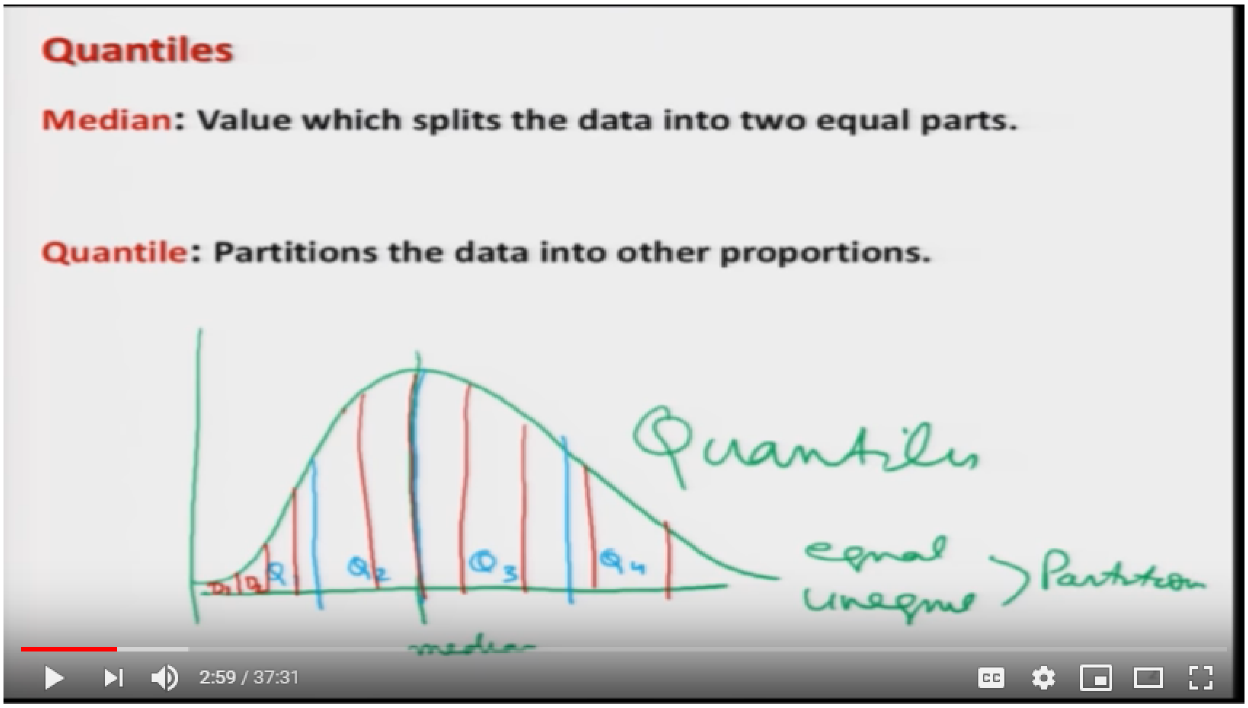
Lecture - 16
Quantiles

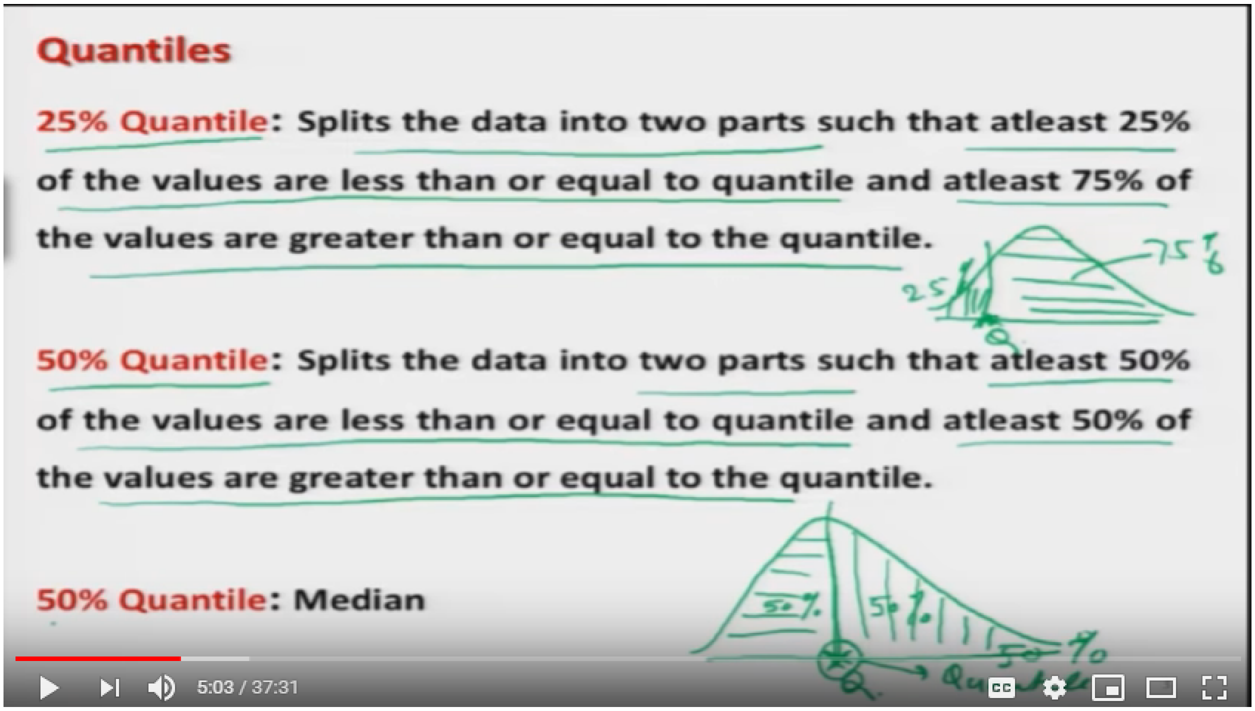
Welcome to the, next lecture on the course, Descriptive Statistics, with R Software, you may recall that in the earlier lecture. We started a discussion on the partitioning values. And we had discussed the concept of median and we had learned how to compute it manually. And then how to do it on the R software. Now I would like to extend this concept further, as we have seen that the median is the value, in a frequency distribution, which divides that total frequency into two equal parts. So, now the question is this, why only two equal parts, they can be more, they can before they can be ten they can be hundred, also this partitioning can be equal or this part is things, can also be unequal widths. So, now let us try to understand this concept, in this lecture. And which are generally called as,' Quantiles'. So, quantile are nothing, but they are the values, which try to divide the frequency distribution, into different partitions.

Refer slide time :(01:30)



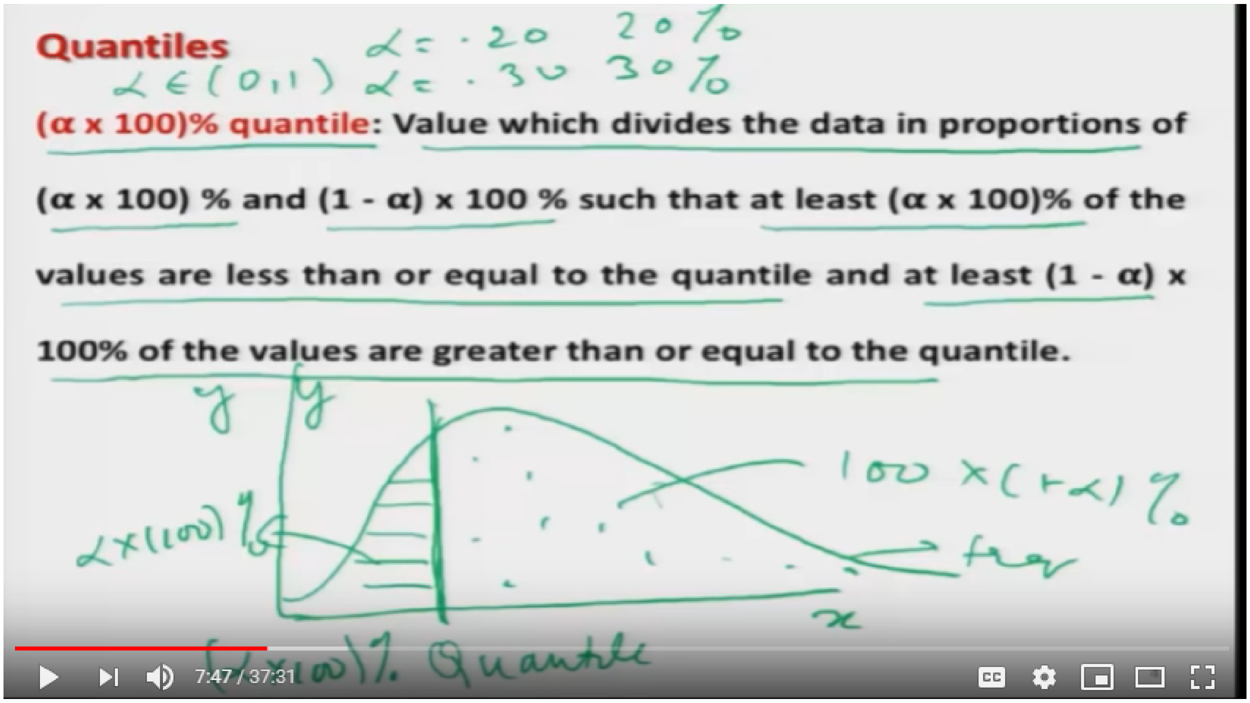
So, in case if I try to plot here, the frequency distribution, like this one then. We have understood that in case if I try to divide, it into two parts, then this is called,' Median'. And if I say, now I want to divide it into, four equal parts, for example, this can be first part, then second part, then third part and then here, fourth part. So, you can see here, this is a first part I can denote, it with here q1 this is the second partition I can denote it by here q2, third partition q3 and fourth partition q4. And similarly, I can divide it into say here, more equal parts like as ten equal parts, I convert it here 1, 2, 3, 4 and here 5, 6, 7 8, 9, 10. So, this will become here first part partition, second partition and so, on and similarly I can increase the number of partitions also. And these partitions can be of equal, widths or safe unequal. Right? So, in general these partitions are called as,' Quantiles'. So, just like, the median partition, the total frequency into two equal parts similarly, quantiles partitions, the total frequency into say, some other proportions and these proportions are decided by us.

Refer slide time :(01:30)



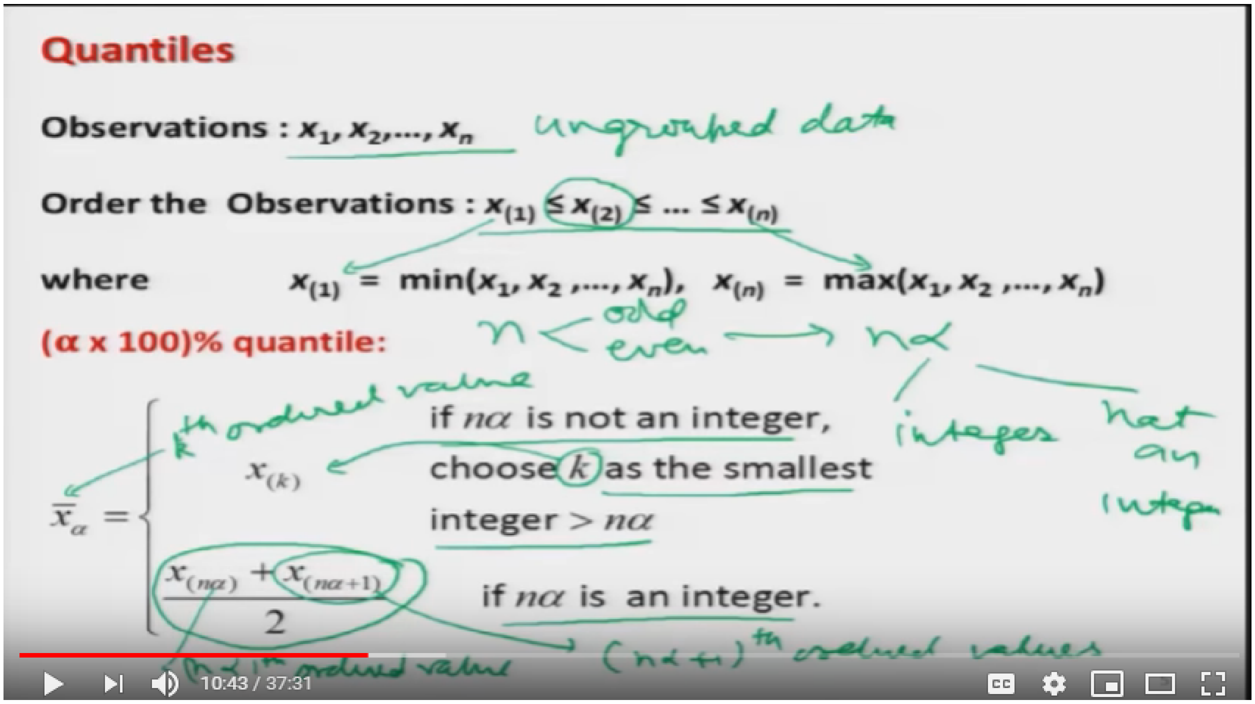
So, for example if I say 25 percent quantile. So, if you remember, the definition of median, the median was the value, which was trying to split the data into, two parts such that at least 50 percent of the frequencies are lower than that value and fifty percent values, are more than that value than, the value of the median. So, similarly if I try to say here 25 percent quantiles, than 25 percent quantized, they split the data into two parts, such that at least 25 percent, of the values are less than or equal to that quantile. And actually 75 percent of the values are greater than or equal to that quantile. So, if I try to plot it here it will look like, this that this is here, the quantile here. So, you had Q and this is the 25 percent part and this is here on the right hand side, this is the 75 percent part. So, this is here the point I'll value. And similarly if I try to define the 50 percent quantile, than 50 percent quintile split the data into two parts such that Ashley is 50% of the values are less than or equal to the quantile. And at least 50% of the values are greater than or equal to the quantile. So, once again if I try to plot here, the frequency distribution, then the value of the quantile is dividing, the total frequency into two parts, such that this part is the 50% of the total frequency and this part with vertical sign, this is also the 50% of the total frequency. And this value which is partitioning the total frequency, into two parts this is called here as a ‘, Quantile’. And this is essentially the 50 percent quantile and you may notice that 50 percent quantile is nothing but, your median. So, this is the,

Refer slide time :(05:08)



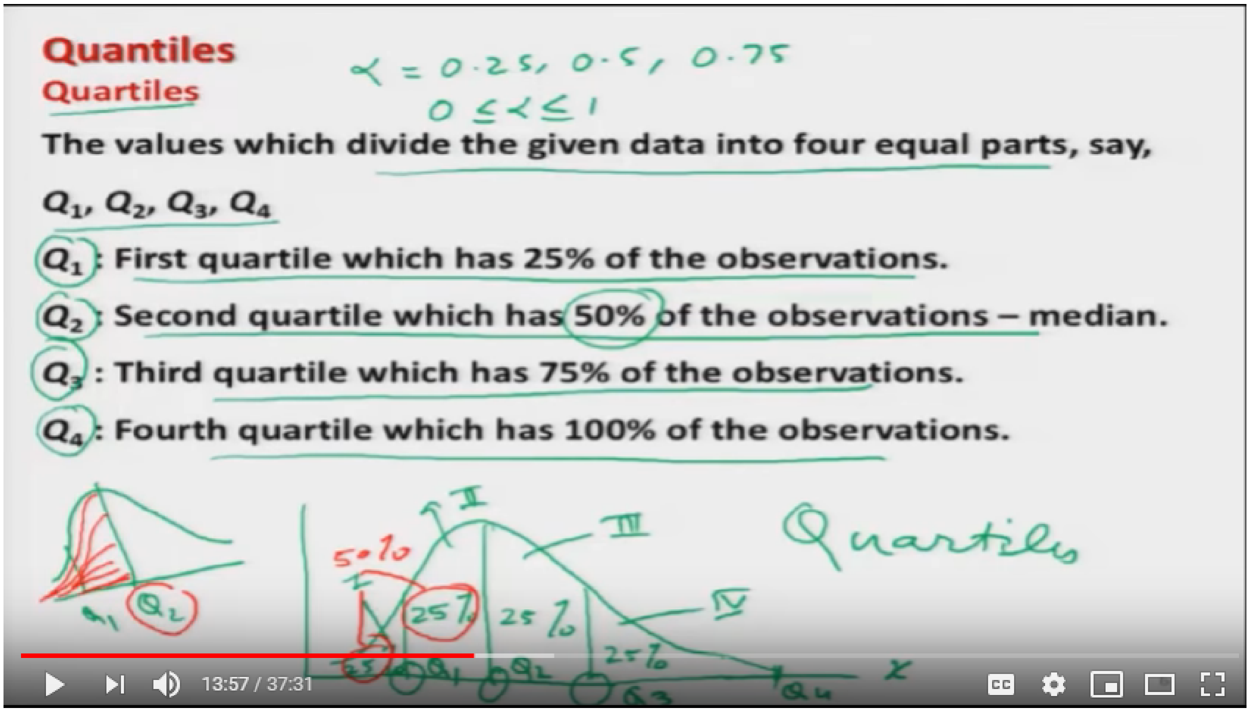
basic idea of the quantiles. So, now as we have defined that 25 percent quantile, 50 percent quantile similarly I can choose any value say 3 percent, 5 percent 11 percent, 18 percent to 25 percent, 25 percent 90 percent and so, on whatever we want? So, in general I can, extend this definition to a general definition of quantized. So, I can see here, when we choose here the value here, alpha. So, I can see here, the definition of alpha in 200 percent, quantile means if I try to choose the value of alpha, between 0 & 1. So, if I take the value of here alpha to be say here point to 0, then this becomes 20 percent quantile, if I try to take alpha to be 0.3 zero, then this becomes 30 percent quantile and so on. So, this alpha into 100 percent quantile is the value, which divides the data in two proportion, 1 consisting of alpha into 100 percent, another partition containing, the 1 minus alpha into 100 percent. And this division has been made in such a way, such that at least alpha into 100 percent of the values are smaller than or less than the value of that quantile and at least 100 into 1 minus alpha percent, of the values are greater than or equal to that quantile. So, in general if I try to plot it here. So, for example I can see here that this region, this region consists of alpha into 100 percent, of area this reason with the hair dots, on the right hand side, this is hundred into 1 minus alpha percent, area and these are the values of the, frequencies. So, this is a graph on the, x and y axis like this, sorry this exercises here, x axis and here y axis. And this value, which is doing it here, this will be you're here quantile or more specifically alpha, into hundred percent, quantile. So, now I can choose different values of alpha, the values of alpha can be a single value, the value of alpha can be in a sequence, where the values are at equal width or the values of alpha can be a sequence where the values are different, they are not of the equal width. So, firstly you can try to understand the basic definition, of this quantile and in order to understand this definition, means I can suggest you, please try to recall the discussion, in the earlier lecture, when we had discussed, the concept of ordered data and unordered data. And based on that we have, we had defined the median for ungrouped data.

Refer slide time :(08:16)



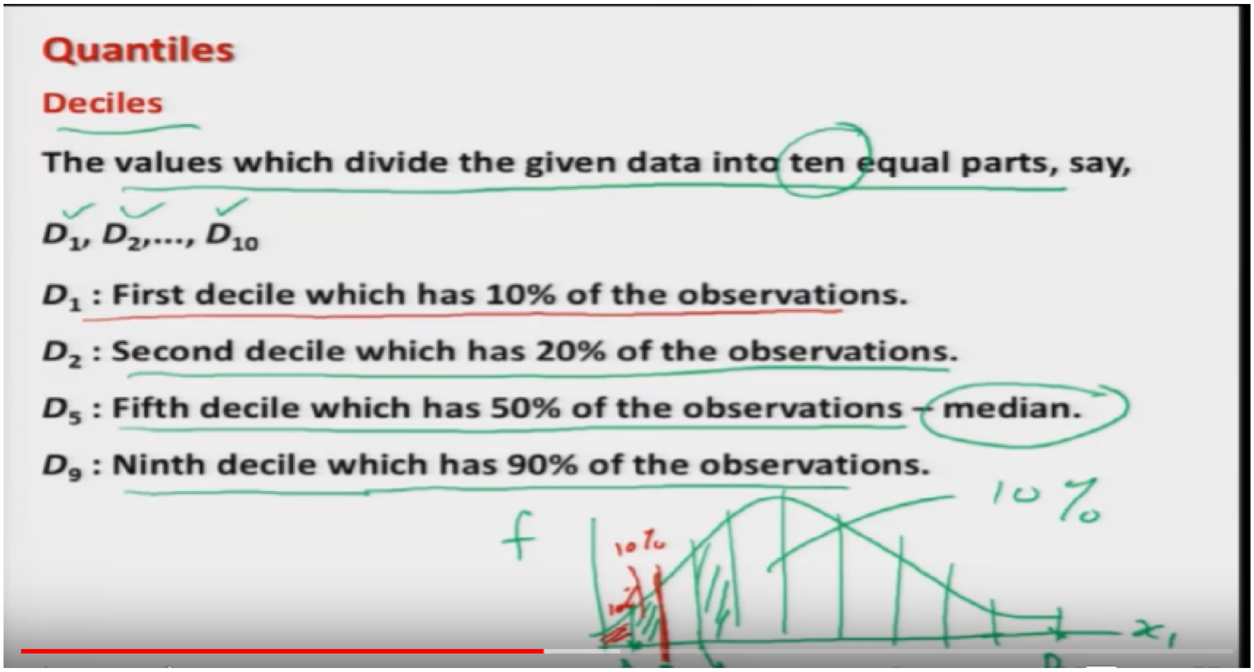
Now here, in this case also let me, consider, the data is given to be X 1 X 2 xn and yeah this is ungroup data. Right? And this data has been ordered. So, this value X 1 is K noting the minimum value among X 1 X 2 xn. And this xn is denoting the largest value or the maximum value, in the data and similarly here, this X 2 means the second ordered value that is converting, the second highest value. And now in case if you recall the, definition of median, for the two cases when the number of observations, were odd or even, Act definition is now extended to an integer and alpha. So, now the definition of the alpha quantile, is duct following, first try to decide, whether an alpha, this is an integer or not an integer. So, if an alpha is not an integer, then I have to choose the K which is the smallest integer greater than an alpha. And then corresponding to this value of K, try to find out the K, it ordered value. And this gate value or the K it ordered, value is going to provide the alpha into 100 percent, quantile similarly in case if, an alpha is an integer, then the quantile is going to be the arithmetic mean of two values, 1 and alpha 2 ordered value and second and alpha plus 1 it, ordered value. So, both these values, they can be obtained, from the data. And simply try to take the arithmetic mean of these two values and that will give you the value of the hundred into alpha, quantile. So, after understanding, this definition. Now I can see here that just by choosing the different value of alpha, I can divide the total frequency into different number of groups. So, for example, in case if I try to,

Refer slide time :(11:05)



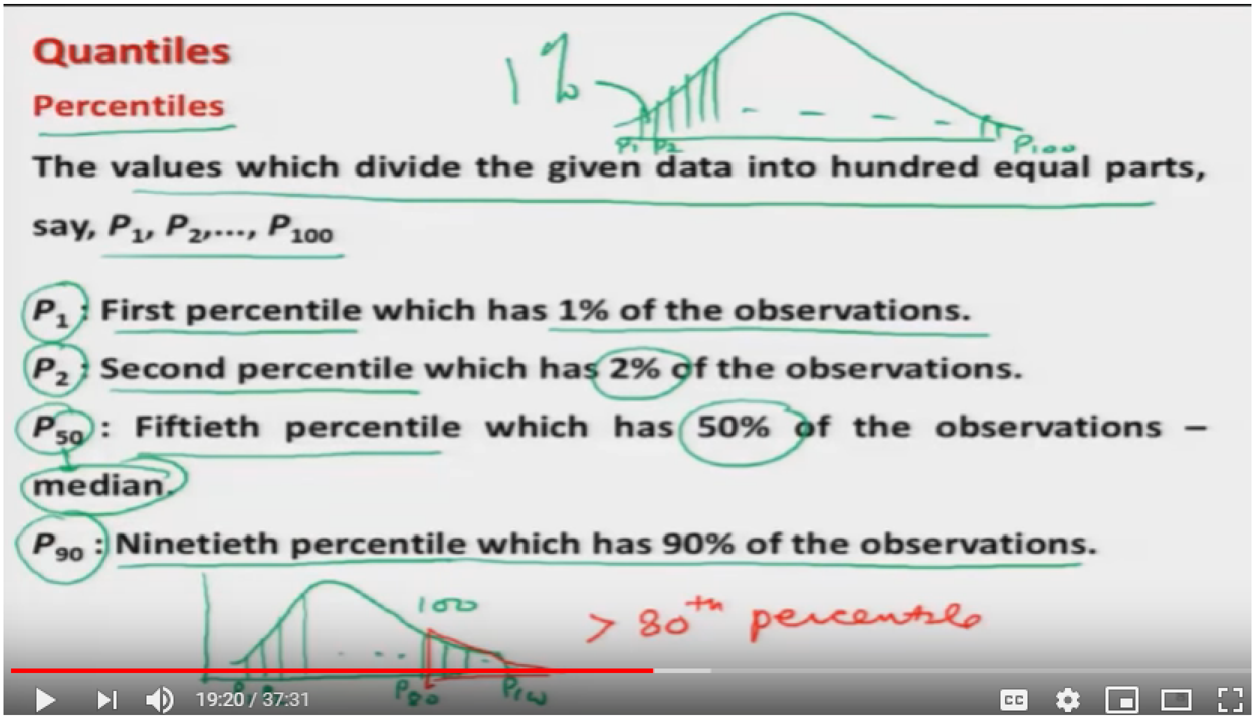
choose the Alpha to be here, 0.25, 0.5 0.75 and yeah means obviously the last value of alpha will be 1, because alpha is always lying between, 0 and 1. So, what essentially, I am trying to do if I try to create here the frequency, distribution like this. So, I'm trying to divide the entire frequency, into four equal parts here, alpha is equal to 0.25, corresponding to 25percent, the second section, another 25%, third section another 25% and another section 25%. So, these are the fourth partition, this is the first partition, this is the second partition, this is the third partition and this is the fourth partition. And corresponding to, which I can write here the values on the x-axis, say here this value to be q1, this value to be q2, this value to be q3 and whatever is here the fourth, in the final value this is here q4. So, these values are called as,' Quartiles'. So, these quartiles are the values or they are the particular values, of the quantiles, when the entire frequency distribution is divided into four equal parts and this is converted by q1, q2, q3 and q4. So, this q1 wait we can note the first quartile, which has 25% of the observations. And q2 is the second quartile, which has 50% of the observation, what does this mean? If you try to see here, my frequency distribution is like this. So, this is here first quartile and this is here second quartile. So, now it if you try to see, this q2 is going to be the value, which is trying to take care of entire this data. So, that is why, the two values here 25% and 25%, they are added and they are corresponding to 50 percent of the data. So, that is why I am calling it here that the second quartile has the 50% of the observation. And this is the same as the median. And similarly the third quartile is represented by q3 and which has 75% of the observation and then obviously that goes without saying that, the fourth quartile q4 will take care of all the hundred percent observation. So, now the rule is very simple, I am trying to divide the entire frequency, into four equal parts and those partitions will be called as,' Quantized'. And they are the particular value, of quantiles. Okay? Similarly in case if you want to divide the total partition into, ten equal parts for example here we have divided in four equal parts, now I am changing it to ten equal parts.

Refer slide time :(14:25)



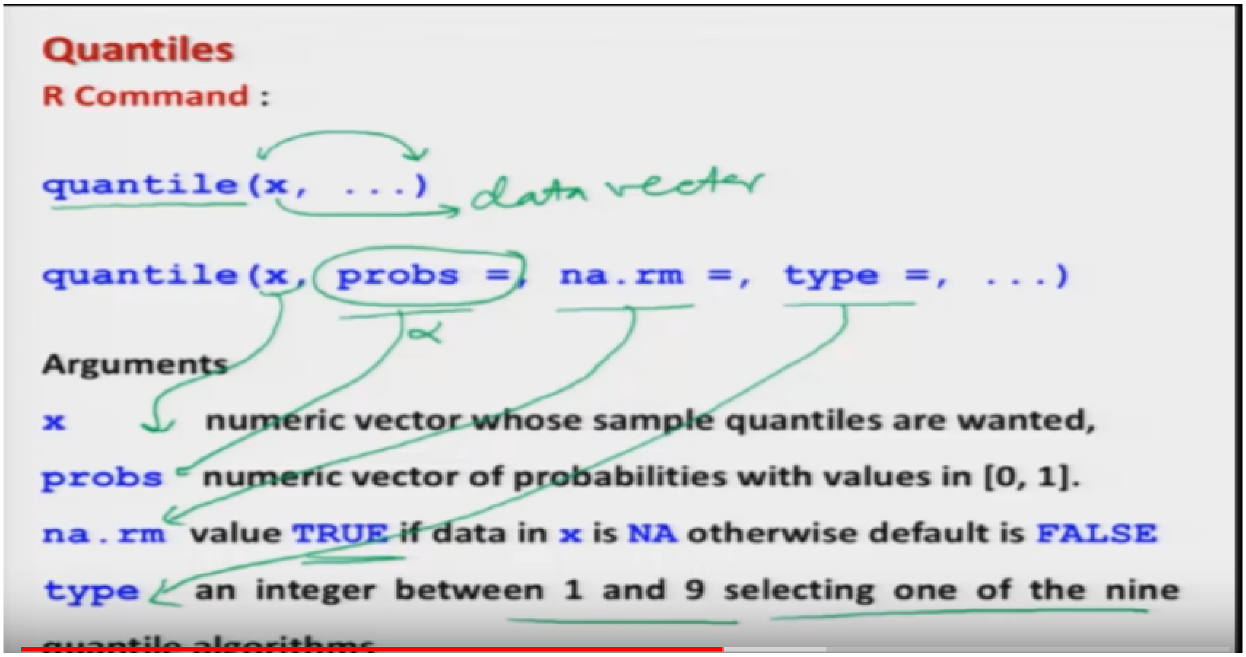
Then in this case this is called', Deciles'. So, deciles are the values we divide the entire frequency distribution into ten equal parts and we denote them as here D 1, D 2, D 10. So, this will look like this if I have this frequency distribution, then I am trying to divide it into say five, six, seven, eight, nine, ten. So, this partition value, on the x axis this is here say D 1, this is here D 2, this is here D 3 and so, on this is here D 10. And every partition here, this partition, this partition, every partition will take care of only, ten percent of the frequency. So, I can see here that this, first partition, this is trying to take care of only the 10% of the frequency. So, I am trying to decide, what I am trying to say define the D 1 as the first decile, which has only 10% of the observation, similarly we and I come to the second, quantile in this case this is here D 2, this is trying to take care of 10%, plus 10% frequency. So, I can see here that the second decile is the value of the quantile, which has 20% observation or 20% of the total frequency. And similarly if you go for third, fourth and so on, similarly third yes I will take care of 30% of the value, fourth will take care of the 40% of the value, fifth will take care of the 50% of the observations and that is going to be the same as median. And similarly the, ninth in castle will take care of the 90% of the observation. So, when we are trying to divide the total frequency, into ten equal parts, we call the quantile deciles. Now similarly in case if I try to, divide the total frequency into hundred equal parts, like that suppose I this is my here frequency,

Refer slide time :(16:30)



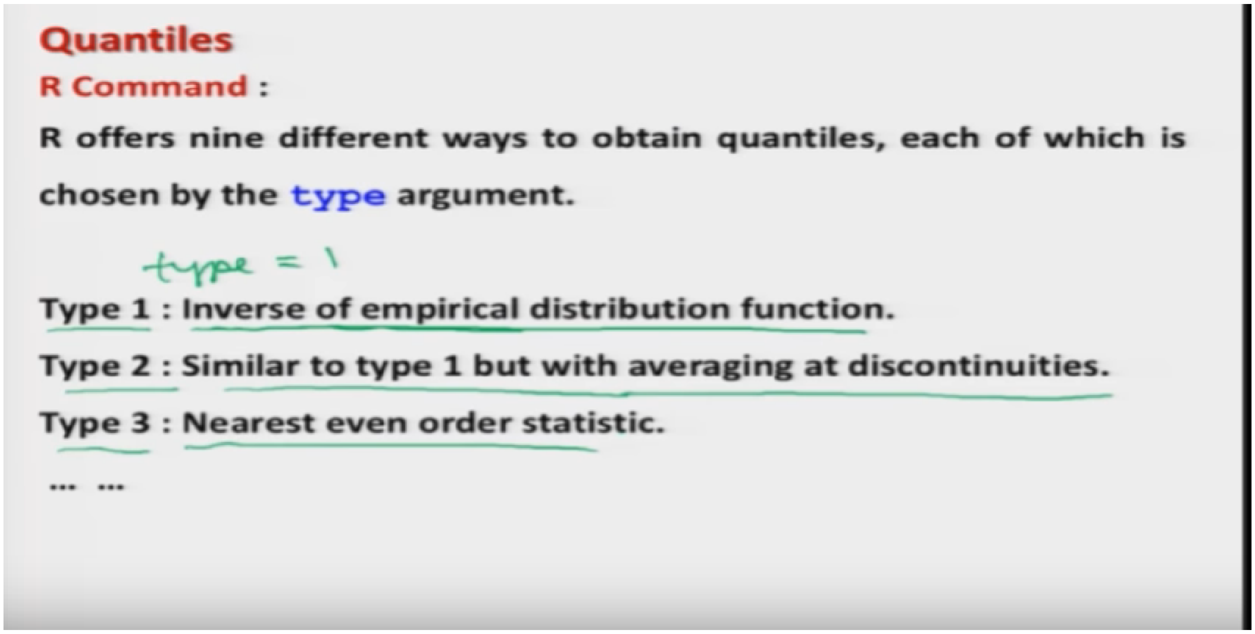
graph and I try to make it a one, two, three, four see here, handed partitions. So, hundred partition means, every partition will take care of one percent of the total frequency. So, here it will be here, they notice here p1 p2 and the last one will be here P hundred. So, the percentiles are the values of quantiles, which I divide the given data into 100 equal parts. And they are denoted as P 1, P 2 P hundred. So, this first percentile is denoted by, say p1 and this takes care of 1 percent of the total frequency or the 1 percent of the observation, similarly the second percentile, will take care of 2 percent of the observation. And that is denoted by here P 2 and similarly if I try to take the 50th percentile, this will take care of the 50 percent, of the observation and P 50 is the same as, median and similarly if you go for to say here 90^{th} percentile, then this will take care of the 90 percent of the total frequency and it is denoted by P 90, you have heard that some of the examinations, have a condition that the candidates, should have obtained the marks, which are lying in the top 20 percentile or top 30 percentile, what does this mean? If you try to see, what is top 20 percentile? The maximum value of the percentile can be hundred. So, top 20 percentile means, those who are lying in the top 20 percentile that is from 80th percentile 200 percentile. So, what they are asking, they are saying that suppose, a large number of students have appeared in the examination. And they have prepared the frequency, curve of the marks obtained. And now they have were divided, those frequencies into say here hundred equal parts. So, this first value is denoting the first percentile P 1 second is denoting P 2, somewhere here is P 80 that is the 80th percentile and finally it is here, P hundred that is the final percentile. What they are asking, is that any candidate? Who has scored the marks? Which are lying in this region, they are eligible to appear in the examination. So, that means, they have got the marks, which are greater than 80th percentile. So, this is the basic interpretation and this is what they want?

Refer slide time :(19:24)



Now after this, let me explain, how to compute this different types of contexts in the R software. So, the basic function is here quantile, QUANTILE and inside the arguments, we have to give different types of thing, but the compulsion is that I have to give here, a data vector. And I am denoting the data vector here say X. Now after this, we have several options, but I am going to illustrate here, the use of two props. And then na dot rm and I will also discuss about type, this first parameter is going to give you the data vector, of the data for which the quantiles are needed, PROBS, this is going to denote a vector, of probabilities, between 0 & 1. So, this is essentially the value of alpha. So, it depends, which of the quantiles you want to obtain. So, by controlling, the value of here this probs, we can generate different types of quantiles, like as quartiles deciles percentiles or something else. Third option is n a dot rm that we know that in case if there are missing values, in the data, then if I say that na dot R M is equal to true, then the quantiles, will be computed on the basis, of the data that is available inside the vector X, after this there is another parameter, here type, type is going to take a value between, 1 and 9 that can be 1, 2, 3, 4, 5, 6, 7, 8 or 9 and this is going to info rm the quantile, command that, which of the nine available, algorithms to compute the quantiles, is to be used, what does this mean? Well you see once, we have got that data, the data has to be ordered, using some algorithm and then based on that, the algorithm has to partition the values. And then the algorithm needs to choose the correct value of the quantile. So, this all these things, this entire process is based on, certain algorithms, different people have given, different types of algorithm, to compute the quantiles. And it is possible that when we try to compute the quantiles, based on different algorithms, their values may, differ a little bit. But that is essentially the choice of the experimenter or those who want to compute it that which of the algorithm they want to use. So, this R has this facility that you can choose, say any of the algorithm that is available, inside R software to compute the quantiles. Okay? So, for example.

Refer slide time :(22:59)

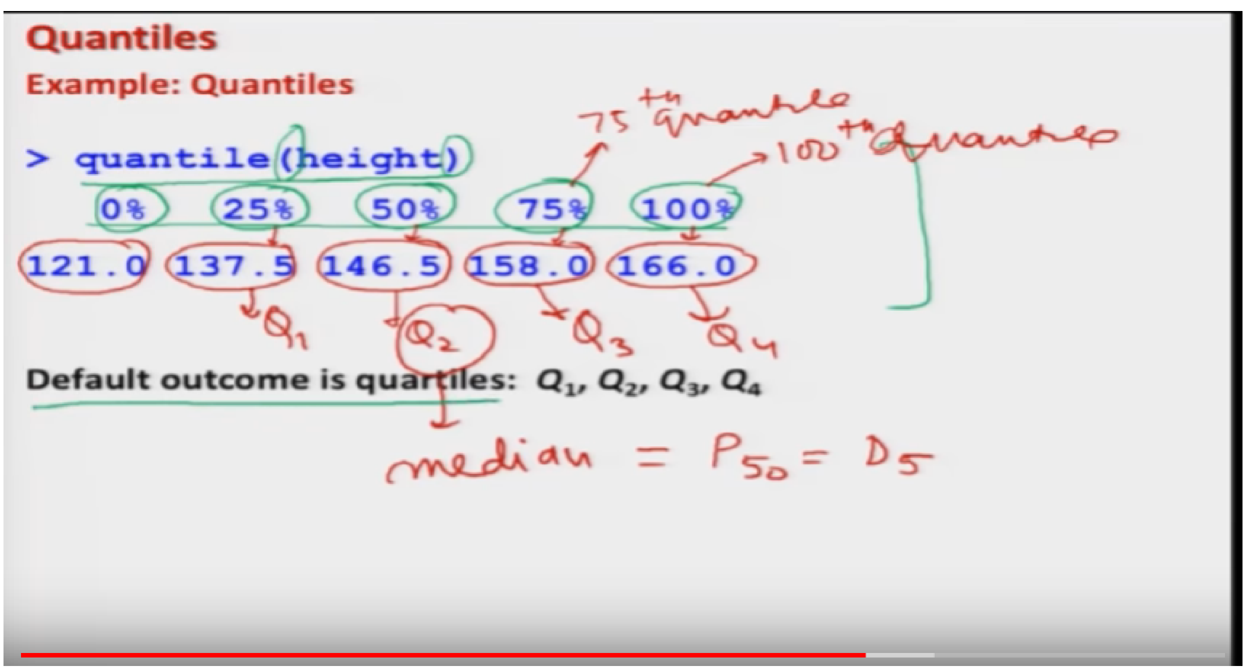


If I say, type 1 means it is something like, I will type, type is equal to 1, then this is going to use the, algorithm based on the concept of inverse of empirical, distribution function, if you choose type equal to 2, this is similar to type 1, but here in this case the averaging, has been done at the points, of discontinuities. And similarly if you choose type equal to 3, then this is based on the nearest even order statistics. But definitely I am NOT going to discuss, this type of algorithms here, but my simple objective is this how to use them on the given set of data. So, now I will try to take here, an example and I will try to show you that how you can compute, the quantiles or in particular quartiles, deciles percentiles or anything else.

Refer slide time :(23:55)

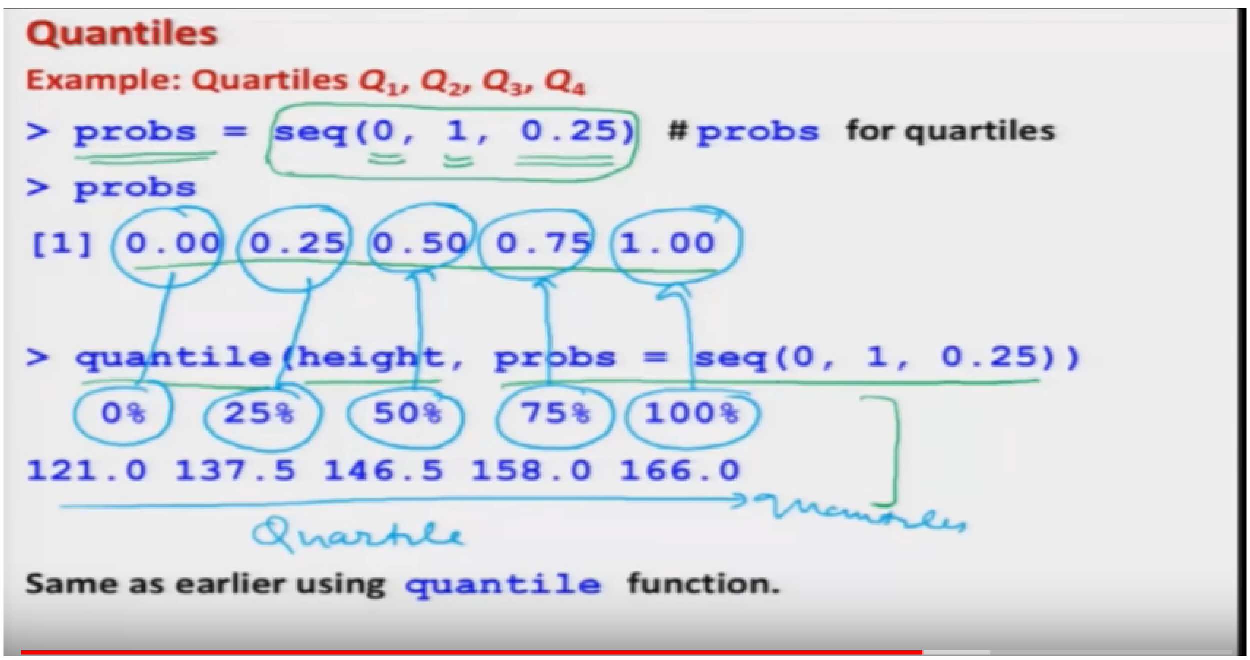
So, now I consider the example, which I also had considered earlier that there is a data on the heights of 50 persons, which are recorded in centimeters. And this data has been stored, inside a variable here, height. And now we would like to compute,

Refer slide time :(24:14)



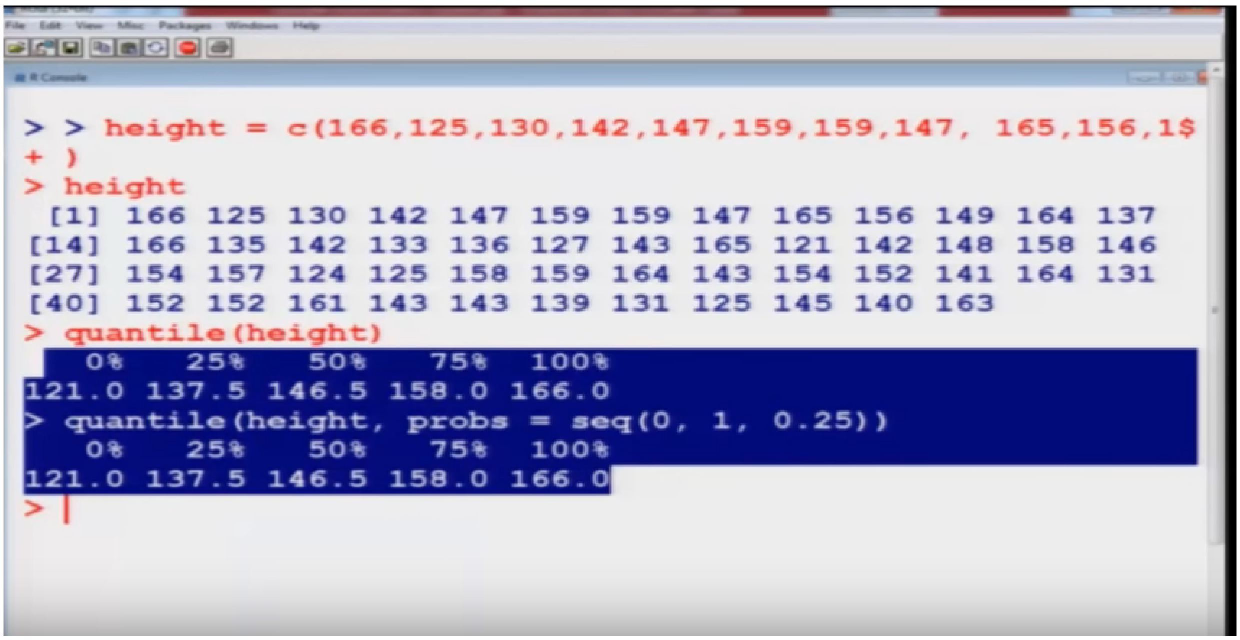
different types of coin ties over this data. So, when I say, quantile and here height, the height has to be the data vector, which has to be given inside the arguments, brackets. And this is here the outcome, you can see here, there are two rows first row, is this one. So, it is trying to show, 0 percent, 25 percent ,50 percent, 75 percent and percent and just below, this there are values here hundred 21.0, 137.5,146.5, 158.0 and 166.0. So, this is indicating, the value of the quartile at zero percent. And then the second value, is two converting the value at the twenty five percent, third value is denoting the value one four to six point five, at fifty percent quantile, this seventy five percent value is indicating the value one five eight point zero. So, this is indicating the seventy five or seventy fifth quantile. And this is indicating the last value, is indicating the hundred quantile, whose value is 166. So, essentially if you try to see here, this value is indicating, the first quartile q1, this is indicating the second quartile q2. And this is third value is indicating the third quartile q3 and q4 is the last value, which is the value of the fourth quartile. So, you can see here, this is also the value of the median. And this will also be same as the value of P 50 or say D five. Right? So, what you have to observe here that when you are trying to use the command quantile, then that default option here is quantiles. Now in case if you want to,

Refer slide time :(26:23)



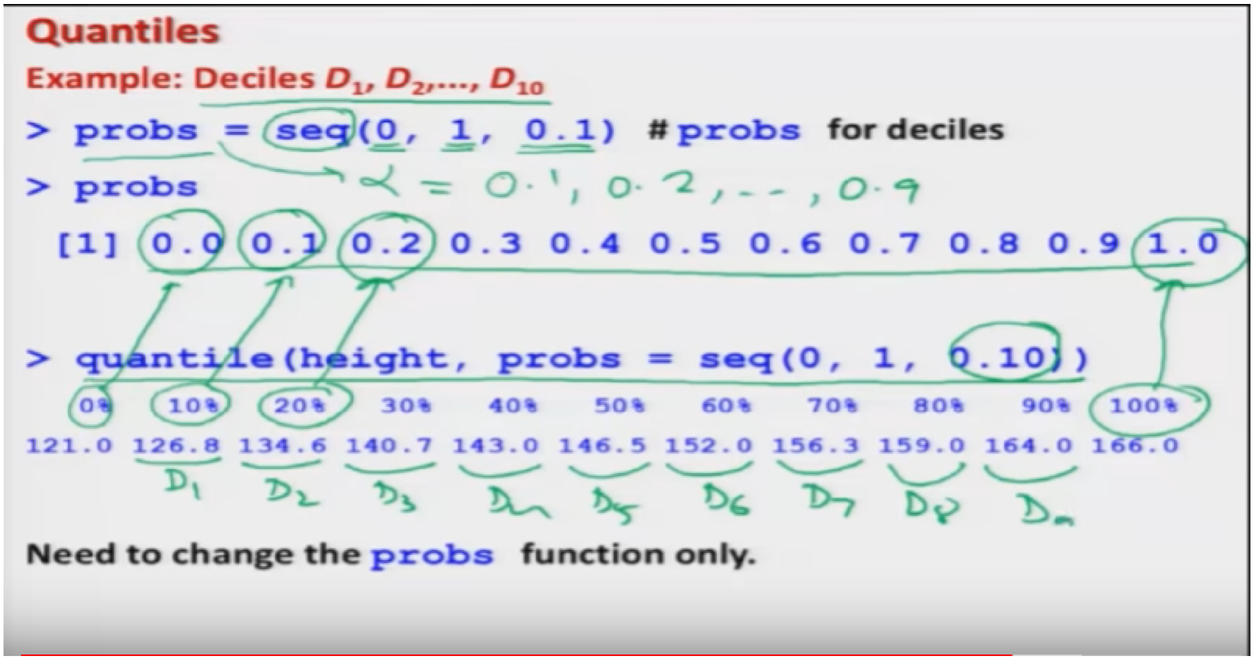
find out other things, like as deciles and percentile, then you simply have to control, the parameter probs. But here I would like to show you before I go to decide. So, percentile that in case if I try to choose, the probs here, to be a sequence between, our sequence from, starting from zero to one at an interval of 0.25. So, this command seq, it is going to create a data vector, like zero point zero, zero point two five point five points and five and one. So, now I'm asking to compute the quantile, of the data in the height vector using the probabilities, which are here. so, you can see here, the outcome will look like this and now what you have to understand that how the things, are happening this 0%, this is the same as this zero point zero, zero this 25%, is the same as 0.25, the second value in the probs vector, 50% is the same as zero point five, which is the third value in the probs vector, 75% is the same as the value points seven five, which is the fourth value in the probs vector and 100% is the last value one point zero, zero in the probs vector. And you can see here that these are the values of the, quantiles and essentially if you try to see, they are the quartiles. So, you can compare these values and the values, which you have obtained directly, by using the quantile, they are the same.

Refer slide time :(28:01)



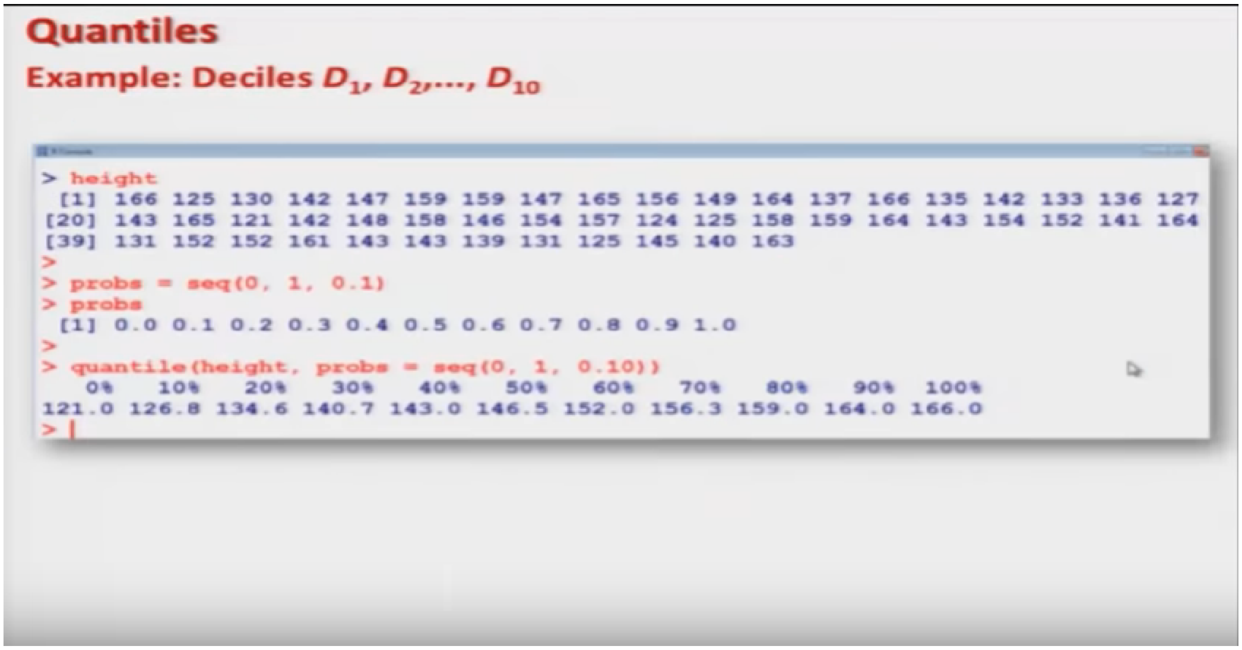
And this is here the screen shot of, which I will try to show you on the R console. So. let me try to first copy, the data on the R console. So, you can see here, this is the height, data and I'm simply trying to fight here quantile, off here, height. Right? And this is giving me the this, value and if I try to use here mice, this probs or the sequence of between zero to one, at an interval of 0.25, you can see here that these values are the same values, which are here. Right? Okay?

Refer slide time :(28:45)



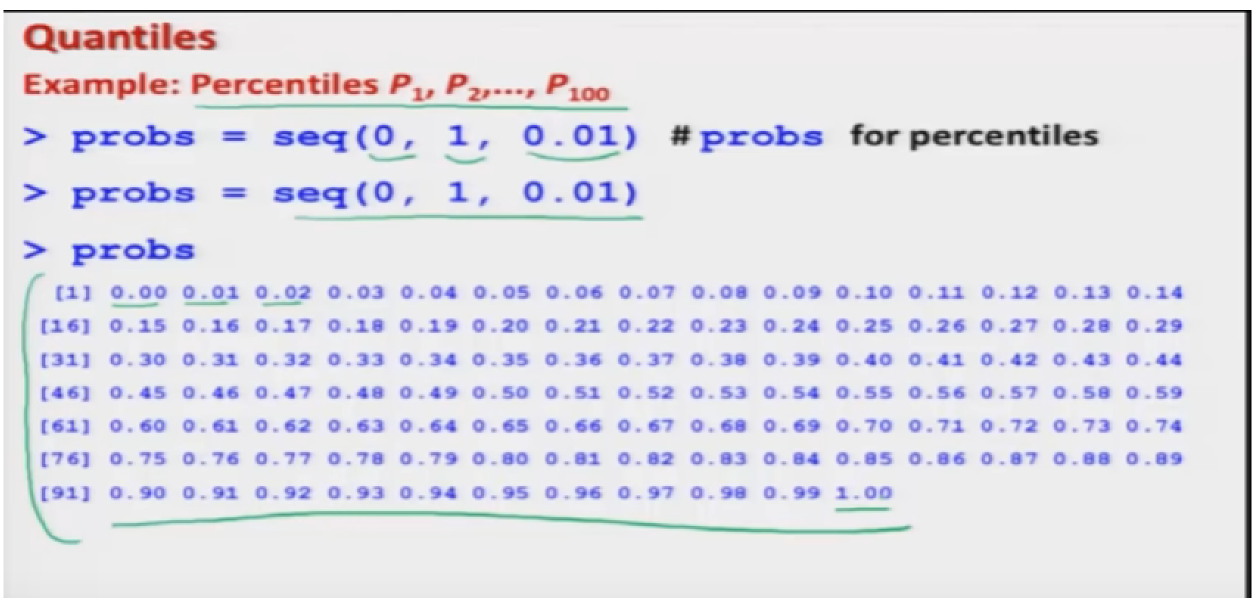
Now after this, I try to show you that how you, would try to find out the, other types of quantiles, like as percentile and deciles. So, first we try to understand, how to generate or how to compute deciles. So, now you have understood, it is very simple the command, is the same what you have to do? You simply have to change the probs vector or the data inside the probs vector. So, this aisles are essentially controlling, the value of alpha to be as 0.1 0.2 up to here 0.9. So, I simply have to generate here, a sequence from starting from zero and ending at one at an interval of 0.1. So, using the command seq I generate, such a sequence and this value is coming out to be here like this you can see here. And now I simply use the same command, which I used earlier but in this case, I have simply replaced the value 0.25 earlier kuno 0.10. And you can see here, here I am getting the 10 value of that deciles. So, for example as we have explained earlier the zero percent is corresponding, to zero point zero, zero 10 percent, is corresponding to zero point one, twenty percent is corresponding to zero point two and similarly a hundred percent is corresponding to one. And these are the values, of say this here D 1, D 2, D 10 this is here D 1, this is here D 2, this is here d3, d4, d5, d6, d7, d 8, d 9 and here D 10. Right? So, these are the values, of deciles.

Refer slide time :(30:33)



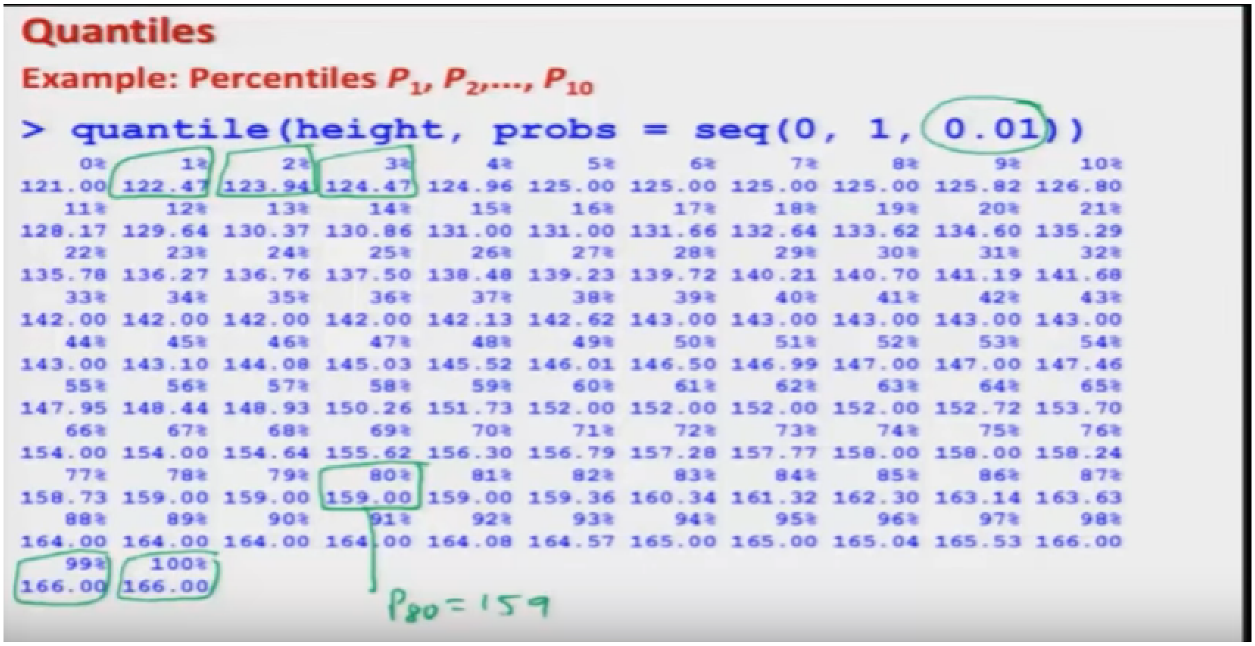
And similarly this is here that screenshot; I will show you on the screen also.

Refer slide time :(30:37)



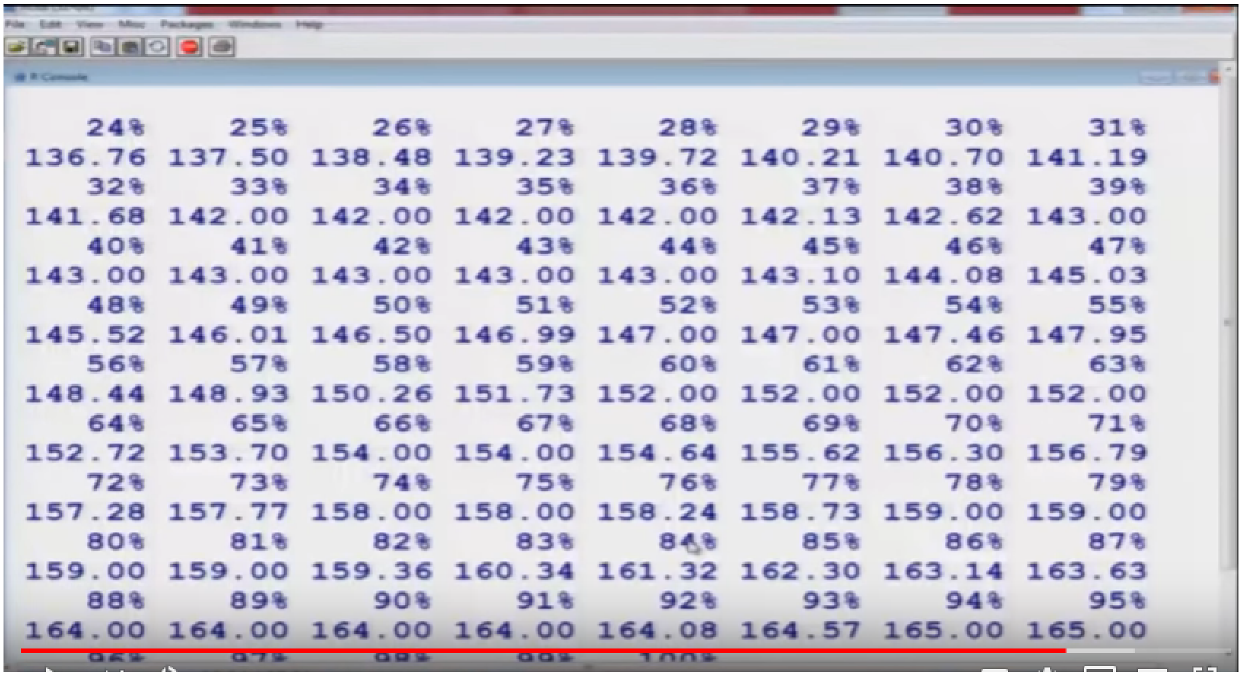
Now we try to compute the percentiles, P 1, P 2, P hundred. So, now it is pretty simple, I simply have to generate, a sequence from 0 to 1 at an interval of point 0 1. So, I try to generate, this sequence and you can see here, these are the values which are obtained here 0.0 0 .0 1.0 2 up to here 1.

Refer slide time :(31:04)



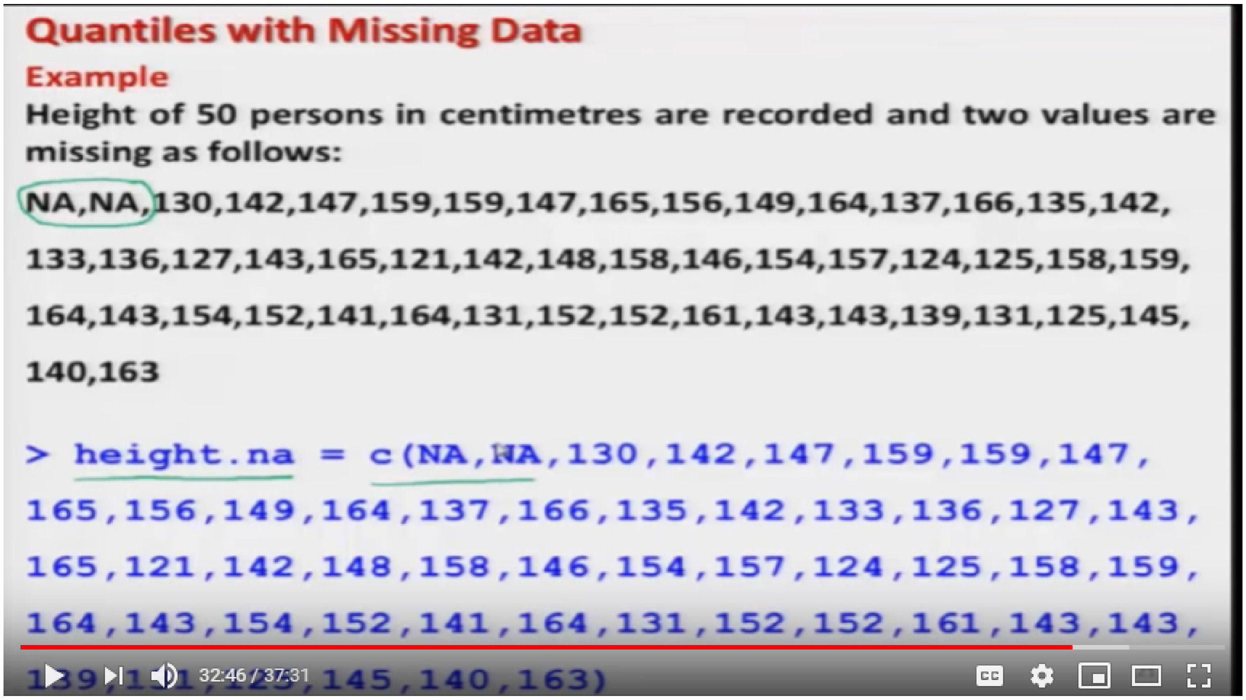
So, now just using the earlier, command means I generate the quantile. So, I simply have a replace, the value in the probs by here point is 0 1 and you can see here that the values of all hundred percentiles, they have been obtained, for example this the value of first percentile, this is the value of second percentile, this is the value of third percentile. And similarly in the last this is the value of here 99 percentile and this is the value of here, handed percentile, similarly in case if an examination needs that the candidates, must have the marks more than 80th percentile, then this is going to be controlled by this 80% and the p80 here, is 159. So, this means the candidate needs, to have marks, more than 159. Now before,

Refer slide time :(31:59)



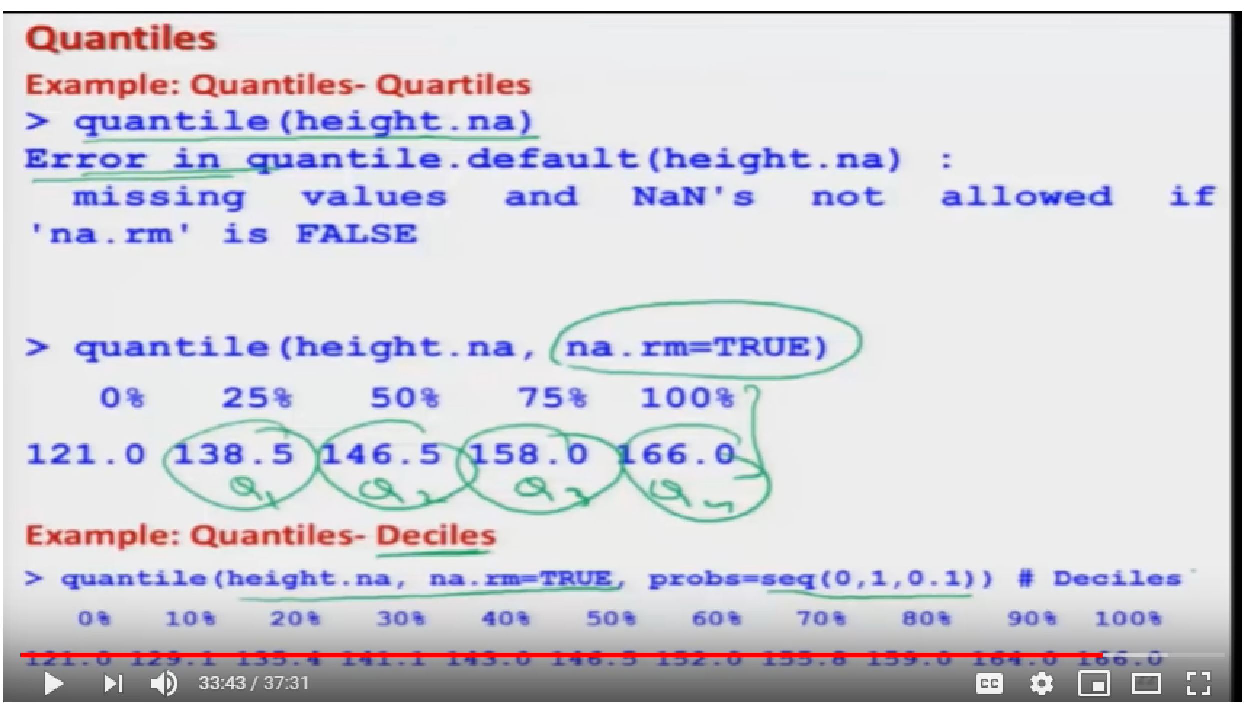
I go further let me try to show you, these things on the R console. So, I try to compute here. The deciles you can see here, the ideal that it can value and similarly if I try to change here, only the last value, then this will give me hundred values, which are here the percentiles and they have been presented on the slide. So, you let me come back to my slide.

Refer slide time :(32:27)



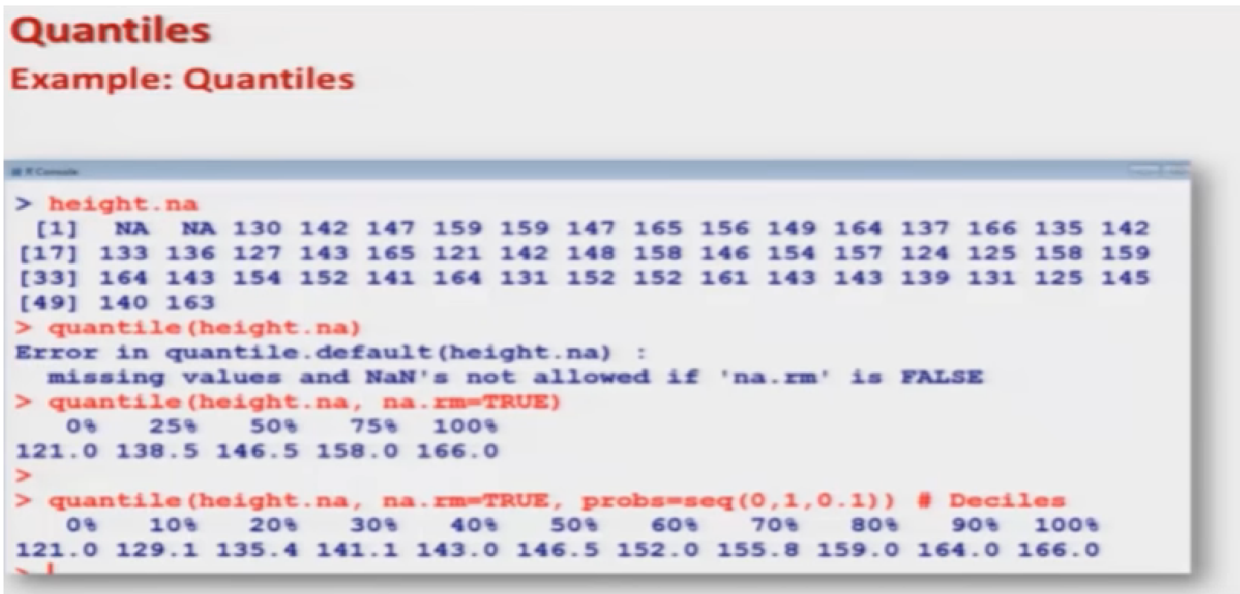
Now I will try to continue, with the same example and I will try to show you that in case if the data is missing then how we are going to compute, the quantiles. So, I have replaced, the first two data values by here na and I have stored this data into hi dot an here, as we have done it earlier and now I will try to compute,

Refer slide time :(32:50)



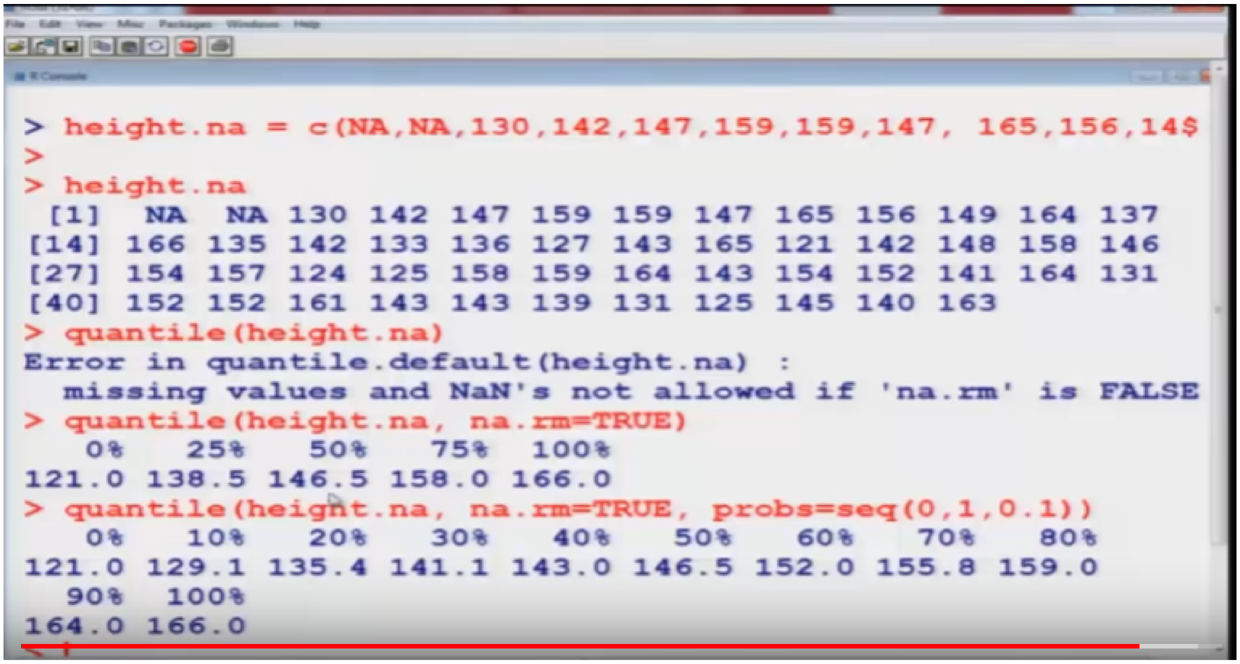
the same thing, whatever we had to compute, for this hi dot na. So, you can see here, as soon as I give here only the quantile function without, any specification of any dot rm, this will give me an error. Right? And so, I try to add in the quantile command and a dot rm is equal to true and this gives me this data. That is the first quartile, second quartile, third quartile and fourth quartile. Right? And similarly in case if I want to find out that deciles in the same data, I have to use the same command here. And I have to use the probs, which is the sequence of 0 to 1 at an interval of 0.1. So, this will give me the deciles so, you can see here these are the deciles.

Refer slide time :(33:48)



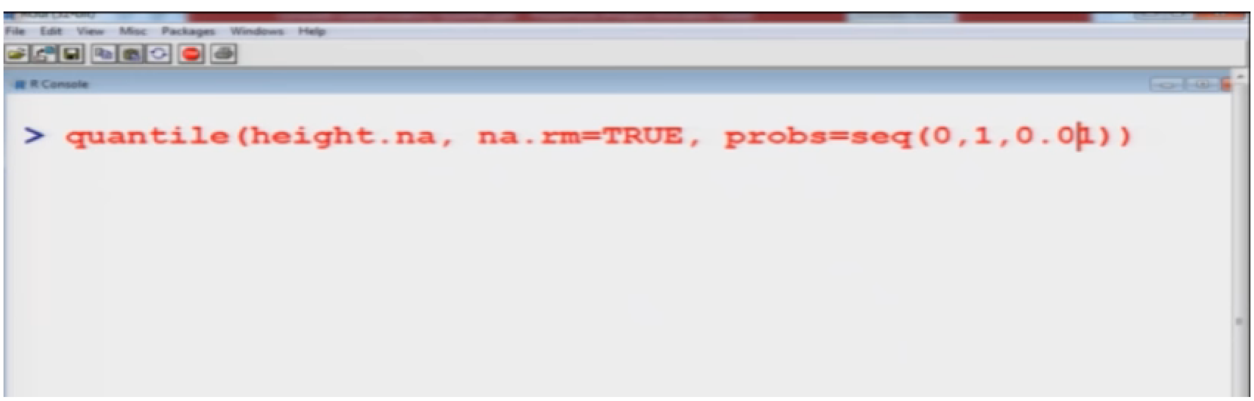
That we have obtained and this is the screenshot, of the same operation which I just shown you.

Refer slide time :(33:55)



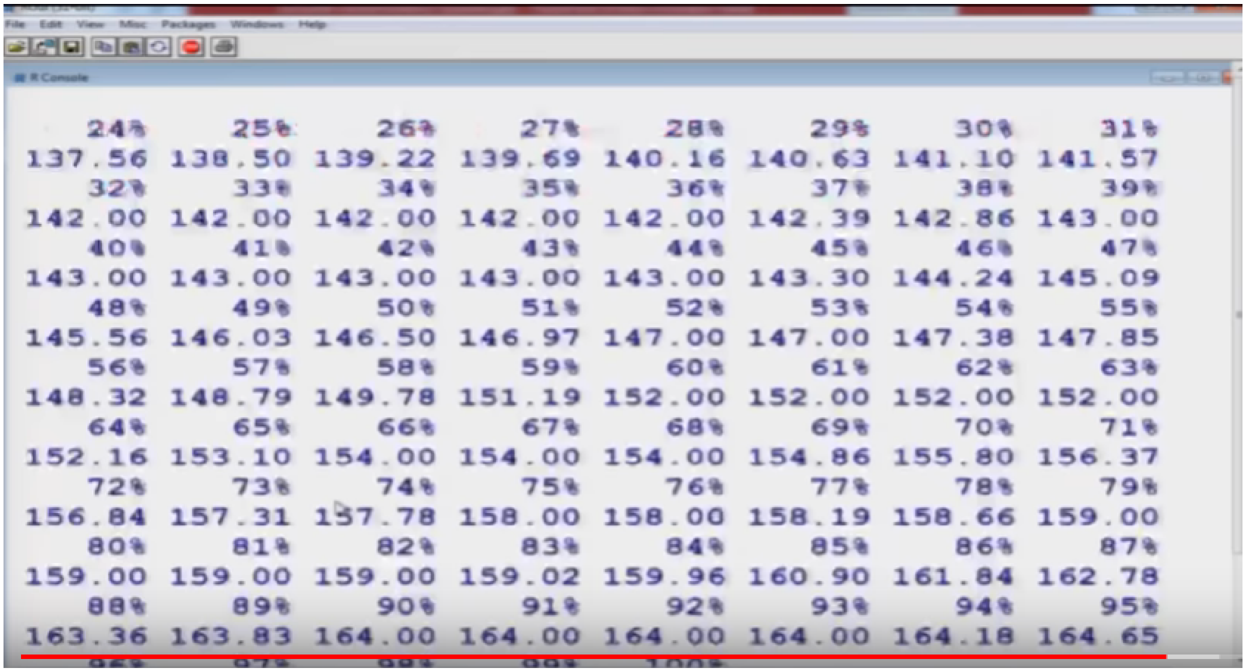
And now I would try to show you, this thing on the R console, to make you confident. So, I try to store this data I say here I dot na. So, you can see here this is my here, I dot ne and if I try to find out the quantiles, of height dot any this will give me an error. So, now I have to add here that n a dot rm is equal to true. And this will give me the values, of the quartiles and similarly in case if I want to find out say deciles then I have to add here the function, probs is equal to sequence starting, from 0 to 1 at an interval of 0.1, which is here like this. So, you can see here this is here the deciles.

Refer slide time :(35:01)



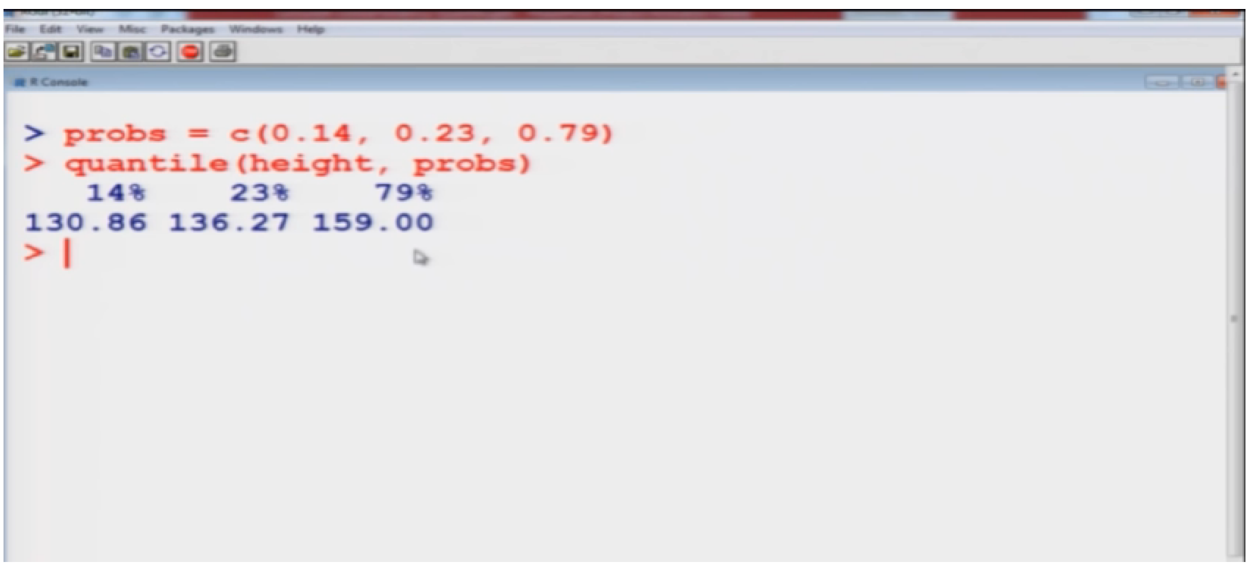
And similarly in case if you want to compute here, percentiles in this case you simply have to change, this value to be here, point 0 1. Okay?

Refer slide time :(35:06)



You can see here, these are the different percentile.

Refer slide time :(35:13)



Now I would like to show you that in case if you want to compute, some other percentiles, which are not in the equal width, suppose I want to compute here the 14th percentile. So, I will try to give it here, the value 0.14 suppose, I want to compute the 23rd, percentile. So, I will try to give it here the value as 0.23 suppose, I want to compute the 79th percentile. So, I will try to give it the value here point seven nine. Okay? And now if I try to write down here, quintile of the data, Heights and with probs, given by here double. So, you can see here you are getting here the fourteen percent, twenty third percent and seventy nine percent, percentile. So, this way actually you can compute different types of percentile and if you learn the statistic, there is a topic say testing of hypothesis, where'd we define the level of significance, those the definition of level, of significance is nothing, but it but it is related to computing a particular percentile. So, this concept is going to be very, very useful for you, when you go for further courses, any statistics. So, I would like to stop here and I would once again request you, to, to look into this concept, try to practice them, on the R software and we will see in the next lecture again. See you and good bye.