2,3,4), c(1,2,3,4) )
[1] 1
> cor( c(1,2,3,4), c(-1,-2,-3,-4) )
[1] -1
> |
```

You can see here this is the covariance between 2 is identical vector and now this is the covariance between 1 vector and another vector is just changing by the negative symbol. These values comes out to be the same, the only difference is the negative sign. Now in case if I try to find out the correlation. So, the function is cor between the 2 identical vectors this is coming out to be 1 and the correlation between the 2 vectors 1, 2, 3, 4 and another vector having same magnitude, but opposite sign this comes out to be minus 1. So, you can see that the things are working, okay.

(Refer Slide Time: 23:33)



Example:

Daily water demand in a city depends upon weather temperature.

We know from experience that water consumption increases as weather temperature increases.

Data on 27 days is collected as follows:

Daily water demand (in million litres)

```
water <- c(33710,31666,33495,32758,34067,36069,
37497,33044,35216, 35383,37066,38037,38495,
39895,41311,42849,43038,43873,43923, 45078,
46935,47951,46085,48003,45050,42924,46061)
```

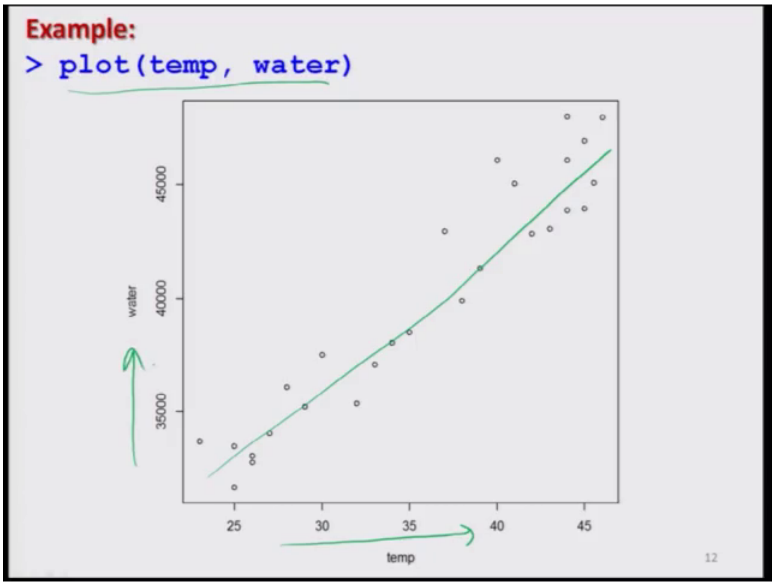
Temperature (in centigrade)

```
temp <- c(23,25,25,26,27,28,30,26,29,32,33,34,
35,38,39,42,43,44, 45,45.5,45,46,44,44,41,37,40)
```

Now, in order to understand this thing for that I try to take some more examples. So, now we take one example and from there we try to do a complete analysis. And we are trying to take the same example that I considered in the last lecture where I had a plotted the different types of plots between the consumption of water and weather temperature. So, just to recall, we had collected the data on 27 days and then we had collected the data on say a daily water demand for those 27 days in million milliliters and this data was collected inside a vector water, and it was combined and the data on the temperature was combined in say another vector here temperature say temp and this was measured in centigrade.

And in the last lecture I have given you the detailed information about this data set, but now my objective is that I expect that there can be a sort of linear or say non-linear relationship between the daily water consumption and temperature because I expect that as the temperature increases, the consumption of water increases.

(Refer Slide Time: 24:50)

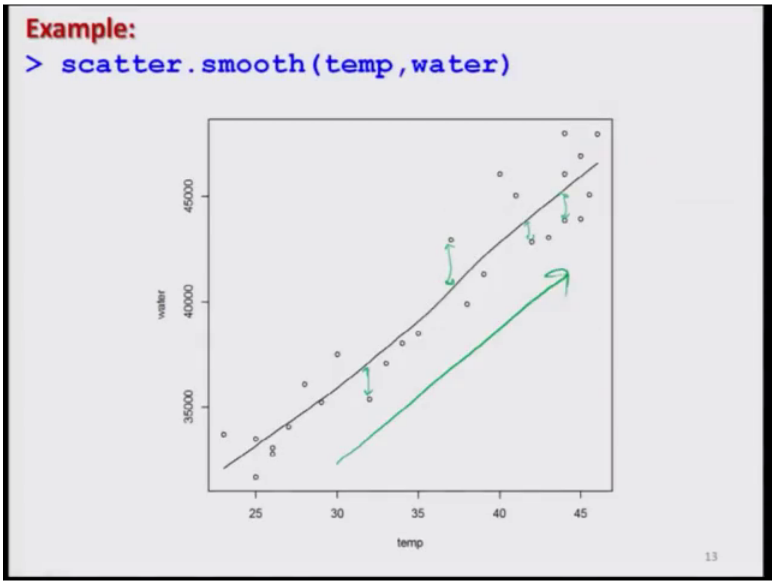


Example:

```
> plot(temp, water)
```

So, I try to first make a plot. So, if you try to see I have made here a plot between temperature and water and you can see that there seems to be a reasonable linear trend and I can say that the values are increasing and what we expected that as the values of temperature are increasing the consumption of water is also increasing. So, this gives us a confidence.

(Refer Slide Time: 25:11)

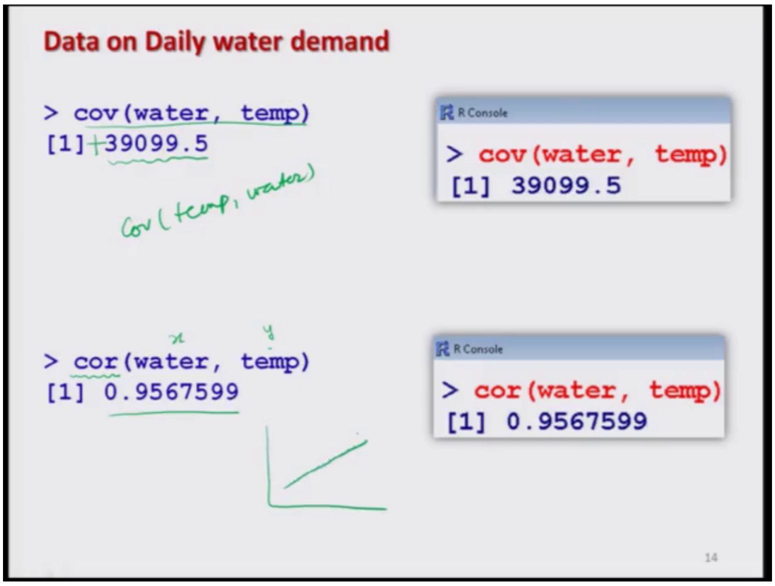


Example:

```
> scatter.smooth(temp,water)
```

And can I try to make a scatter is smooth and we can see that here that that if I try to plot this line then the points are lying quite close to the line, but there is some different like a here, here, there is some difference you can see here. So, now I want to measure this this degree of linear relationship.

(Refer Slide Time: 25:35)



Data on Daily water demand

```
> cov(water, temp)
[1] 39099.5
```

```
R Console
> cov(water, temp)
[1] 39099.5
```

```
> cor(water, temp)
[1] 0.9567599
```

```
R Console
> cor(water, temp)
[1] 0.9567599
```

And so, I try to first find out the covariance between water and temperature. You see finding out the covariance between water and temperature and covariance between temperature and water, this will yield the same thing there is no difference, okay.

So, in case if you try to find out the covariance between water and temperature, this comes out to be like this 39099. And now we try to find out the correlation, but before that you can see here the sign of the covariance here is positive and this confirms that whatever we have observed in this figure that the trend is positive that is confirmed, right. Now we try to find out the correlation by the function cor between water as x and temperature as say here y. And this correlation comes out to be here 0.95, what is this mean? That the relationship is increasing and the quantitative measure of the degree of linear relationship is 0.95, right.

Had all the points be lying exactly on the same line then this correlation coefficient would had been 1. But that is not really practically feasible in the real-life situation. So, you can see here what were is the deviation of these points from the line that is now quantified by the correlation cor.

So, we stop here and I would request you to take some example and then try to practice the covariance and correlation function. And now in the next lecture we will come up with some more topics. So, see you in the next lecture, till then goodbye.