Introduction to R Software
Prof. Shalabh
Department of Mathematics and Statistics
Indian Institute of Technology, Kanpur

Lecture - 38
Boxplots, Skewness and Kurtosis

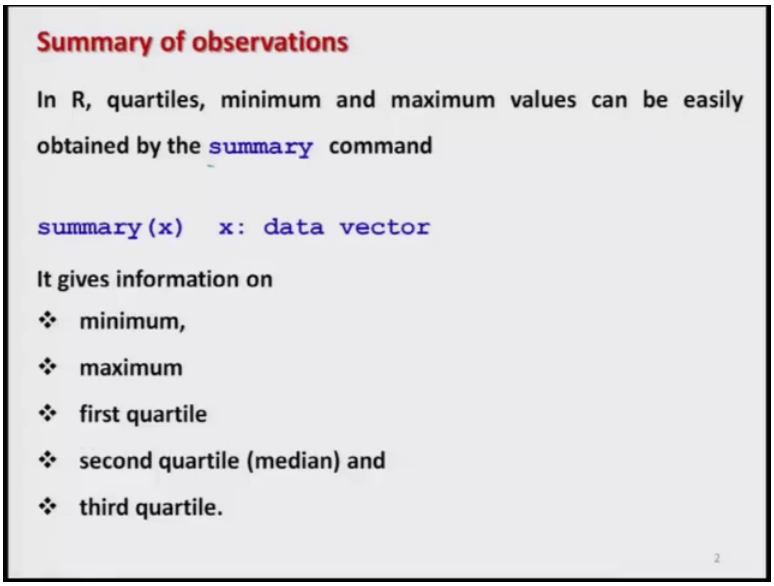
Welcome to the next lecture on the course introduction to R software. You may recall that in the earlier lectures we started a discussion on some statistical function. And we had discussed different types of statistical function which are available in the base package of R to compute frequency, measures of central tendency, measures of dispersion and so on.

So, we will try to continue with the same discussion and we would like to know something about box plots, skewness and kurtosis in this lecture. So, we try to start our lecture; you have seen earlier that when we talked about the descriptive statistics, then there are different types of things which can be computed. We had computed different types of functions like as minimum value of the observation, maximum value of the observation, different types of quantiles like a quartile decile percentile, and we have consider say some graphics also.

Now the question is that when we are trying to compare 2 different data sets, then it would not be a very good idea to compare them only on the basis of one characteristics. For example, it will not be a good idea to say since the mean of first the data set is higher than the mean of second data set, so, it is a good data set. Or in case if I individually consider that the variance of the first data set is greater than the variance of data set 2, then which one is preferable.

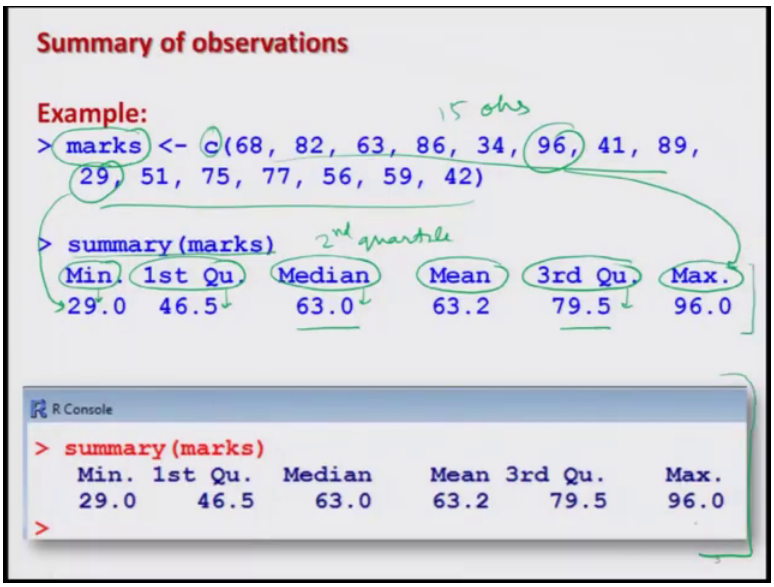
Actually in general all these feature should be considered together. In order to do that thing first option is that we try to compute all the things individually, like as minimum maximum different types of quantiles for each of the data set and then we try to compare them to make a final conclusion. In R there is a function what we call as summary function. And this function helps us and gives all these values from the same function. You may also recall that in the earlier lecture we also talked about the summary function and at that time I had told you that we will be considering it in the future lecture. So, this is the moment when we are going to consider the summary function.

(Refer Slide Time: 03:01)



So there is a command here, s u double m a r y. And this command gives us the information on the minimum maximum first quartile, second quartile which is actually median and third quartile from the same function and the syntax of this function is summary and inside the argument we have to write the data vector. So, first try to take example over here and try to see how it operates.

(Refer Slide Time: 03:33)

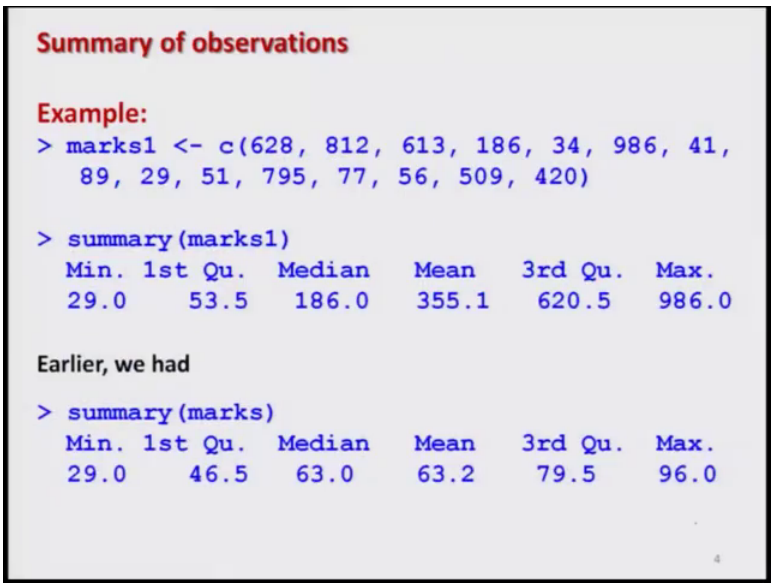


So, I simply try to consider here a data set of 15 observations. And these observations are on the marks of a students which are combined here inside a vector marks using the

command c. Now when I try to operate the summary function it gives me here summary of marks and you can see here this type of outcome comes. First value is the minimum value. So, it is saying that the minimum value is 29. That you can see from here this is this value. And then it is trying to compute the first quartile, second quartile which is actually the median.

And then it is trying to consider the third quartile also. And fourth quartile is; obviously, the entire data set. So, here you can see it is giving us first quartile is 46.5, second quartile that is the median is 63 third quartile is 79.5. And it also computes the mean, that is the arithmetic mean of the data set. And it also find out the maximum which is here 96 right. You can see over here. So, this summary function you can see here this is giving us all this information here in a single shot and this is the screenshot of the same thing. Now what is the advantage?

(Refer Slide Time: 05:07)



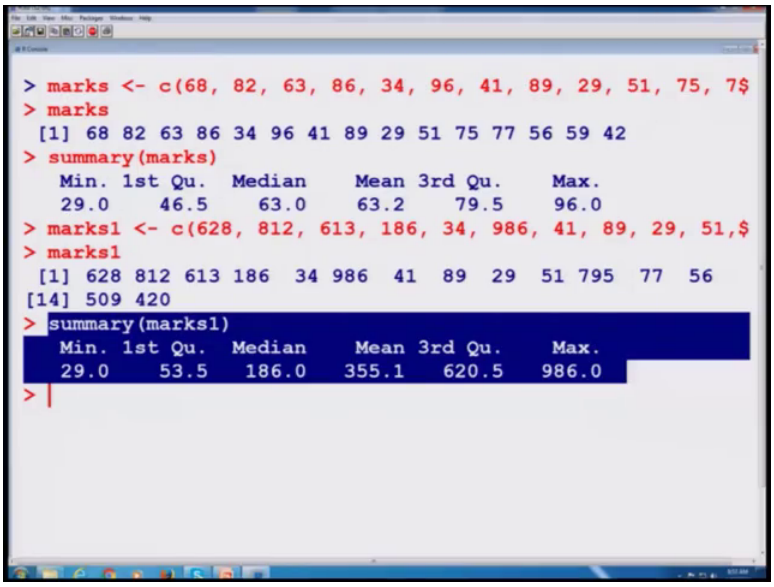
Suppose I try to take an another data set and I call it say here marks 1. What I have done in the same data set I have added some extreme observation, you can see here 6 28 8 1 2 and so on 795 986 and so on.

And I try to obtain the summary functions once again on the data vector marks 1. So, the summary function gives us this outcome. Minimum first quartile median mean third quartile and maximum. And in the last example we had the vector marks for which we had obtained the summary function like this one. Now I can compare the minimum of

the 2 data set. First quartile of the 2 data, set median of the 2 data set, mean of the 2 data set third quartiles of the 2 data set, and maximum of the 2 data set. And this will give us a sort of numerical comparison among all the values. For example, I can see here that 29 and 29 which are the minimum values in both the data set they remain the same, but first quartile is change median is change mean is change third quartile is change and maximum is also change.

Because you can see here is the maximum this thing. So, this summary function helps us in taking a conclusion by looking at these 6 values together right. And now let us try to do this thing on the R console also. So, first I try to write down my here my marks.

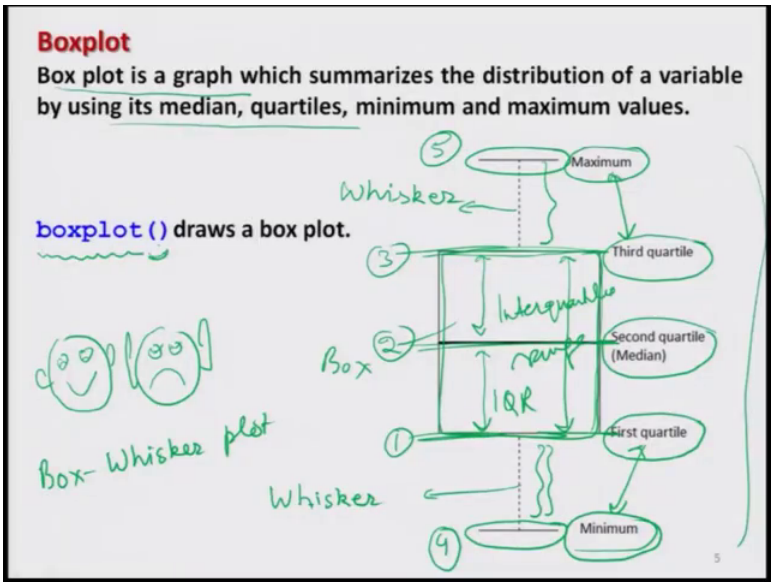
(Refer Slide Time: 06:58)



So, now let us try to do this on the R console. And first we try to paste the marks vector. So, you can see here is the marks and now when I try to find out the summary function summary of marks it gives me this thing something like this. And when I try to find out that summary for the marks 1, then I can see here that the marks 1 is like this and summary function of marks 1 gives us this thing.

So, you can see here this is the outcome of the summary function from the marks and this is from the outcome of the summary function from the marks 1. So, this is the same outcome which is given over here right.

(Refer Slide Time: 07:47)



Now, let us come back to our slide and try to consider another topic. We have seen that in the summary function, we are getting all the numerical values. The numerical values of mean median first quartile, second quartile, third quartile, minimum maximum and so on. And by comparing those numerical values we can have a fair idea that what is really happening in the 2 data sets.

And we can compare them, but do not you think that this will be a very convenient thing if all this information can be represented graphically. Why? What is the advantage of this graphic representation? Now you have seen different types of symbol. For example, say smiley for example, in case if I try to make here 2 faces like this one, and here I write simply like this and here I writes simply like this just by looking at the 2 faces, you can see that one face is representing the happy face and another face is not so happy.

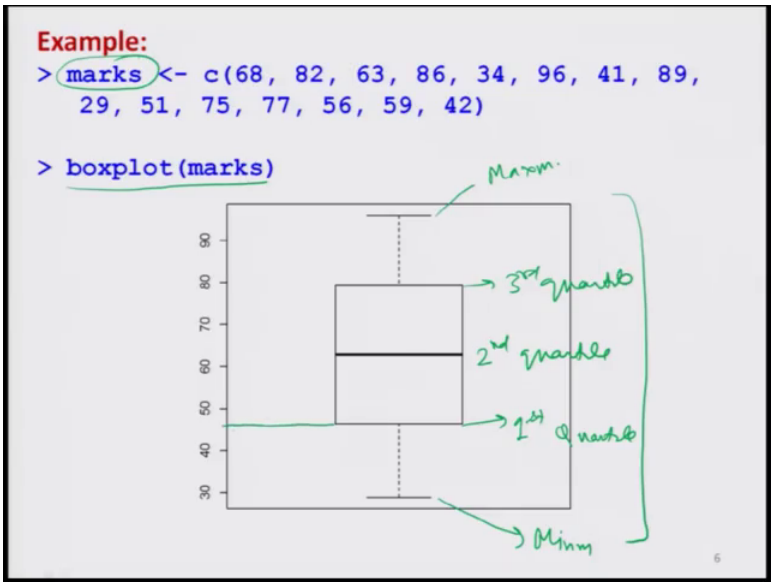
So, this is the advantage of graphical representation actually and here you can see means everything is there means you can also look at their eyes something like this or here if I try to make different types of eyes. So, you can look in the say mouth eyes and ear also here the ears are like this. So, you can compare everything just in a single shot. The same concept is translated to a graphic called as boxplot. Boxplot is a graph which summarizes all this information whatever we have got using the summary like as median quartiles minimum maximum mean everything in a same graphic.

And this looks like this over here, you can see here this is the boxplot right. Why this is called as a box? Because you can see that here is a box, this is here a box. So, in this picture this is called actually a box. And this is called a whisker. And this is also called as whisker, and that is why sometime this is also called as say box whisker plot. That is another name of the boxplot. Now we try to understand what this boxplot is trying to indicate? You can see here first we try to go inside the box. There is one line here one line here and one line here. So, let me call line number 1 line number 2 and line number 3. These 3 lines that is the first and third line at the sides of the box and line number 2 is in the box they are trying to indicate the quartiles.

For example, line number one is indicating the value of first quartile. Line number 2 that is inside the box that is indicating the second quartile, the value of the median. And line number 3 is indicating the third quartile that is the value of the third quartile from a data set. And similarly here 2 more values one is here minimum say line number 4 and say here another is here say line number 5 say here maximum. So, this line number 4 here this is denoting the minimum value of the data. And line number 5 is indicating the maximum value of the data. And this different that is the length of the whisker is indicating that how far is the minimum value from the first quartile, and the line number 5 here that is indicating how far is the maximum value from the third quartile.

You can also decide what is the difference between third and second quartile what is the difference between first and second quartile or even what is the difference between first and third quartile. So, if you try to see the difference between first and third quartile, this difference is I think, but your interquartile range. This gives you an idea of interquartile range that we had computed by the function IQR. So, you can see here that this boxplot is a graphic that is trying to give us various type of information under a same plot. And in order to draw a boxplot the command in R is simple b o x p l o t, boxplot and inside the argument you have to give the name of the data vector only. For example, let us try to take the same example that we did earlier that I try to take here a data vector here marks and we try to create the boxplot.

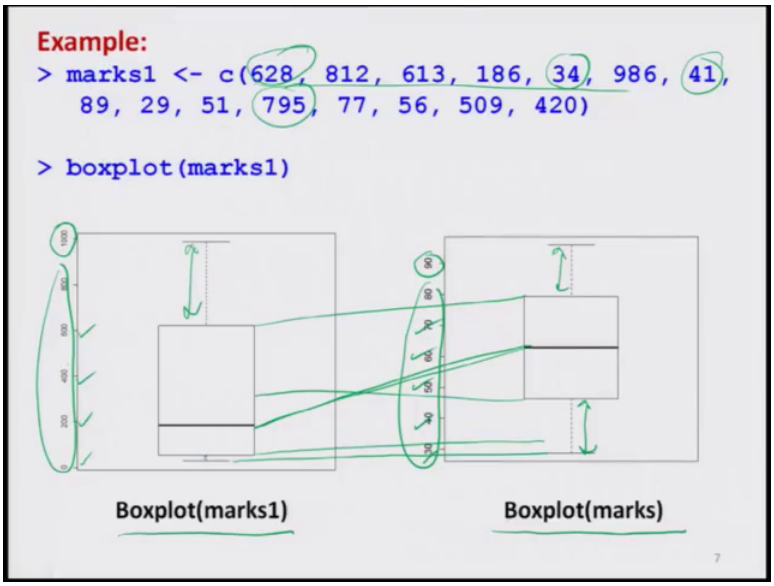
(Refer Slide Time: 13:29)



So, you can see here I have used my boxplots of say marks, and then it is trying to give us like this right. So, you can see here this is giving us the maximum value, this is giving us the minimum value and this value is the first quartile. This is my second quartile and this is my here third quartile. And you can read their values from here, you can see this is the value here, this is the value here, and this is the value here, this is the value here, this is the value here right, but reading the values is not that much interesting in the boxplot. It is more helpful when we try to compare 2 different data sets. So, what I try to do in the next slide is that, I try to consider the second data set that we considered marks 1. And I try to compare the box plots of marks 1 and marks. So, you can see here that in the means earlier vector.

Here marks there are no extreme values these all values are say relatively scattered around the mean, but in marks 1 these values are quite far away. For example, a lower value is something like 34 41 and the maximum value is or the upper values are something like say 628 795.

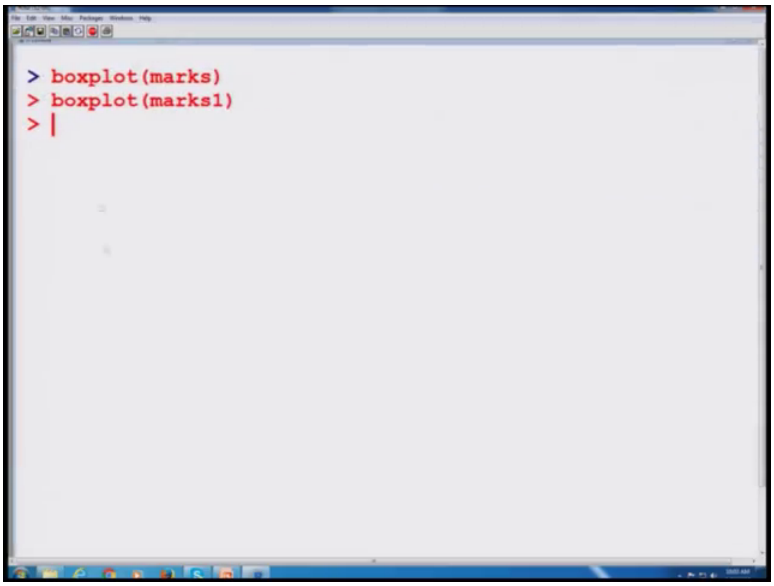
(Refer Slide Time: 15:30)



So, now I want to compare these 2 data sets. So, I try to first make the boxplot of the marks 1, which is given here. And then I also try to place the boxplot of the marks vector and now you can see I can compare. Here you can see this value here is 90 whereas, the here this value is 1000. And you can see here that the minimum value are remaining the nearly thus mean same, but this first quartile is here, it is here second quartile is something like this third quartile is something like this and this length of whiskers are also deviating. Here the length of whisker is just like this, like this.

But here you have to be careful that it is not only the length, but you also have to see that magnitude. The magnitude on this y axis they are also differing a lot. For example, in the box plots of marks the values are 30 40 50 60 70 and so on. Whereas, in the boxplot of marks number one, it is say 0 200 400 600 and so on right. So, by looking at these 2 box plots you can compare the 2 data sets, with their basic characteristic like as minimum maximum quartiles and so on. So, why not to do it on the say here R console and try to see what do we get well we already have entered the data marks and marks 1. So, I do not need to enter it here.

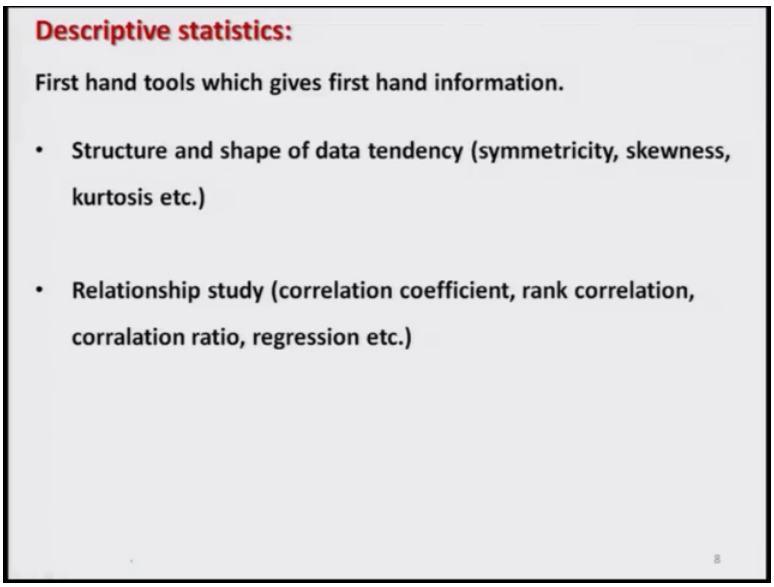
(Refer Slide Time: 16:52)



But I simply try to go with here boxplot of say here marks. So, you can see here this comes out to be like this.

And now I try to computed the created the boxplot of marks 1 you can see this will change. Try to observe on the boxplot and this is the same screenshot which I have given earlier. So, anyway. So, we come back to our slides.

(Refer Slide Time: 17:24)

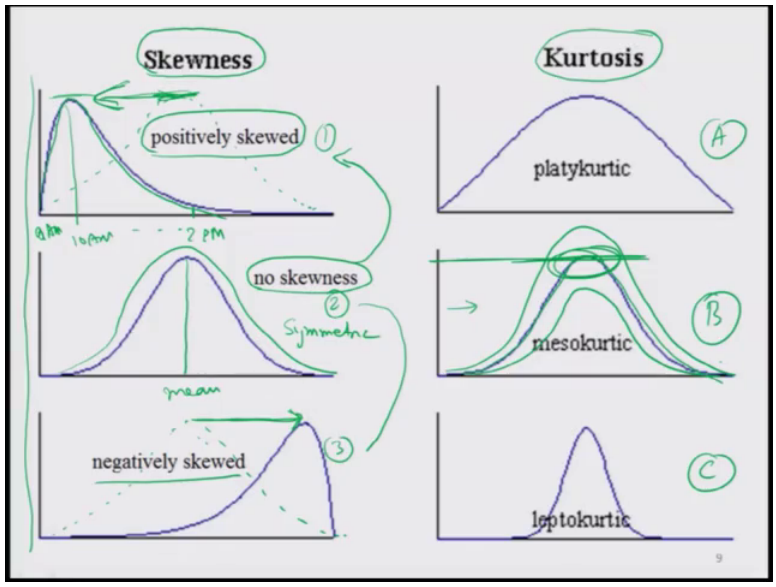


Now, we try to take another topic. You may recall that when we started a discussion on descriptive statistics. So, we have taken 4 topics measures of central tendency measures

of a variation. The third was the structure and of the data and fourth was the relationship study. So, we already have done the measures of central tendency and variation. So, now, we come to the third aspect. And where we want to study the structure and shape of the data and we want to see it is tendency. What do we mean by this? There can be various features inside the data.

So, here we are going to study on 2 aspect. What is symmetry that is measured by a skewness and another property is kurtosis. And after that we will be take up this topic also that how to study the relationships through correlation etcetera. So, now, we try to understand.

(Refer Slide Time: 18:27)



What do we mean by these 2 characteristic which are characterized by the terms skewness and kurtosis right? You can see here in the left panel; on this side we have here 3 types of curves. Let me call it here curve number here 1, curve number here 2, and curve number here 3. So, that you can follow me easily. First we try to see in the curve number 2. You can see here this curve starts from here and goes like this. And you can see here this is more or less symmetric. Symmetric around what this value, which is essentially the we can assume that this is the mean value.

Now, in case if you try to compare this figure number 2 with figure number 1, the figure number 2 if you try to plot over the same screen same figure this will look like this. So, you can see that this hump part that is now shifted on the left hand side. And similarly if

you try to compare the figure number 2 and 3, you can see here that in figure number 3 a symmetric curve would have been like this one. So, you can see here that here the mean value is shifted to here. And this hump is shifted from centre to the right hand side. So, this is essentially the feature which is characterized by a skewness, s k e w n e double s. Skewness is a measure of symmetry. Whenever we have a data set we would like to know whether all the values are symmetrically distributed or not. Symmetrically distributed around the say, for example, mean say half of the values are on the left hand side of the mean and half of the values are on the right side of the mean. Then I would say my data is nearly symmetric.

So, this is the character that we try to study here. And in this case say in the case number 2, we call it there is no skewness. And we say that the curve is symmetric. And on the same lines when we go to the figure number here one, where the hump is shifted on the left hand side this is here we call that this curve is positively skewed. And the opposite happens in the figure number 3 where the hump is shifted on the right hand side, and we see where that the curve is negatively skewed. So, this is the characteristics of skewness, the skewness can be say symmetric; that means, no skewness or it can be positively skewed or say negatively skewed.

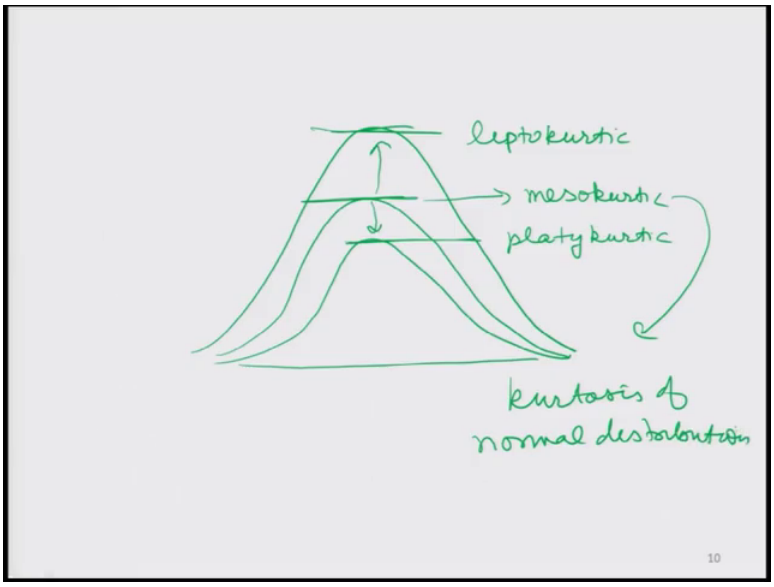
For example, if I take a simple example suppose I try to measure the number of vehicles crossing a particular say square or a particular point. Then you will see at about 10 o clock or say 9 o clock in the morning, which is a office hours many people are going to the office then maximum number of people are crossing that point say between say 9 and 10 a m, but about say 1 p m or 2 p m, most of the people are inside the office then the number of vehicles crossing that point will be very less. So, for example, if you try to look in the figure number one, if I say this is 8 a m and then it is something like here 10 a m and then this is something like here say at 2 p m this is my time. So, you can see here that the number of vehicles they increase from 8 a m till 10 a m, and after that they start decreasing. So, this type of data feature can be studied by a concept of skewness.

Similarly, we have another aspect that is called kurtosis. What is kurtosis? Now we try to look into these 3 graphics on the right hand side and let us call them as graphic number a graphic number b and graphic number here c. So, first we try to see into the graphic number here b. You can see here this is a symmetric curve, and it has a hump is here this is somewhere here. Now there can be 2 options. This hump can be more than this or this,

hump can be say smaller than this. This characteristic of looking at the hump of the curve that is studied through the concept of kurtosis right. So, now, I have to define 3 aspects one a standardized value of say kurtosis or the property of kurtosis. And then 2 other values or 2 are the names which are denoting that the hump is lower than the curve or upper than the curve.

So, this concept of defining whether the hump is lower or say upper that is actually based on the normal distribution. We are not doing it here in this course, but you can look into any statistics books normal distribution is a very popular distribution. And we try to compare the hump of say curves with respect to normal distribution. In order to understand this thing more clearly one have to be draw here a curve which is so, called the normal curve.

(Refer Slide Time: 24:51)



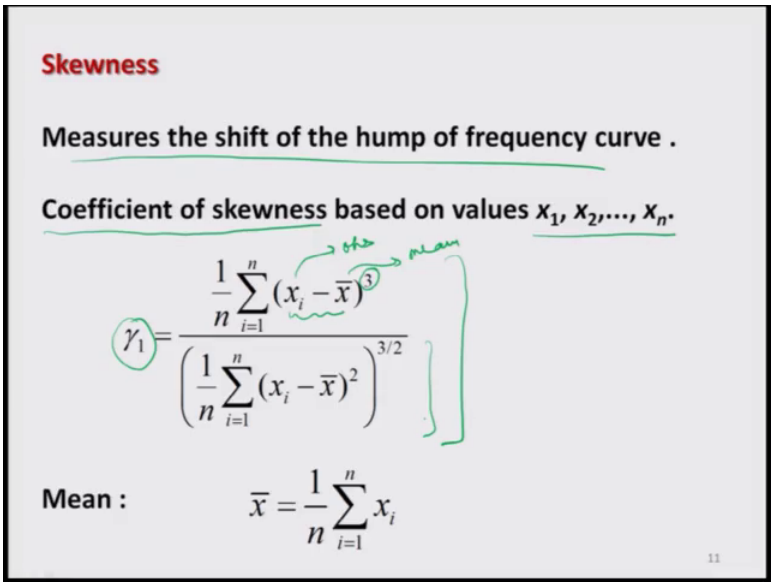
Now, there are 3 option. Let us try to look at this hump. One option is that this hump can be lower than this thing, this is in the lower direction. So, this is another characteristic. And another option is this this hump can be more than this somewhere here. So, this is bigger than this one.

So, this is called as mesokurtic. This is called as say here platykurtic. And this is called here as a leptokurtic. These are the 3 nomenclature to define the property of kurtosis. When we talk of this mesokurtic, mesokurtic is nothing this is the kurtosis of normal distribution right. So, this is how we try to classify the distribution of the data with

respect to hump. We simply try to plot them and we try to study the hump through the property of kurtosis. One will be called as platykurtic another will be called as mesokurtic and another is called as leptokurtic as we have discussed. Now the question is that by looking at the curve you cannot really decide much. We need to quantify it both the skewness and kurtosis. So, in order to quantify the departure from symmetry or the nature of hump we have coefficients.

And these are called as coefficient of skewness and coefficient of kurtosis. So, we try to understand how they are defined and based on that we will try to see how they can be computed in R software.

(Refer Slide Time: 27:01)

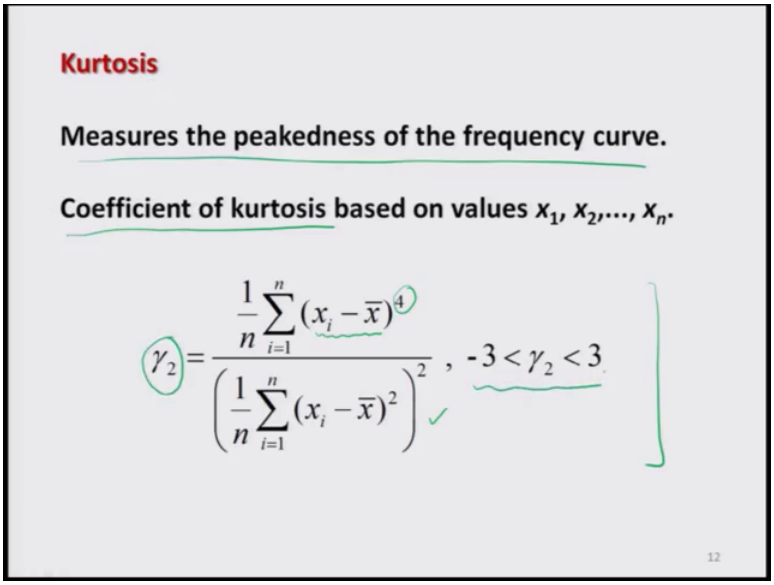


So, first we come to skewness this coefficient of skewness measures the shift of the hump of frequency curve either it is on the left hand side or on the right hand side and based on that suppose we have got some data values x 1 x 2 x n, then the coefficient of skewness is defined like this. And this is usually denoted by gamma one by this symbol. And here you can see it is a very simple thing this is simply trying to compute the sum of cubes of the deviation of the observation from their mean, these x is are my observation this is my mean.

And then it is trying to take the deviation of each observation from the mean and it is trying to make it is cube, and then try to take the average of these cubes. And similarly in the denominator also we have a similar thing. So, actually this quantity gives us an idea

that whether the hump is on the left hand side or right hand side. How? That we will try to see.

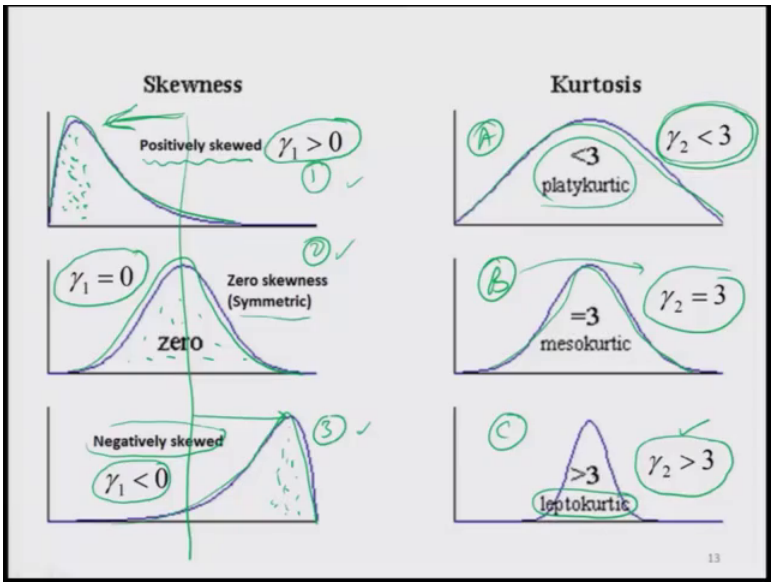
(Refer Slide Time: 28:03)



But before that similar to coefficient of a skewness we have the coefficient of kurtosis also. This coefficient of kurtosis actually measured the peakedness of the frequency curve right. And this is also a similarly similar measure, but here we are trying to take the deviation of every observation from the mean, and then we are trying to take it is fourth power and then try to take the average. And similarly in the denominator also we are defining a similar quantity and this measure is denoted as usually as a gamma 2.

And this value of gamma 2 lies between minus 3 and plus 3. Now what is the meaning of this coefficient of skewness and coefficient of kurtosis that is gamma one and gamma 2.

(Refer Slide Time: 28:54)



Let us try to see from this figure. So, in case if I have say symmetric curve, say let us call give them the same name as we have done it earlier you can see here I given the name 1 2 3 and a b c. So, we write the same nomenclature here 1 2 say here 3. And here a here b and here c. So, first we try to look into the left hand panel with the figures 1 2 and 3 first we try to look into figure number 2. Here the curve is symmetric. So, in that case in case if we try to compute the value of gamma one this will come out to be 0.

So, any value of gamma one goes to 0 will indicate that the distribution of the data is symmetric or say nearly symmetric. In case if gamma one comes out to be greater than 0 as in the figure number one then that would indicate that the hump is shifted on the left hand side. I can draw here this line this is the mean. So, this is shifted on the left hand side and in this case the gamma one will come out to be greater than 0. Any value of gamma one greater than 0 will indicate that the curve is positively skewed; that means, the hump is shifted on the left hand side from the mean value.

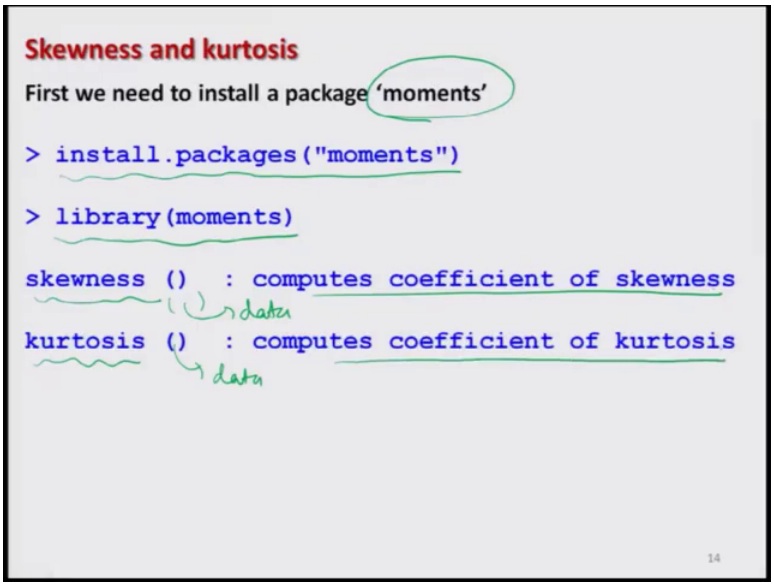
And similarly when we try to look into the figure number 3 in this case the gamma one value is coming out to be less than 0. And in this case we call that the curve is negatively skewed this means that when I am trying to compute the coefficient of skewness on the basis of given sample of data and if the value comes out to be negative, then I can conclude that the hump of the curve is shifted on the right hand side and it is something like this where more frequency is concentrated on the right hand side of the curve.

Whereas, in the figure number one more frequency is concentrated on the left hand side of the curve. Whereas, in the second case the observations are symmetrically distributed around the mean. So, the computed value of gamma one whether it is 0 greater than 0 or a smaller than 0 gives us an idea whether the curve is positively skewed symmetric or negatively skewed.

Similarly, now let us try to look into the figures a b and c. So, in order to know whether my data distribution is platykurtic mesokurtic or leptokurtic. We simply try to compute the value of gamma 2. Now I have 3 choices gamma 2 will lie between minus 3 and plus 3. So, one choice is this that gamma 2 can be equal to 3 as in figure number b. So, in case if I am getting gamma 2 equal to 3 that is indicating that my curve is mesokurtic. That is the hump of the curve or the peakedness of the curve is the same as the peakedness of a normal distribution. And when gamma 2 comes out to be smaller than 3 as in figure number a then we say that the frequency distribution is platykurtic like this. And similarly when the gamma 2 value comes out to be greater than 3 as in the figure number c then we say that my curve is leptokurtic. So, this is how we try to define the coefficient of skewness and kurtosis and this is how we use them.

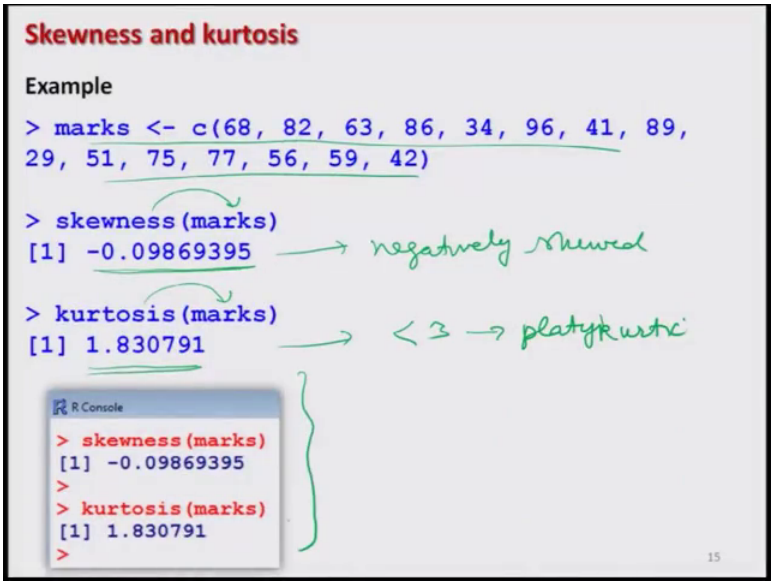
So, now the question is how to compute this coefficient of skewness and kurtosis in R software. So, in order to compute them in the R software, first we need to install a package. This is not built in the base package. So, in order to compute it, we try to consider a package moments, m o m e n t s and we try to first install it.

(Refer Slide Time: 32:56)



And then we try to upload it, by writing library moments and after this you simply have to write down a skewness and inside the arguments whatever is the data vector and then kurtosis k u r t o s i s and inside the arguments you need to write down the data or the data vector. So, skewness will compute the coefficient of a skewness that we discussed earlier. And similarly the kurtosis will also compute the coefficient of kurtosis that we discussed earlier right. So, now, I try to take this example that we did earlier.

(Refer Slide Time: 33:49)



So, I simply try to take the same data vector here, which I consider earlier and I try to simply write skewness of marks and kurtosis of here marks. You can see here that this value is coming out to be negative and whereas, this value which is smaller than 3 right. So, based on that what does this mean? This is saying that this is negatively skewed. And what about this? This is smaller than 3 smaller than 3 means what? Look here, this is platykurtic. So, I can say here that in this case this is platykurtic and here is the screenshot of the same thing.

So, now I would request you that you try this example yourself on your own computer. And also try to take the data set marks 1 and try to compute the coefficients of skewness and kurtosis yourself and try to see how do they differ you will clearly come to know that what is really happening with the skewness and kurtosis of the data. So, I stop here and in the next lecture, we will come up with some new topics till then you practice and we say goodbye.