Descriptive Statistics with R Software

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

Lecture- 27
Association of Variables - Quantile – Quantile and Three Dimensional Plots

Welcome to the next lecture on the course Descriptive Statistics with R Software. You may recall that in the earlier lectures we started our discussion on association of two variables and we had discuss several types of two dimensional plot. And those two plots were trying to give an idea about the direction and the degree of association between two variables.

Now, I am going to consider two dimensional plot, but they are used in different concepts. So, in this lecture I am going to talk about Quantile Quantile Plot and after that I will give you a brief introduction and brief description about the Three Dimensional Plots which are available in R software.

Now, there is a situation that two samples have been obtained from some population. Now, we want to know on the basis of given sample of data that whether that two samples have been drawn from the same population or different population, this type of a situations are very very useful in several statistical tool. For example, one good example is in the case of testing of hypothesis we try to conduct one sample test, two sample tests and so on. In those types of test there is a requirement that sample is coming from a particular type of population and those populations are characterized by some probability density function.

For example, you must have studied in the books that are popular sentences let x 1, x 2 x n be a random sample from normal population with mean mu and variance sigma square. So, definitely in this case we are trying to say that there is a population which is very big and practically unknown to us and this population is characterized by the normal density function.

Now, I am drawing a small sample say 20 observation, 30 observation or say 100 observations and so on. And I would like to know whether the sample is coming from a normal population or not and this type of information is needed because the tools like the

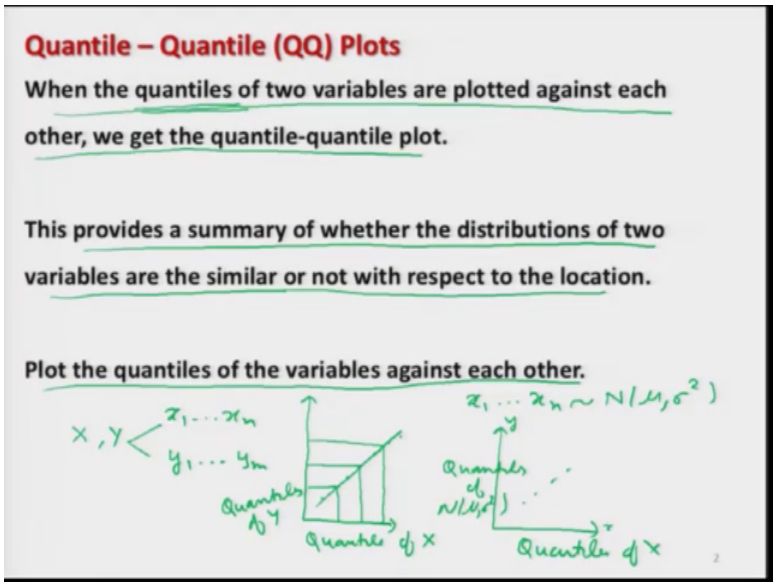
t test z test and so on and different types of tests which are use in the testing of hypothesis, they are constructed assuming that the population is characterized by normal distribution.

So, unless and until this assumption is verified, the further statistical inferences will be questionable. So, one question I would like to address here is that how to judge that a particular given sample is coming from our normal population and if I try to extend this concept. Then I would say suppose 2 persons have brought two samples to me and I would like to know whether these samples are coming from the same population or from different population.

And in order to makes such an comparison one option is that, I can compare the quantiles of the samples and then I can conclude that whether the two samples are coming from the same population or not, in case if they are coming from the same population I would expect that the quantiles of both the populations are going to be the same. And similarly when I want to test whether a sample is coming from a normal population or not.

Then I would try to compare the quantiles which are computed on the basis of the sample and the quantiles of the normal probability density function, if they match then I can say yes my sample is coming from a normal population with certain mean and certain variance. So, these types of plots are called as quantile quantile plots. So firstly, let us try to understand the concept interpretation and then I will show you how to use them on the R software.

(Refer Slide Time: 05:18)



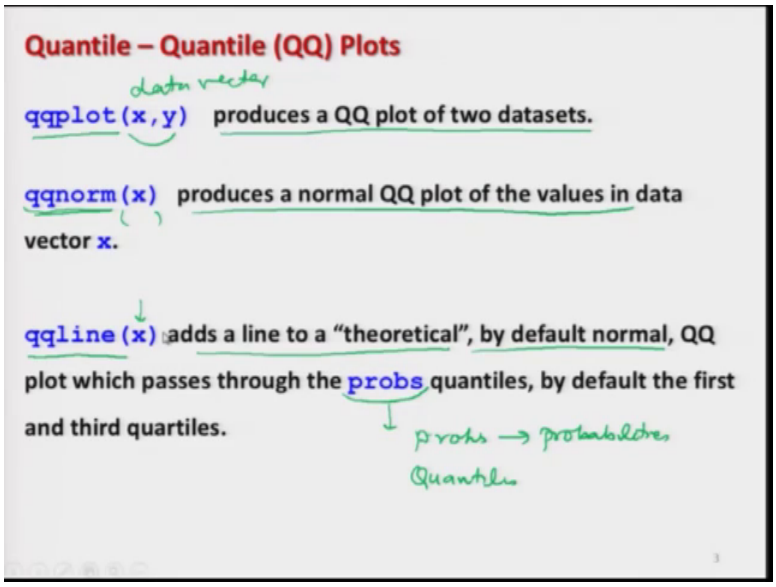
So, in the case of quantile quantile plots what we try to do? That we try to consider two variables and we try to find out their quantiles and when the quantiles of the two variables are plotted against each other in a two dimensional plot, then we get the quantile quantile plot. For example, if I say if I have got here two variables X and Y and based on that I have got two samples say x 1, x 2, x n and say y1, y 2 say y m these samples may be of same size or of different size.

And then I would try to plot the quantiles of X and on the Y axis quantiles of y and then I would try to see that how they are matching and in case if they are matching, then I would say yes they are coming from the same population otherwise not. So, it is something like this 25 percent quantile of x and 25 percent quantile of y and say 40 percent quantile of x and 40 percent quantile of y, say 70 percent quantile of x and 70 percent quantile of y if they are matching, then I can join this line and that is going to be a straight line.

And similarly if I want to compare or if I want to know that a sample x 1, x 2, x n is coming from a normal population with mean mu and sigma square, then in this case I would try to plot the quantiles of x on 1 axis and say quantiles of normal distribution normal mu sigma square on the y axis and then I would try to see the pattern.

So, in case if I try to do so these types of graphics they provide us a summary whether the distribution of the two variables are similar or not with respect to the location and we try to plot this quantiles of the variable against each other.

(Refer Slide Time: 07:40)

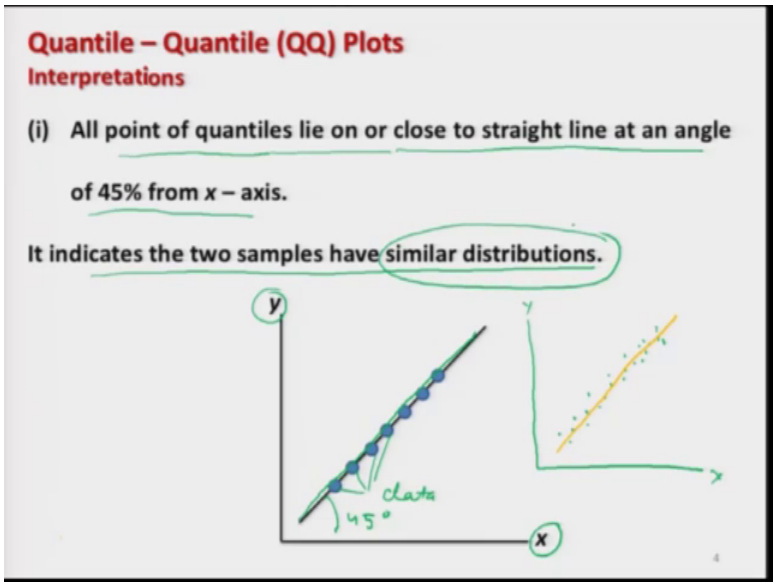


In order to plot such quantiles in R software we have a command here qqplot and then inside the arguments we give the data vector. So, I have here two data vectors x and y and we use the command qqplot and inside the argument x comma y and this is going to give us a QQ plot of two data sets. Similarly, there is another command qqnorm this qqnorm produces a normal quantile quantile plot of the values in the data and in this case they are compared with the quantiles of the normal distribution.

So, here the command is qqnorm and inside the argument you have to give the data vector x. Similarly, there is option that inside this normal qqplot we can add here a line and this line is based on the theoretical quantiles and by default this is normal. And these quantiles whatever are being plotted, they can be control by the probs function. You may recall that we had use the probs function to define the probabilities or they were trying to indicate that the quantiles have to be computed for which of the probabilities.

So, you may look into the lecture on the quantiles we have where we had used this probs function. So, the command here is qqline and inside the arguments we give the data vector. So, this will also plot a line inside the qqplot.

(Refer Slide Time: 09:44)

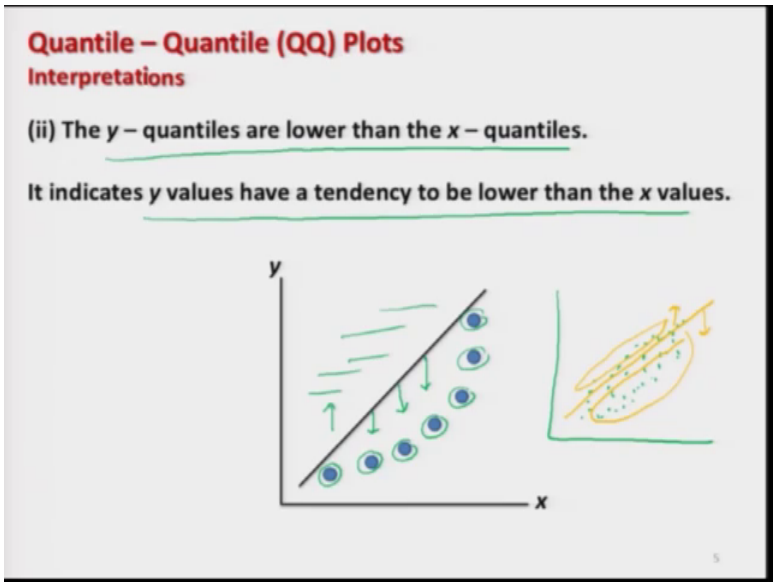


So, the first question comes that how to make interpretations from this quantile quantile plots? So, I will try to take here different types of possible situations and based on that I will try to show you that how we are going to take a conclusion or draw a statistical inference.

Suppose I try to plot the quantiles of data on x and data on y on a two dimensional plot again the x axis x and y, these dots they are trying to show the data. And if you try to see in this case all this data that is lying on a straight line straight like this and this line is essentially a 45 degree or this line is made at an angle of 45degree from the x axis.

So, in this case you can see that all the points of quantiles they are lying on a straight line which is drawn at an angle of 45 degree from the x axis. So, this is indicating that the two samples have the similar distribution. And in practice it is always not possible to get such a 100 percent clear straight line, but the plot will look like this. So, in this case in case if I try to plot here a trend line, this will look like this and you can see that the points are lying nearly on the straight line. So, in this case we can say that yes the two samples are coming from two populations which have got the similar distributions.

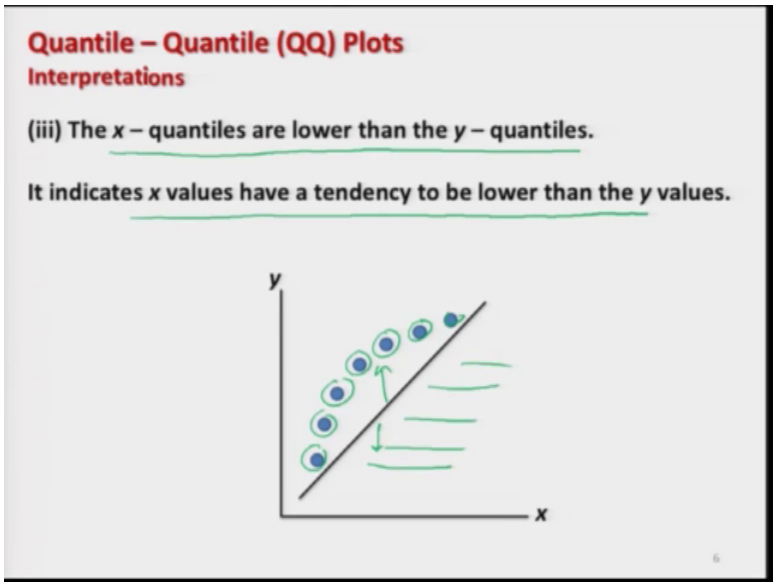
(Refer Slide Time: 11:51)



Similarly, in further case suppose we get a quantile plot like this one where all these data points they are lying below the straight line and no point is lying in this direction here in this region. So, in this case what I can conclude is that the y quantiles are lower than the x quantiles and this has an interpretation that y values have a tendency to be lower than the x values. This obviously, indicates that the distribution from where the samples have been drawn on x and y they are not the similar.

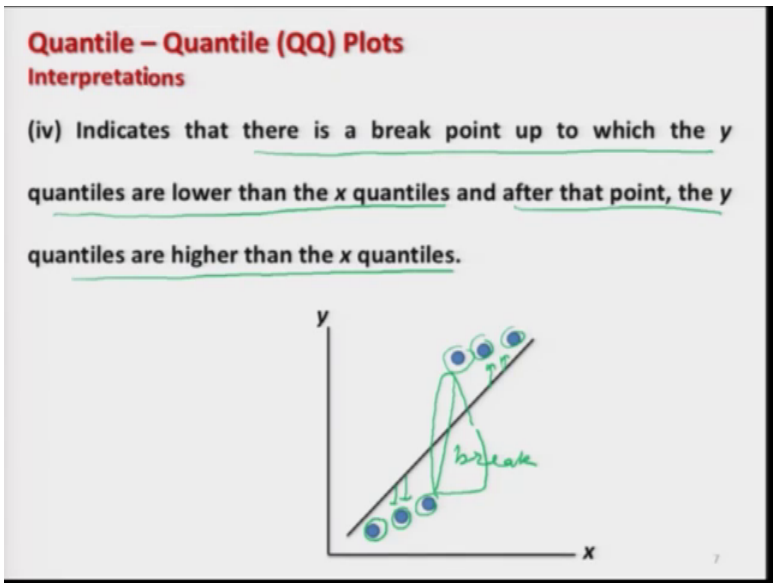
And in practice in case if you are getting a data like this one and suppose the trend line is passing through like this one. So, you can see here; you can see here that most of the points are lying in the lower region below the line and there are very few points which are lying above the line. So, in this case in general I can say that the y values have a tendency to be lower than the x values and hence the distributions are not the same.

(Refer Slide Time: 13:24)



Similarly, the opposite of this can also hold true that all the data points are lying above the line and there is no data point here in the lower side of this line and this is indicating that the x quantiles are lower than the y quantiles and this has an interpretation and it is indicating that the x values have a tendency to be lower than the values of y. And hence the two samples from x and y they are not coming from the same distribution.

(Refer Slide Time: 14:10)



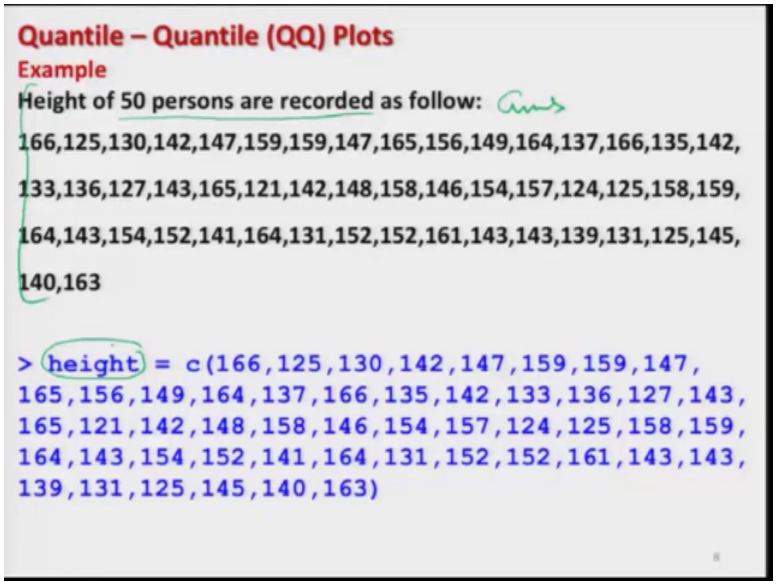
Similarly, in case if you are getting a QQ plot or the quantile quantile plot in this way where you can see that here there are some data on the lower side of this line and

suddenly there is a break. And after that there are few points towards the end and these points are above the line. So, in this case this is indicating that there is a break point up to which the y quantiles are lower than the x quantiles and after that point the y quantiles are higher than the x quantile.

So, you can see here that this is the region where there is a break point and these quantiles are lying in this direction and in the upper part they are lying in the upper direction from the line. So, in this case also we can interpret that the two samples which are coming from two different populations and those populations are not the same.

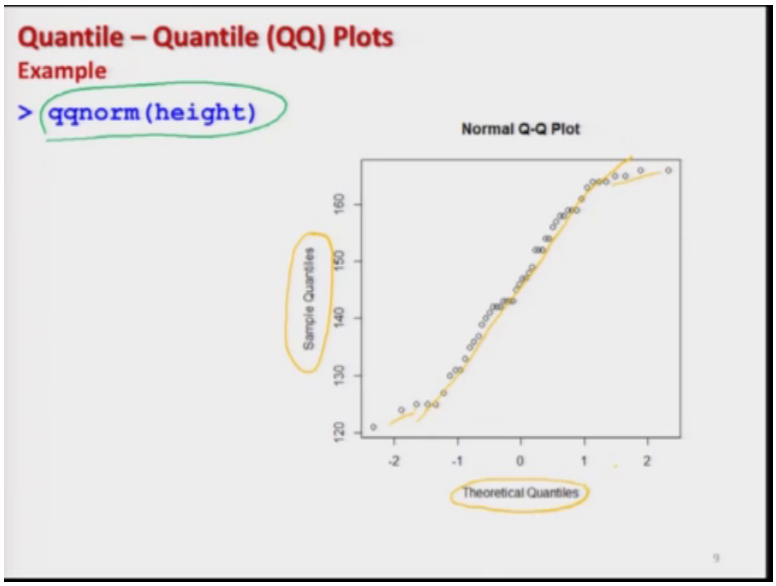
Now, I can do one thing that I am trying to take a two data points say or say two samples one from x and one from y. And similarly, in case if I try to take one of the quantile to be the theoretical quantile from the normal distribution, then I can compare the quantiles of a data set with the quantiles of a normal distribution.

(Refer Slide Time: 15:45)



So, let me try to show you through an example and I will try to make this quantile quantile plot or popularly they are called as QQ plots using this (Refer Time: 15:52) set. So, now this data set is the same example that I have use earlier couple of times and this is about the heights of 50 percents which are recorded in centimeters right and this data is collected inside the data vector height.

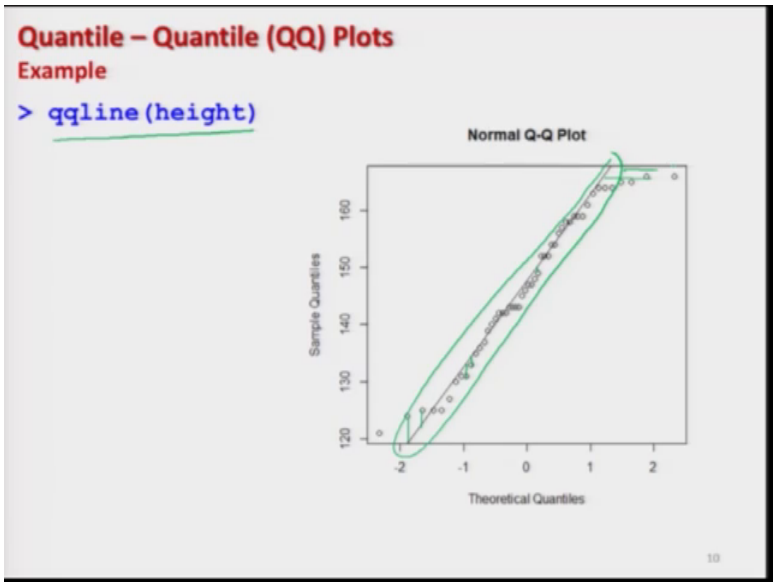
(Refer Slide Time: 16:14)



And now after this I try to first prepare a qqnorm of this data set. So, now you can see here that these points are lying here these dots are indicating the quantiles and if you try to see this line looks something like this. So, you can see that approximately it looks linear, there are some points over here and here which are going beyond the lines, but here you can see that most of the points are lying on the straight line.

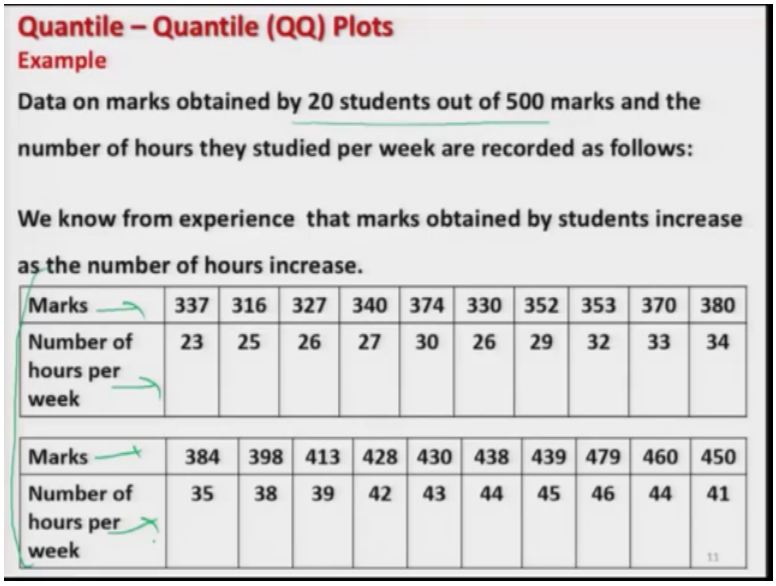
So, possibly I can compute or I can conclude that the quantiles which are computed and indicated on the y axis on the basis of given sample and the quantiles of the normal population which I have been computed using the PDF of normal distribution and these are the theoretical quantiles they are matching they are nearly matching. So, one can safely assume that this data is coming from a normal population.

(Refer Slide Time: 17:27)



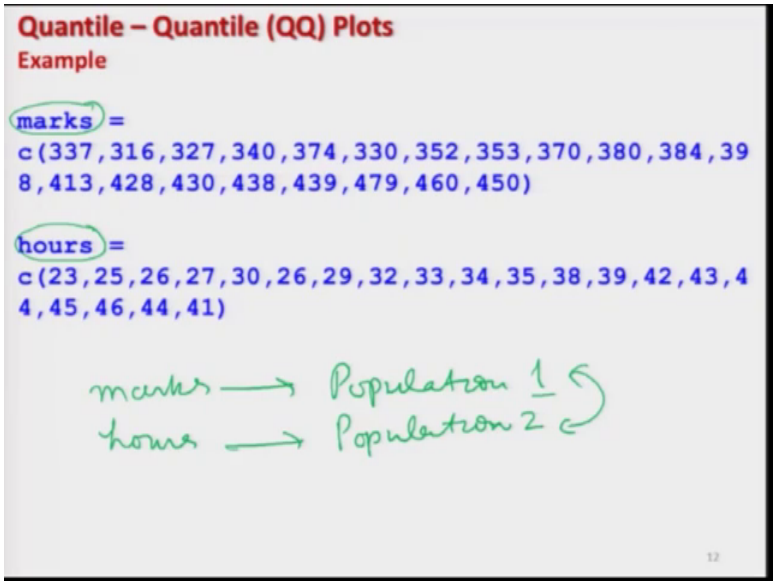
Now, in case if I try to add here a line using the command qqline. So, the command will be qqline and height and you can see here that this line has been added to the same quantile quantile plot. So, that is helping us in comparing that how much is the deviation of this points from this line you can see here this deviation is less and in the starting and towards the end this deviation is here more something like this. So, this will help us in taking a conclusion whether the samples are coming from a normal population or not.

(Refer Slide Time: 18:08)

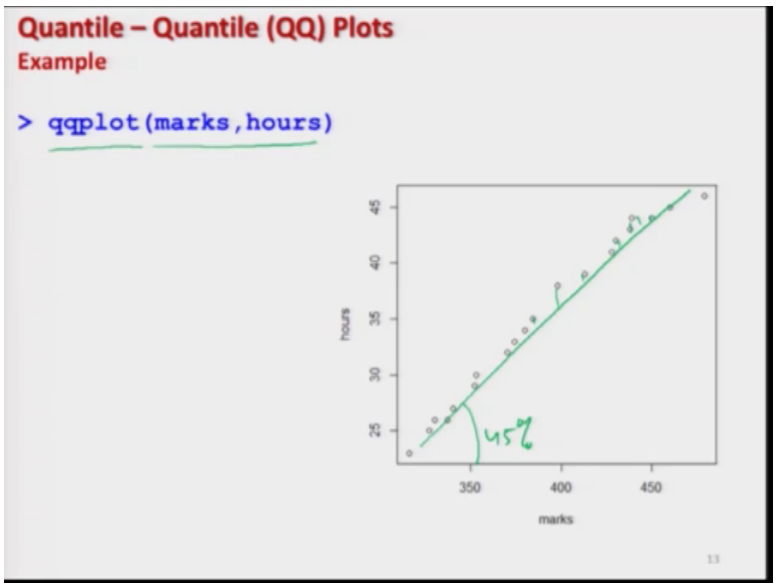


Now, in another example I would try to take the same data set which I had used earlier on two variables. So, in this data set 20 students have given their data on the marks obtained and the number of hours they have studied every week and this is here the first row is giving the marks and the second row is giving the number of hours they studied per week and this data is contained in the two data vectors here marks and hours.

(Refer Slide Time: 18:39)



(Refer Slide Time: 18:48)



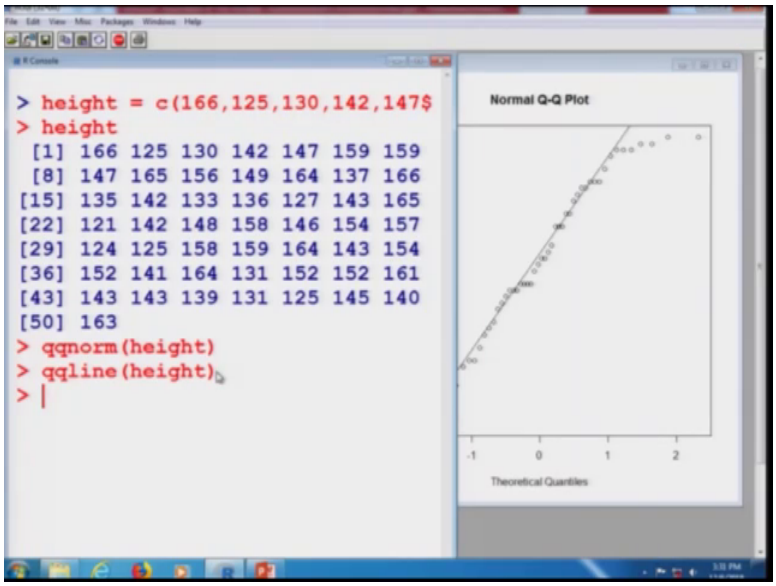
So, when I try to make a qqplot, so the command will be qqplot and inside the argument the two data vectors marks and hours. So, this qqplot is going to help us that I have got

here two data set; one is the marks and another is the number of hours. And this is suppose coming from say population number 1 and hours are coming from the population of hours called as population number 2.

So, I would try to see whether these two populations are same or not, this population 1 and population 2 are they have got a similar characteristic or they have got different characteristics. So, in this case also you can see here that there is a line which can pass through this thing and this angle is going to be 45 degree for this line. So, one can conclude that well most of the points are lying on the line or near to the line. So, I can say that the samples are coming from the similar population.

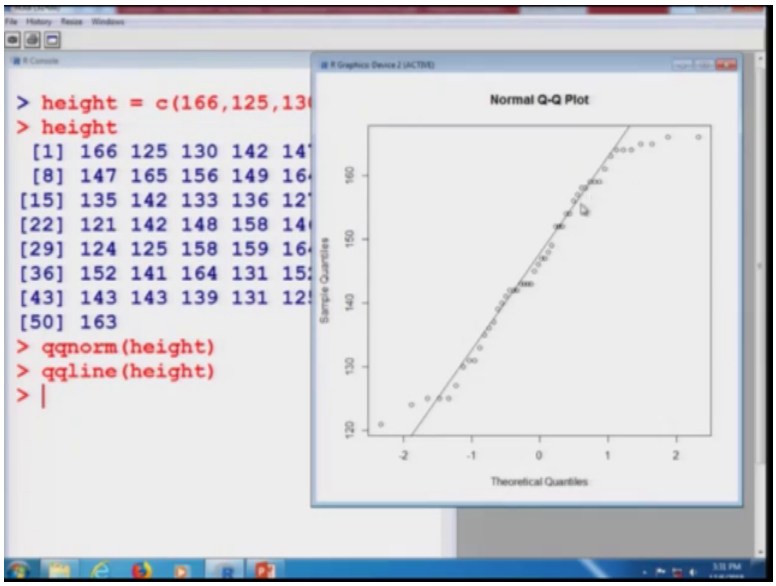
Here you can see here these are the deviation that we need to look and here I am trying to create this trend line and you can see here that how the points are going to lie and whether the trend is linear or something else. So, now looking at this data set we can have this idea. Now, before going further let me show you these operations on the R software.

(Refer Slide Time: 20:30)



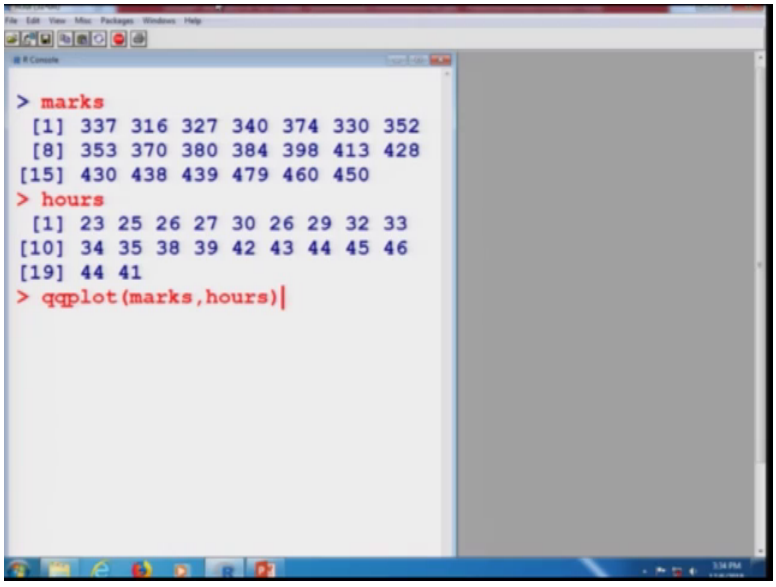
So, I am trying to first create the data vector here height, so you can see here this is the data which is contained here height and now I will try to plot the qqnorm of here height.

(Refer Slide Time: 20:44)



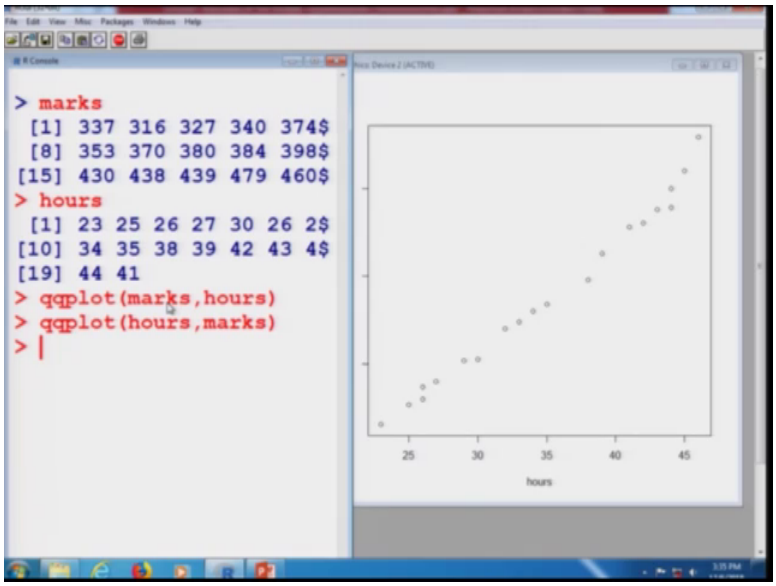
And as soon as I enter you can see here I am getting the same plot which I have shown you on the slides and if I try to make it here qqline means I would like to add a line trend line in this qqplot you can see here now we have this data and this is the qqplot and we have got here this line try to have a look on the cursor.

(Refer Slide Time: 21:11)



Next I would try to make a qqplot with the marks and hours so, I already have stored this data marks and hours is like this and if you try to make here a qqplot between marks and hours you will get here a qqplot like this one.

(Refer Slide Time: 21:31)

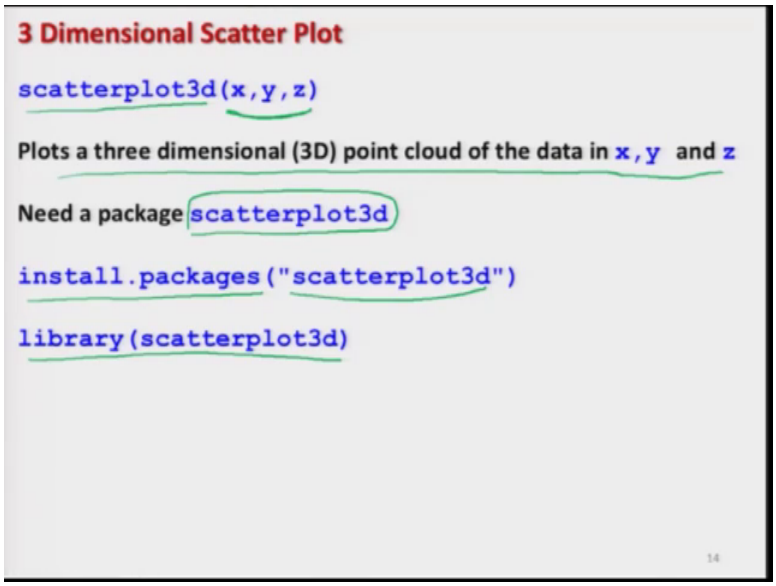


And also you please try to have a look and see if I try to change the order of the variables, say instead of qqplot marks hours I will say qqplot hours and marks you can see here this that correction will simply change, but the information is going to be the same what we are going to conclude. Now, let us come to next topic.

Now, I am going to discuss here briefly how to create the 3 dimensional plot. You see in R software we have a facility to create several types of 3 dimensional plots, well it is not possible for me to give the details of all the plot, but surely I will try to show you how to create those plots and how start it. And I will try to give you an example that how the different types of 3D graphics are made.

So, all the question you guess in what type of situation these 3 dimensional plots are useful. You see whenever we are trying to deal with multivariate data and we want to study the interdependence of the variables over each other, then in that case we would try to make such plots. For example, if I take an example which I have taken in my slides we know for children height, weight and age they all change with time as the age increases height also increases weight also increases, as the height increases the age and weight also increases and so on. So, now how to explore this type of interdependence, so for that we will try to create here the scatter plot.

(Refer Slide Time: 23:36)

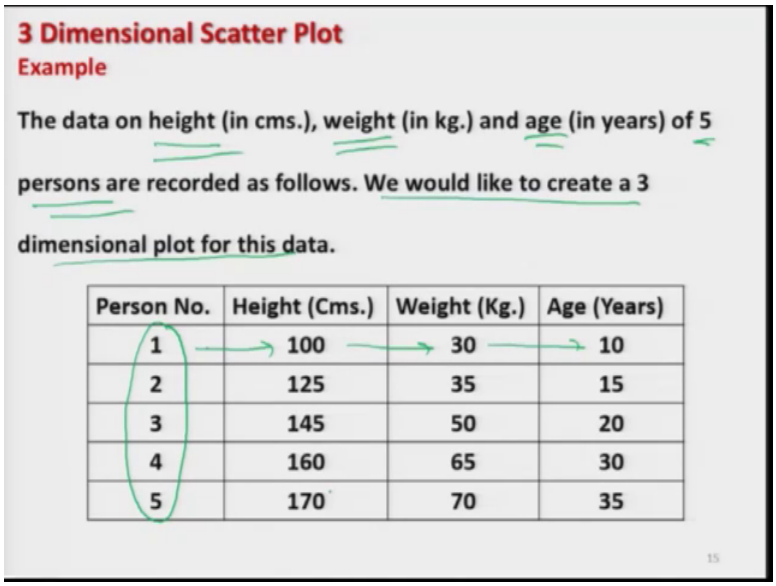


So, now I am going to take here some examples and I will try to show you the commands with those examples. So, first plot which I am going to consider here is the scatter plot that is a 3 dimensional scatter plot and this is created by the command is scatterplot3d s c a double t e r p l o t and 3 and d all in small letters and 3 in numbers.

And inside the arguments I have to give the data vectors for which I need to create this plot and this command will plot a three dimensional point cloud on the data x y and z, but for this thing I need a special package this is not included in the base package of r and the package which is needed here is a scatterplot3d.

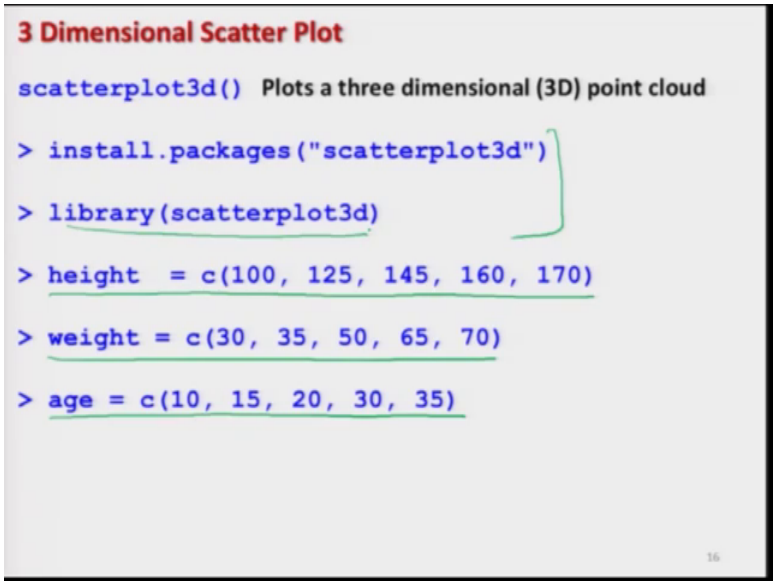
So, first you need to install the package using the command install dot packages and inside the arguments within double quotes type a scatterplot3d and after installing it you need to upload it by using the command library a scatterplot3d. This you know how to get it then otherwise you can simply use this command on the our console and can install it.

(Refer Slide Time: 24:52)



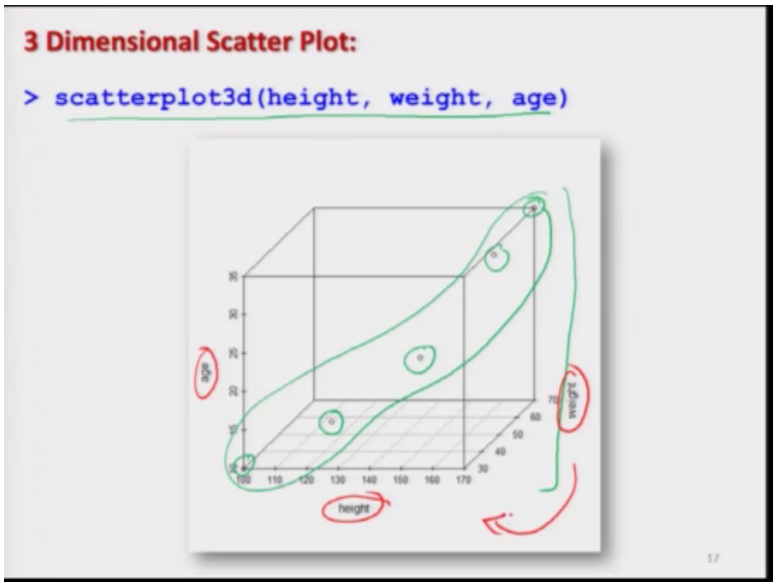
Now, I will try to take the example which I just discussed that we have taken the data on 5 persons for their height, weight and age and we would like to create a 3 dimensional plot for this data set. Well I am taking here only 5 data values this is because I want to show you that how the picture will look like and so that you can see inside the picture that how these values are coming. So, the person number 1 has the height 100 centimeters with 30 kilogram and age is 10 years and so on we have this data set.

(Refer Slide Time: 25:34)



Now, I have copied this data set in three vectors height, weight and age like this height, weight and here age. And before that I have install the package a scatterplot3d and I have loaded on the r console.

(Refer Slide Time: 25:52)

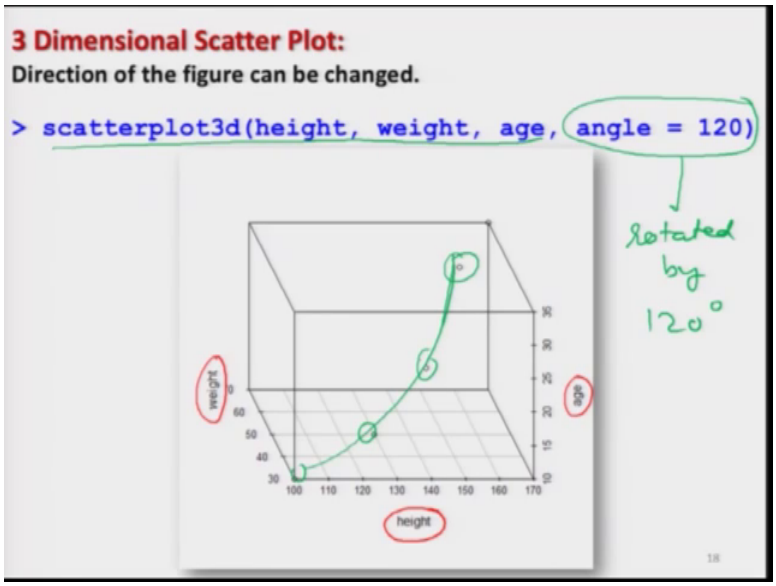


So, now in case if you use the command here a scatter plot3d height, weight, age inside the arguments you will get here this type of picture, you can see here these are the dots here which are indicating the values 1, 2, 3, 4 and here 5. And now you can see here this is a sort of cube or cuboid for which this graph is giving us an information that how the points are lying inside that cuboid right.

So, by looking at such graphs you can have an idea that how the things are happening. It is also possible to create the surfaces which are called surface plots and they will give you an idea that how the variation in the data is happening or how the data is behaving by looking at these observations.

Now, in this type of plot there are various options by which you can draw more meaningful inferences. For example, in case if I want to change that direction; direction means you can see here on one axis we have age, another axis is here height and say another axis here is weight.

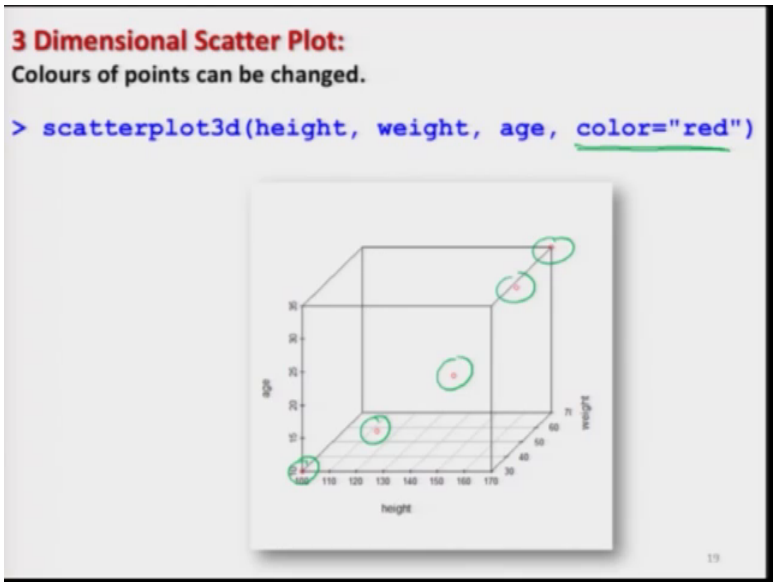
(Refer Slide Time: 27:22)



Now, I would try to change the that direction of this, so I try to now take here weight on the this side height on the x axis and age on the other side. So, then again if I try to use the same command here scatter plot3d with height, weight, age, but now I am giving here an option angle is equal to 120 degrees. So, by giving an option of angle I can control that how much the cube or the cuboid has to be turned or to be rotated.

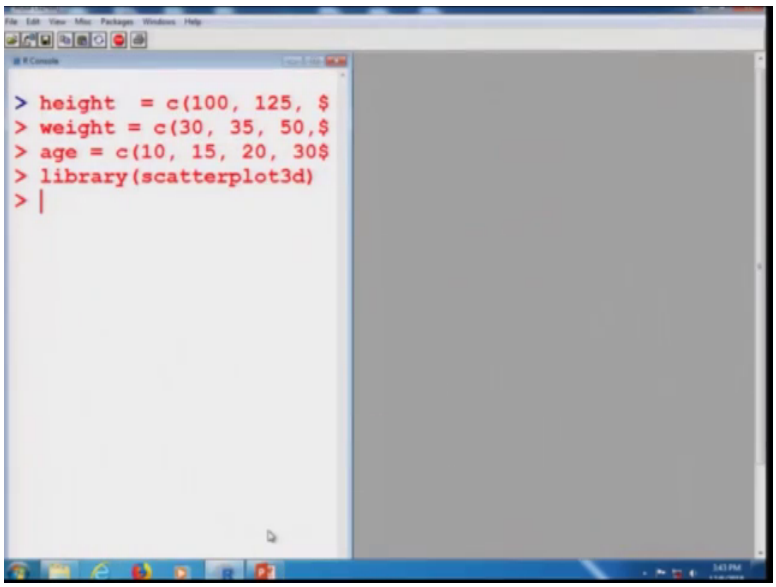
So, the earlier picture is now rotated by 120 degrees at an angle of 120 degrees and but now you can see here these are my points here, so you can see here in sorts of your curve one can see here. Whereas, in this case it was showing like as if there is a straight line. So, by making different types of cuboid with the changing angle you can have an idea that what is really happening.

(Refer Slide Time: 28:23)



And similarly in case if I want to change the color of the points, I can use here an option color is equal to. For example, here I have use red inside the double quotes and you can see here the color of these points is coming out to be now red and this command can also be combined with the angle and before I go further you help me to try to show you this thing on the r console.

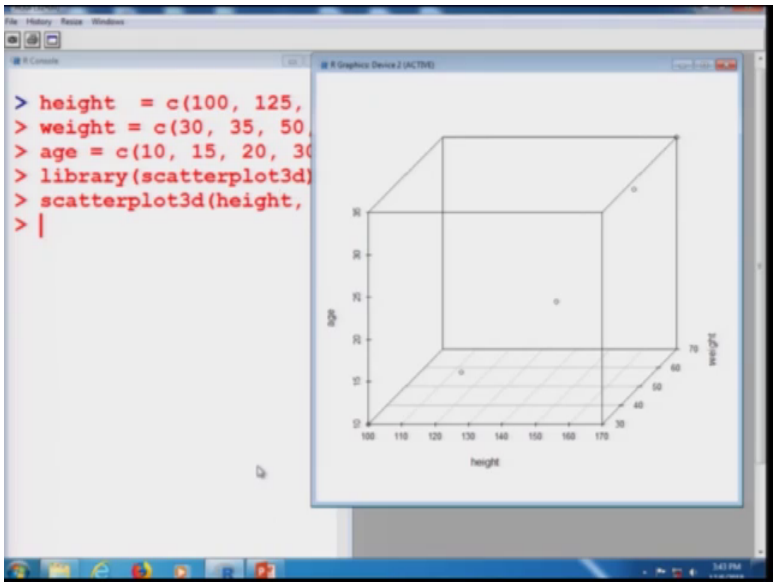
(Refer Slide Time: 28:58)



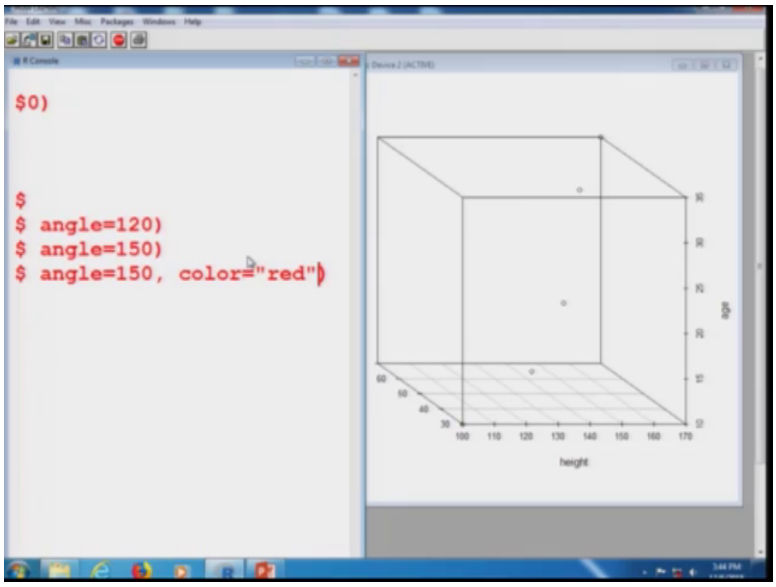
So, I will try to collect here the data, this is my height, this is my data on weight and this is the data on age. And now I need to first upload the library, I already have installed this

package on my computer, but you can do it yourself right. And now if I try to use this command to create the scatter plot of height weight and age you can see here I get this type of picture.

(Refer Slide Time: 29:35)

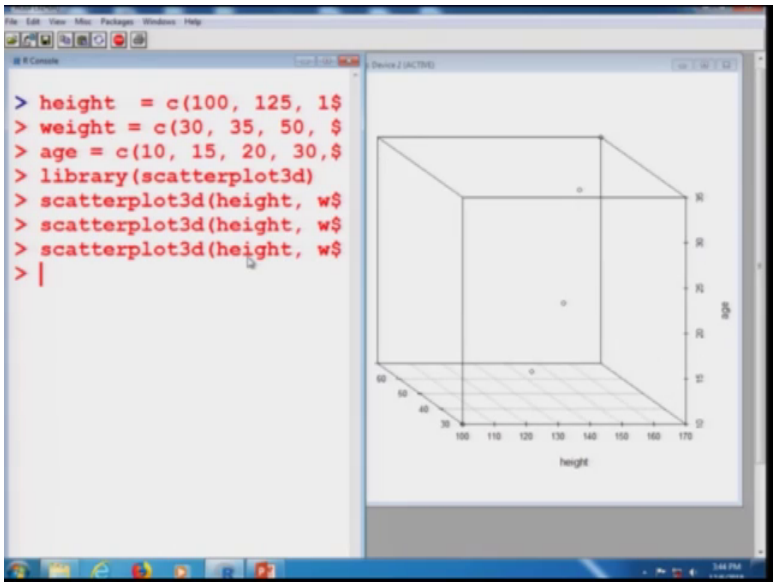


(Refer Slide Time: 29:41)



And if I try to add here say angle is equal to 120, this picture changes here well I can show you the both the things together. Suppose, now I try to make here an angle of say see here not 120, but say 150.

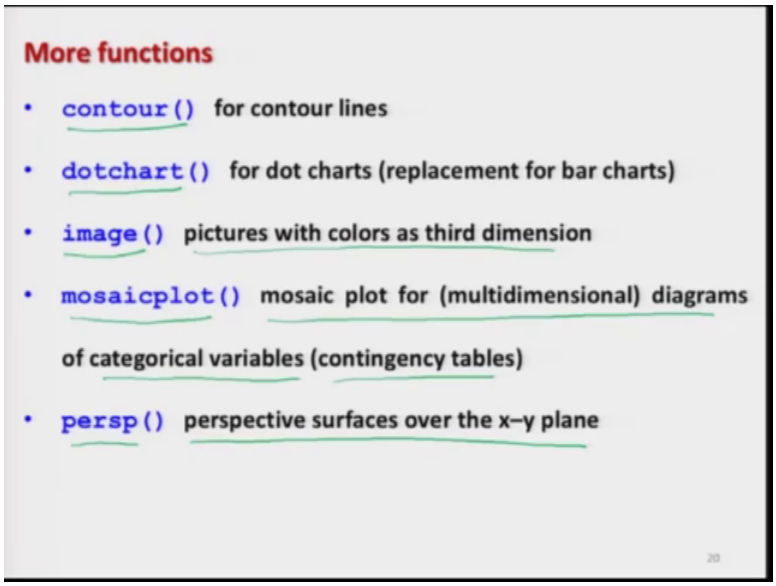
(Refer Slide Time: 30:07)



So, you can see here these points will change you can see here this direction is now changed. Similarly, if you want to add here the color say colors and angle can also be combined color is equal to say here red. So, you can see here this now gives me a red color and suppose if I try to make it here the colors to be blue, now the colors are blue.

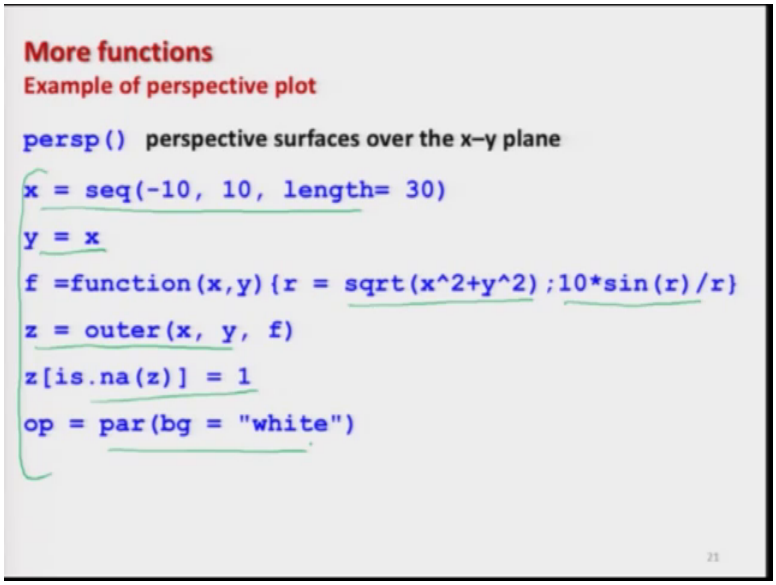
So, by looking at these types of pattern you can have more information. One good thing will be that you please write a program in which the angles are changing continuously say at an angle of 1 degree, 2 degree and so on. So, then you will have a picture which is continuously rotating and then you can have a 3 dimensional view which is possible in hour just by writing a small function right.

(Refer Slide Time: 31:11)



Similar to this 3 dimensional plot we have here some other types of graphics which I am not going to discuss here, but I am just informing you one is here see here contour plots which is give you the plot with the contour line, we have dot chart, we have image plots and this will produce a picture with colors as the third dimension we have a mosaicplot. And which is a mosaic plot for say multi dimensional diagrams a particular in case of categorical variable or say contingency tables that we are going to use later on. And there is another say here perspective plot which is obtained by either command persp and in this case you get surfaces over the xy plane.

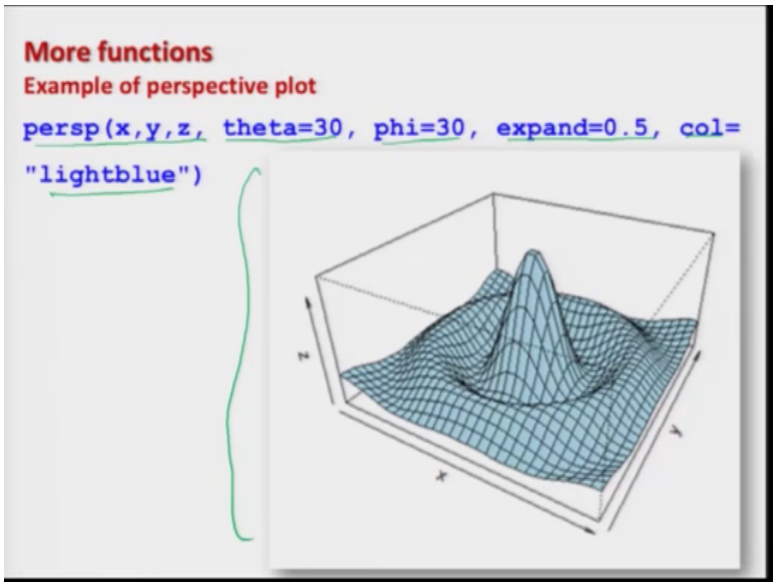
(Refer Slide Time: 31:57)



So, I will simply try to take an example here although I am not going to discuss it in detail, but I will show you that how the perspective plot is created and how it looks like and what is the advantage. For example, I have collected here some data, I have collected the data x as a sequence between minus 10 and 10 with the 30 observation and then y is going to the same as here x.

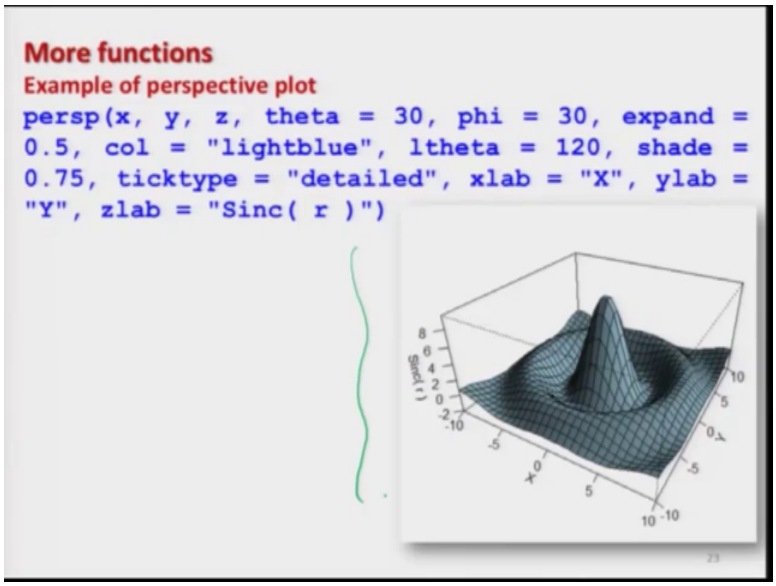
And then I have created a function to compute the value of r is equal to square root of x square plus y square or 10 into sin of r divided by r. And then I am trying to obtain here the z vector as an outer of x y and f and then I am trying to use here a logical operator and then I am trying to give other parameters.

(Refer Slide Time: 32:42)



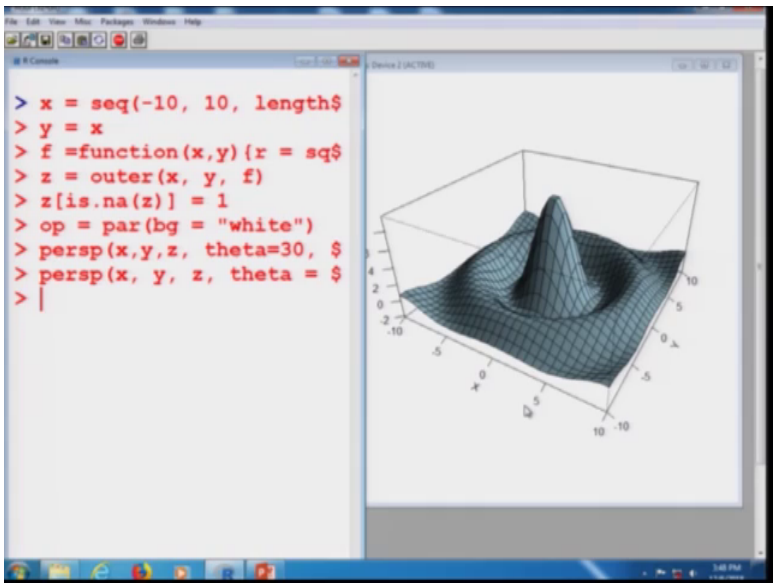
And then I try to create here the plot say perspective x y z with some other parameter theta equal to 35 equal to 30 expand equal to 0.5 and color is equal to light blue and this graph will look like this.

(Refer Slide Time: 32:55)



And similarly if you try to add here some more options over here try to inform the tick types shades etcetera; etcetera you can obtain here a different perspective plot. So, I would simply try to show you it on the r console. So, that you are confident that will these things are possible. So, I will simply try to copy all this commands at the same time.

(Refer Slide Time: 33:22)



And I will remove this thing and I will and you can see here that I have simply copied these commands over here and now I am going to plot here this curve, you can see here

when I try to execute this command over here I get this type of plot. And similarly if I try to use this command which I have done here, this is as soon as I execute this color changer there are shades and it is more informative.

Now, I would like to stop here with all the graphical tools for studying the association between two variables or more than two variables. Well in the given time frame I have taken some representative topics or some important types of plots, but this does not mean that these are the only plots. There are many that plots and you have seen that in case if you want to make your graphics more informative the simple rule is try to take the help from the R software about those syntaxes commands, try to see what type of information they can give you and try to use those commands inside the argument and try to control your graphic in the way you want.

And now you can see here with this one dimensional graphics two dimensional graphic multi dimensional graphics you can produce very good graphics which people try to create from say expensive software's, but the only thing is this that here you have to study little bit you have to understand. But the advantage is that you can control each and every parameter each and every characteristic of the graph whereas, in case of built in packages you do not have much options.

So, now in case if you spend some time try to learn more you will become successful in creating good graphics and we which will give you lots of hidden information contained inside the data. Now, in the next lecture I would try to develop some tools, so that we can get such information in a quantitative way. So, you practice where this graphics take some example try to create graphics, try to experiment with them, try to take different combination of the values of the parameter and see what you get and I will see you in the next lecture till then good bye.