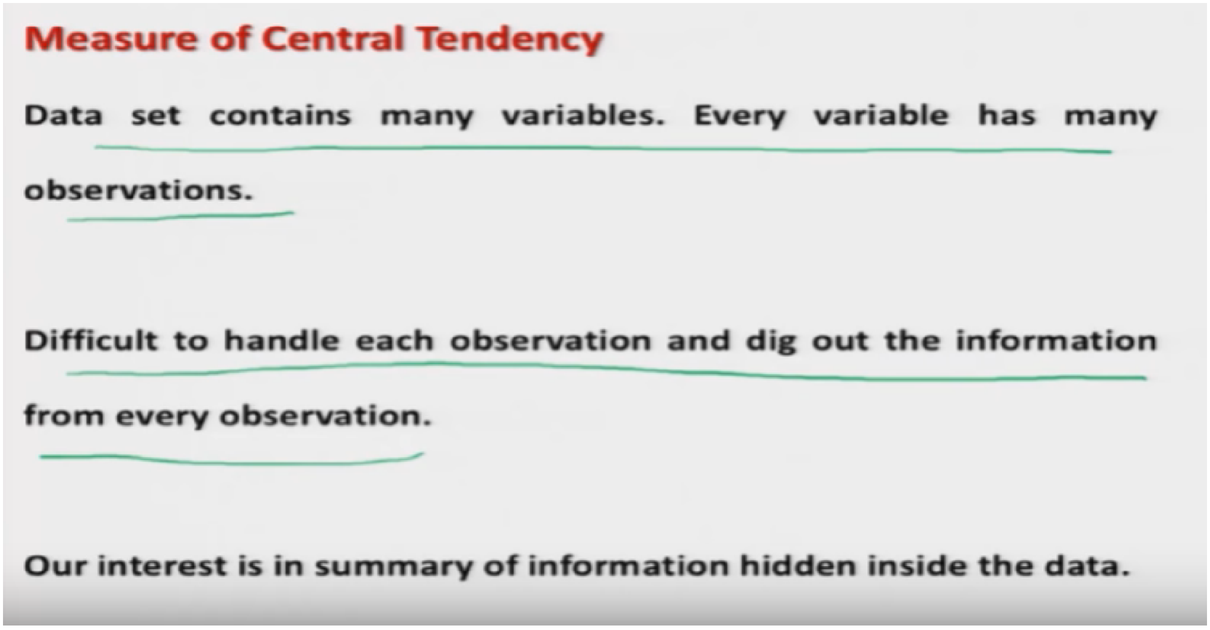
Lecture 14-

Central Tendency of data –

Arithmetic Mean

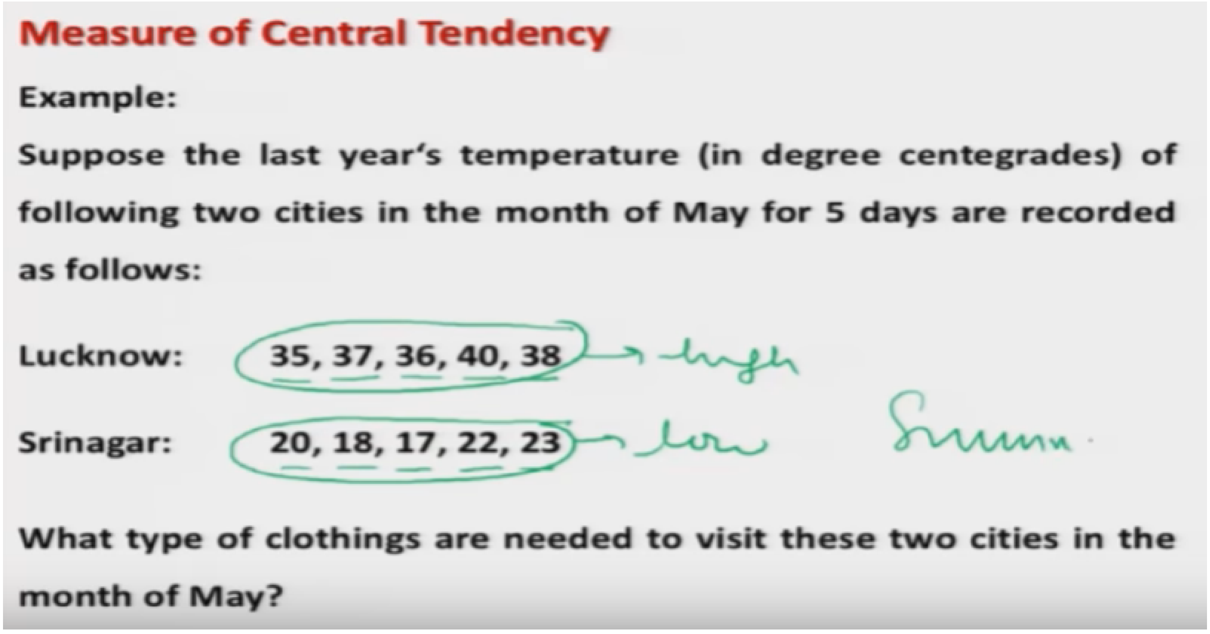
Welcome to the next lecture on the course, descriptive statistic with R Software. Up to now, in all the earlier lectures, we handled how to create,different types of graphics. And they were the part of the graphical tools of descriptive statistics, from now onwards,we are going to handle the analytical tools and in unit direction: that means when we have only, one variable. And we try to handle more than one variable;then again I will try to introduce the graphical tools in two dimensional and analytical tool for to the dimensional data. So,the first step after we get the data, is that we would like to get some quantified information: that is hidden inside the data, as we had discussed the data, is very silent, data cannot speak,data cannot tell you, well I have this value, I have the Chris information and graphical tools, will give you a graphic view, visual, information, from that you have to use your knowledge, your common sense, your information, your aesthetical knowledge, your information from the experiment and you need to combine, them to get a clear-cut conclusion. Now, we would like to quantify: that information.So,when we talk of the information contained inside the data, there can be normalized information which is contained, whatever question is how should we take it out? So, we had discussed that we would try to take the or extract the information on different aspects, of information like as central tendency, variation symmetry etc. So, we are going to start it,with a new topic, in which I'm going to discuss the central tendency, of the data and then I will try to, discuss different types of tools, which are used, to extract the information on the central tendency of the data. So, in this lecture I am going to, discuss the aspect of arithmetic mean.

Refer Slide Time :(2: 58)



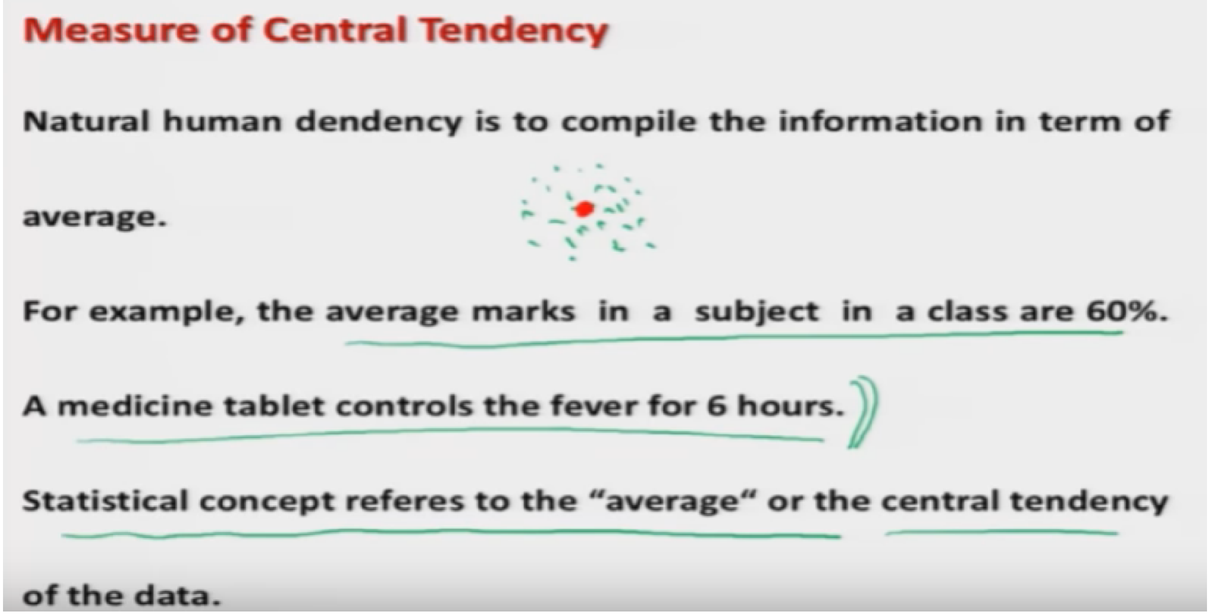
So now, whenever we try to conduct an experiment, there will be several aspects and we try to collect the data,on those aspects. So, finally the data set may contain many variables,several variables and every variable, may have many observation and our basic objective is this we want to know the,information contained inside the data,which is not possible, so we are trying to develop the tools which can help us in, digging out the information, from every observation. Now, the question is this, what we would like to have? Suppose I here hundred Decatur points and every observation, tells me something or alternative is that, instead of having hundred pieces, of information, I have a summary information: that may provide,more information, to a common person and that will be more useful. So, here now we are looking forward to understand some summary measures, which can give us the information hidden inside the data, on different aspects.

Refer Slide Time :(4: 14)



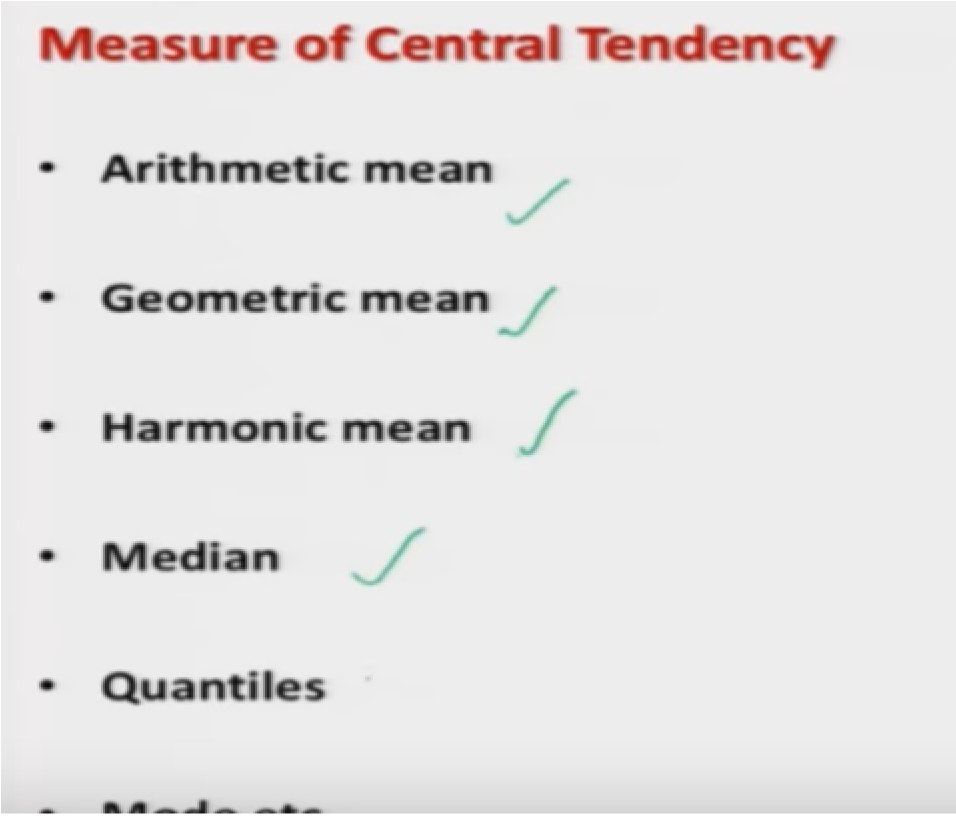
Now, can we take a simple example, to explain my view,suppose I want to,raise as kitty,I have two choices, Lucknow, in Uttar Pradesh, which is quite hot, during the month of May, it is the summer season and similarly other cities, Srinagar in and Kashmir. Which remains cool during the month of May? So, now we have collected the data on the day temperature of last year, on say five days and this data is coming out to be, 35 degrees centigrade,37 degree centigrade, 36, 40, 38 degree centigrades, for Lucknow and 20, 18, 17, 22,23 degree centigrade for Srinagar. Now,what information I can get from this data? This data can be a large, there can be means,I have taken here only, 5 values for the sake of understanding, but these values may be, hundred thousands and, and,and even millions. So, from this data, I would like to know for example: that what type of clothing's I should, take there, in case it is cold then I would try to take some Booleanclothing's and if it is hot and I will try to take some, simple cotton clothing's. Right? So, now how to get this information? By looking at, this information yeah! It is telling me that,the temperature is quite, high and here the temperature is here,usually low. But, we would like to have a summary information, the information in the summary. Now, what we observed? That is the human tendency, to compile the information, in terms of averages for example, in case if I say, in a class, some students might have got 45%,somebody has got 55%, somebody has got 65% and there will be more marks, but then, I am moreinterested, what is the averageperformance of the class?

Refer Slide Time :(6: 35)



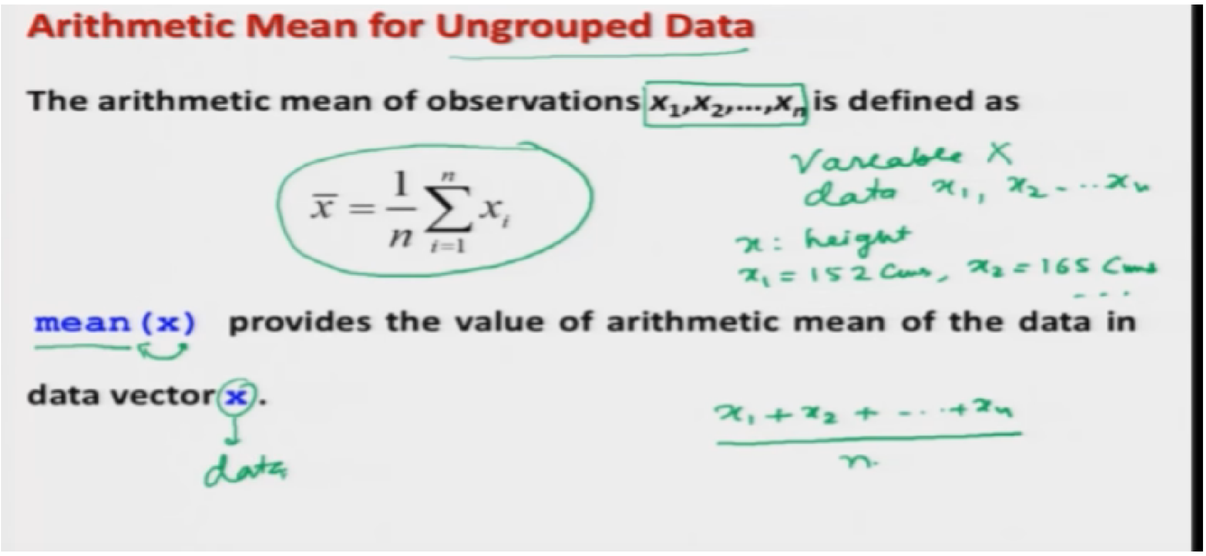
So, if I sayhere: that the average mass in thesubject in a class are 60%, then it makesmore sense. Similarly in case if I amtrying to go for a tablet, of a medicine and suppose the shopkeeper or the doctor tells me: that this tablet can control, the body temperature and bring the temperature down, for six hours, what does this mean?The doctor is delegate the average value and we are very, easily understanding it,this six hours, cannot always be exactly six hours, this may be five hours, this may be seven hours, this may be five point five hours, but and this may be six point five, five hours, but this data, has been collected and doctor has found a sort of automatic mean or an average and he's convinced that, if this medicine is giving to a person, having a fever, then this, tablet can control the body temperature up to six hours or say on an average six hours. This is what we mean?So, when statistics, this concept refers to as average or the central tendency of the data, central tendency of the data means for example, if I have a data, which I plotted here like this, then I would try to see, what is the point here? Around which the data is concentrated for example, here you can see, the Chris is trying to give us a central value, around which the data, is concentrated.In statistics we have different types of measures, to study this, central tendency of the data,

Refer Slide Time :(8: 14)



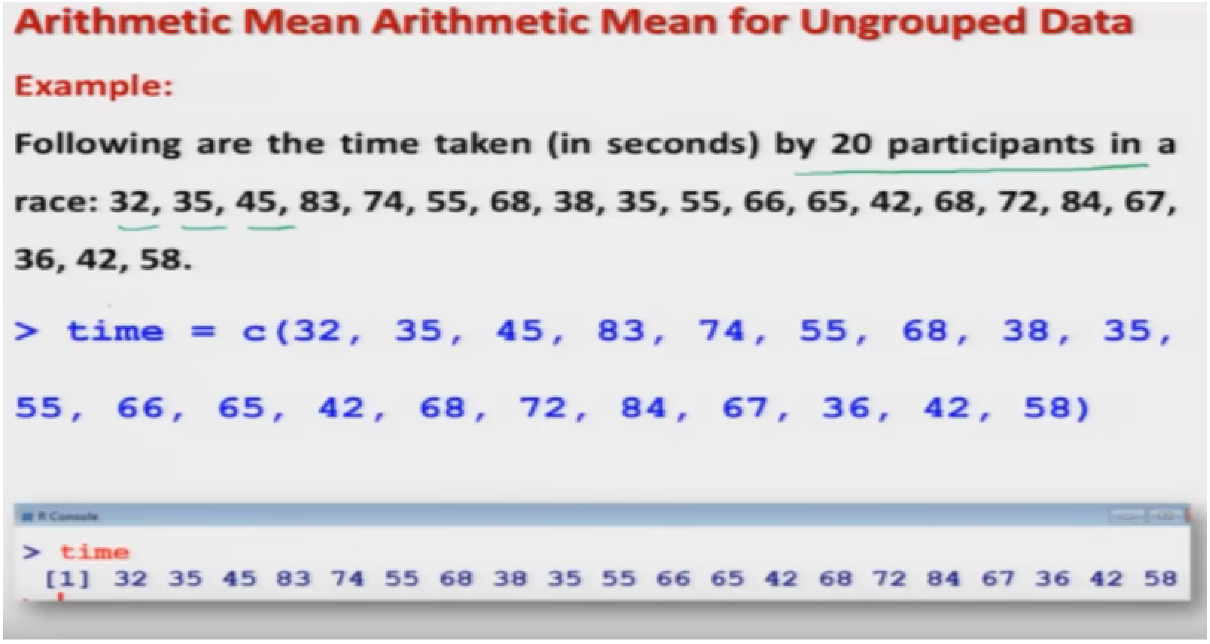
for example, automatic means,geometric mean, harmonic mean, median quintiles mode, etc.

Refer Slide Time :(8: 19)



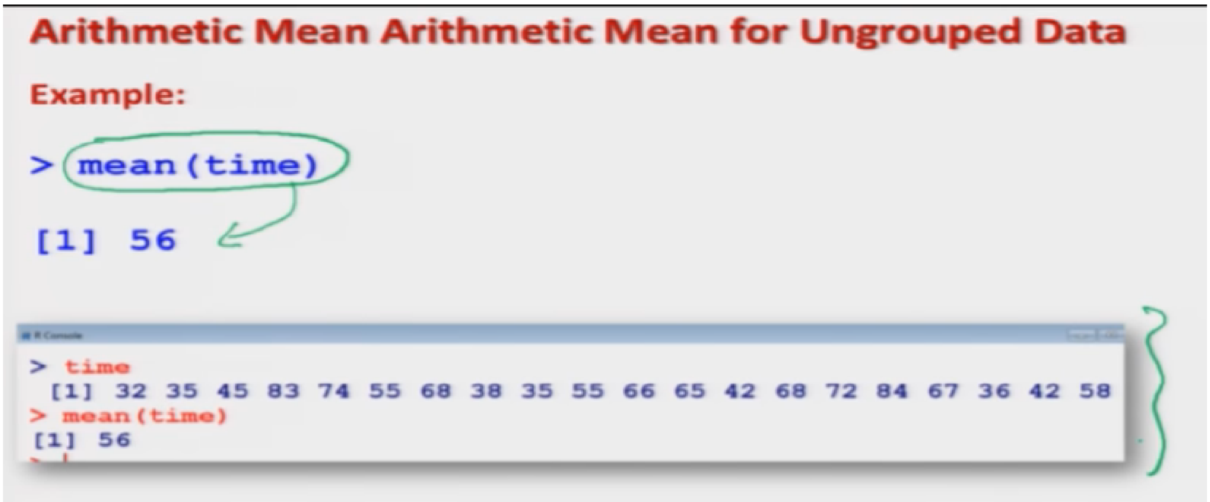
So, we are just going to discuss these measures one by one and I will also try to show you: that how to compute them on the R software. So.first let me try to explain, the automatic mean, foreign and group data and group data means, we have a variable here X and we have collected the data on X, s say here, x1, x2 and soon. See here, X n, so for example, if I say here X, is my hair height, then x1 is going to be the height of first, person say hundred fifty two centimetres, x2 is going to be the height of, the second person say hundred sixty five centimetres and so on. And then we have, total number of an observation,which are denoted by X 1, X 2, XN and these X's are small letters.Okay? Now, the automatic mean, of this observation is defined by like this X bar, is equal to 1 upon n, summation I goes from 1 to n X I, this means, I have first, to sum X 1 plus, X 2 plus, X N and then I have to divide this, sum by here the number of observation and, this is the meaning of this symbol. In order to compute, it in R, the command here is, mean and then I write, mean and inside the argument X, then my data, is contained in the X, using the C command.

Refer Slide Time :(9: 58)



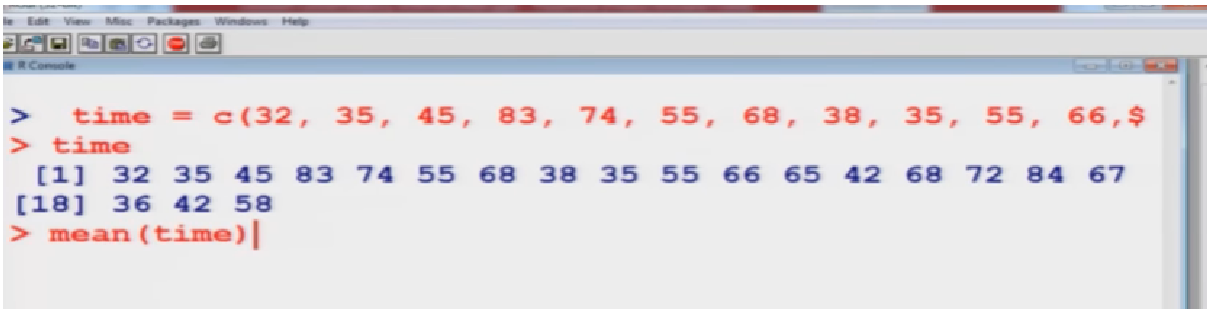
So, let us try to make an example and try to see here,well in case if you want to know, more about this mean for example, I will be discussing here, another aspect that how to handle the missing values, but it but,but there can be, trimmed mean also and there are some more parameters, with this command, I would request you to go to the help, of mean and then try to look into different parameters. So now, coming back to my example, so this is the same example: that we had considered earlier:that there are 20 participants, who participated in a race and their time in second, seconds has been recorded here,like the 32 second, 35 seconds and so on.And this data has been captured, inside a variable here stay here time.

Refer Slide Time :(10: 46)



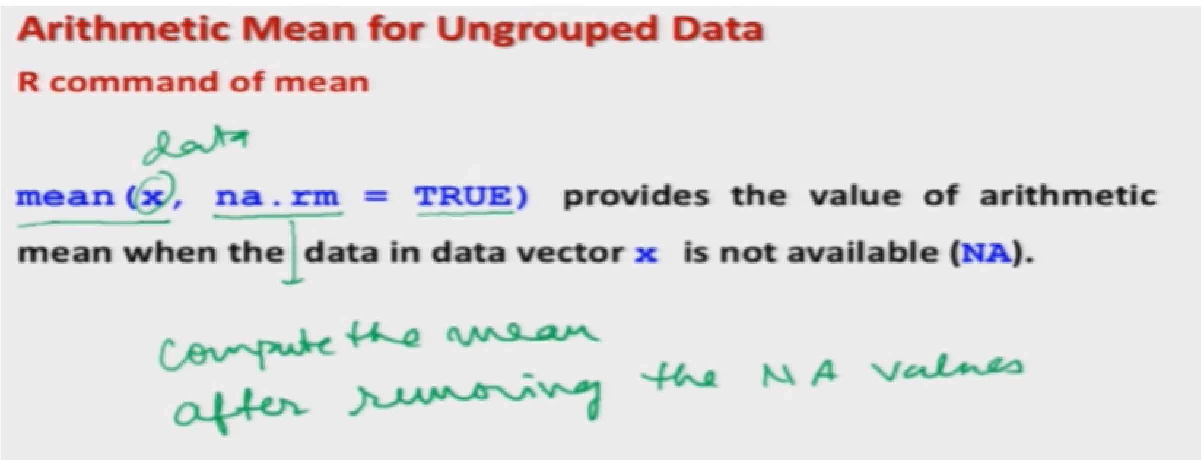
Now, in case if I want to find out here this variable,then I simply have to type here, mean inside the argument, the variable name time and I get it here, as the value, 56. So, you can see here, this is the screenshot. So, I will try to show you here: that how it works on the,R console.

Refer Slide Time :(11: 10)



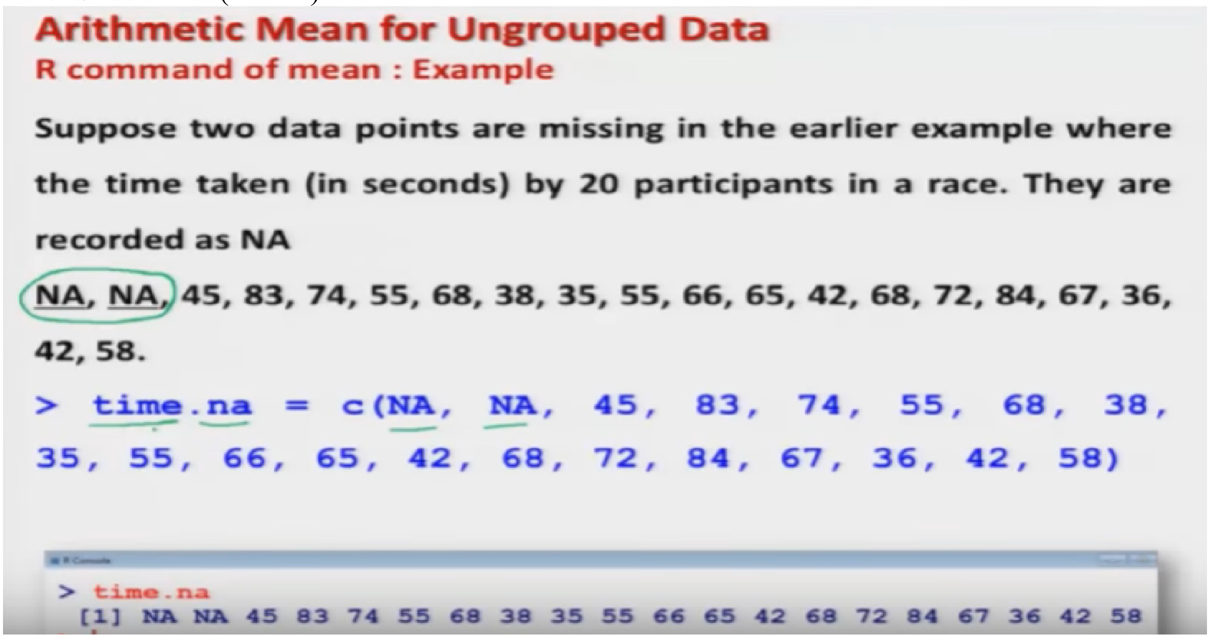
So, let me store the data, over here so that you can see here, this is my here time and when I try to find out here mean of for your time, the variable,in which the data is contained, this comes out to be, 56.

Refer Slide Time :(11: 24)



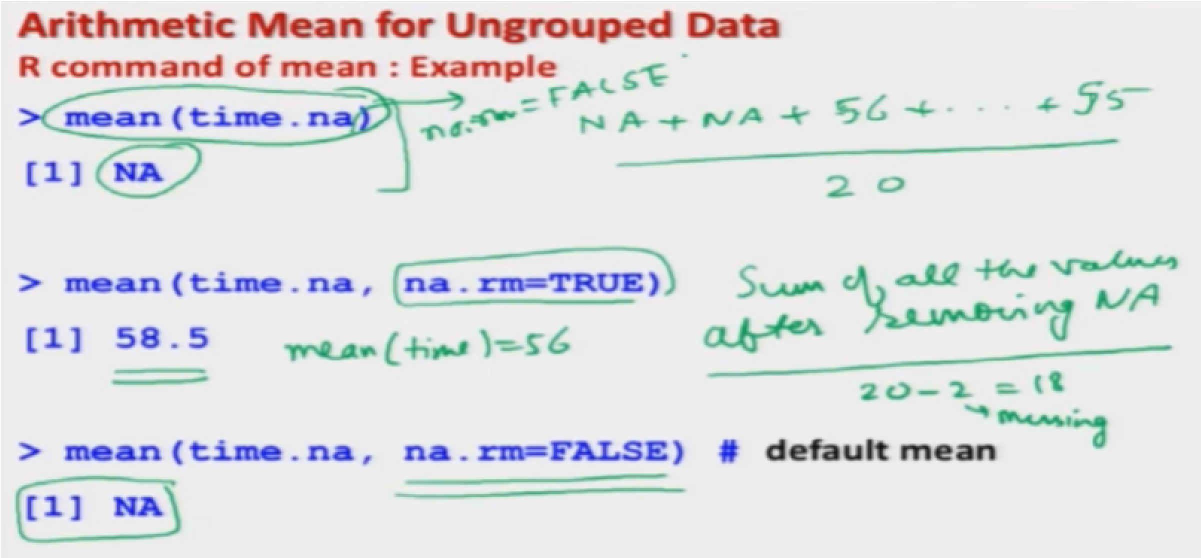
Now,I try to address here, one more aspect, if you remember, when we started our course, then in some initial lecture, I had given you an idea, there,there can be many situations, where some data might be, missing and this data, is represented as capital N and capital A:that is N A and which is a result value.So, in case if the data, is missing, then how would you like to compute, mean and other components, you see, the way I am going to explain here: that how to compute the mean, when data is missing,the same concept will be used, in all other cases: that you want to find out,the variance the standard deviation or the median, the same concept and the samecommand, will be used. So, here I will tryto, explain in detail and after that Iwill quickly, take it. So, when some data,is missing, then in that case, the meancommand,is used to find out the average value,automatic mean of the data in X vector here? But, there is another parameter which is added here and a dot R M, is equal to true, capital T, capital R,capital U, capital E. So, this is, trying to tell that please, compute the mean,compute the mean, after removing the, the n a values, n a or the missing values.This is, what is trying to say? So, in order to understand it,

Refer Slide Time :(13: 12)



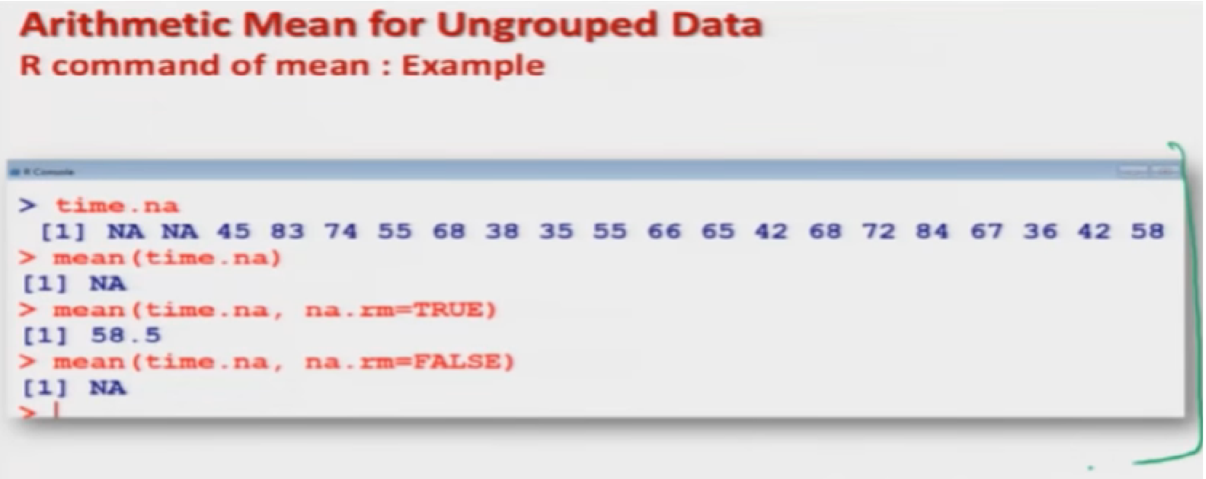
look at me take the same example and you can see here:that I have replaced, the first two values, by n a. I have just made it underlined, so that, you can easily, see it and so in, this case my data vector will contain, first two values here as say and a and, and a. And in order to, store this value I am trying to use a different name and this name I am time to give, the time which I had used earlier, dot n a, dot n a is not a result,what? But, this is simply trying to indicate: that the data on the time variable that we have used earlier, this is the same data with na. Right?

Refer Slide Time :(13: 54)



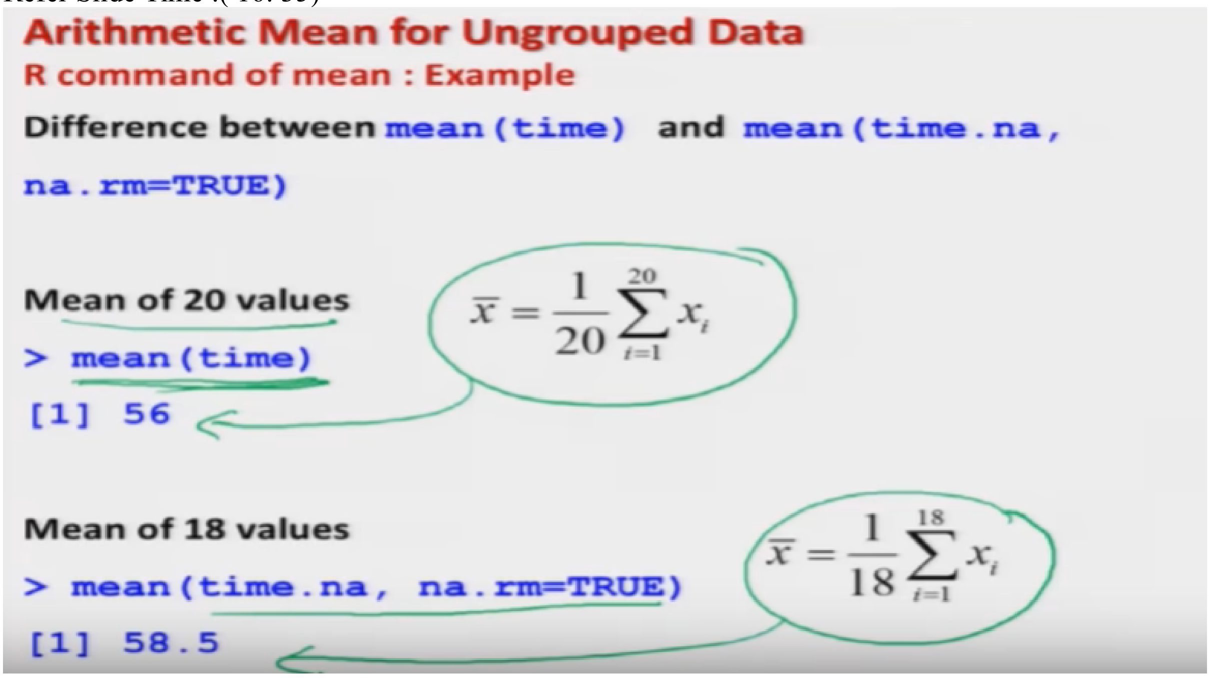
So, in case if I try to do it here, then you can see here, if you try to find out the mean,only of the time dot n a vector,where the data is missing, this will not give you any numerical value. But, it will give you simply here a say n a. Why?Because this is going to find out the sum,of here and a plus n a plus, the value but, the numerical values here, 56 plus dot whatever are the values here,divided by here, 20. Right? So, that is why?This value is coming out to be n a,you can see here: that this na, plus this na,plus this value, plus this value, plus this value and so on, divided by 20. So,definitely this cannot be computed so,this is giving you n a, whereas in case if you try to adhere, the command n a dot R M is equal to true, then, what it is trying to do? It is trying to, find out the sum, of all the values, all the values after removing n a, divided by the number of observation, which now becomes here, 20 minus 2: that is there are 20 observation and 2 observations are missing. So, this is going to be here 18, so this number is going to be divided by 18. And now, you can see here: that, this is giving me a value 58.5, where? If you recall: that the mean of time,which it was 56. So, this is now change, because this has been computed, after removing the missing value. Now, in case if I try to make this n a dot RM to be here false, like as here,then you will see here, this is giving me the outcome na, because once you say, na dot RM: that this is the default function. And when, you are trying to use here the, mean of time dot na,actually, this is, the default that here n a dot RM is taking always, SC Falls as a default.So, whenever we are trying to find out the value of the mean, it assumes by default that all the values are available. In case they are not available, you need to inform your R software: that yes, there are some values which are not, available and please, try to compute the mean after removing those, missing values. Okay?

Refer Slide Time :(16: 32)



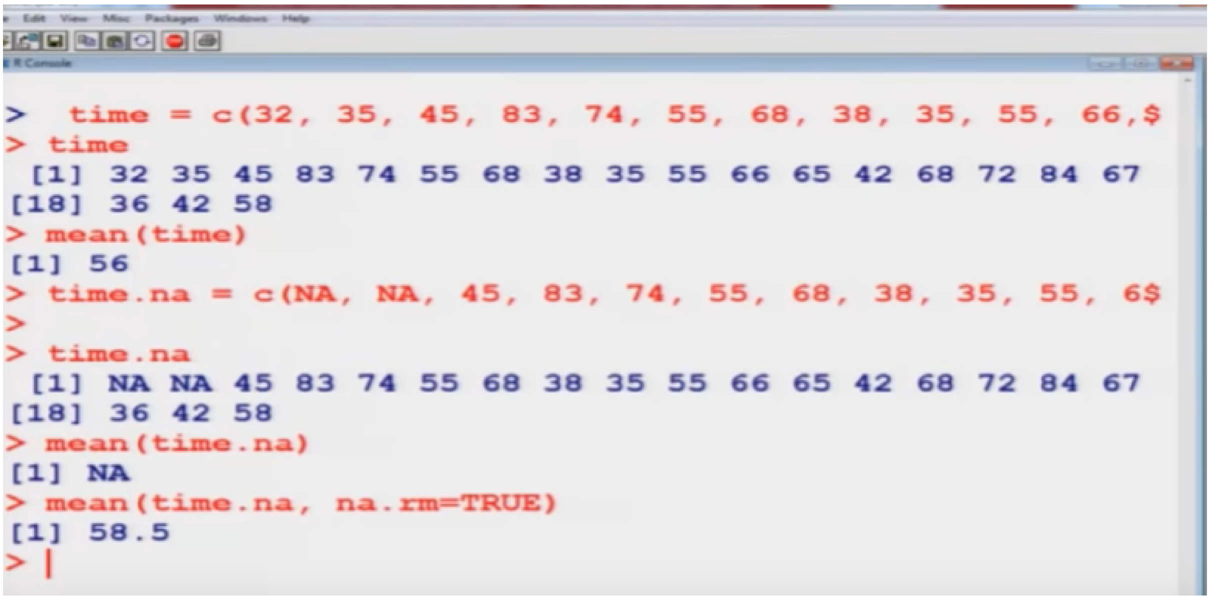
Now, this is here the screenshot, I will try to show you on the R console.

Refer Slide Time :(16: 35)



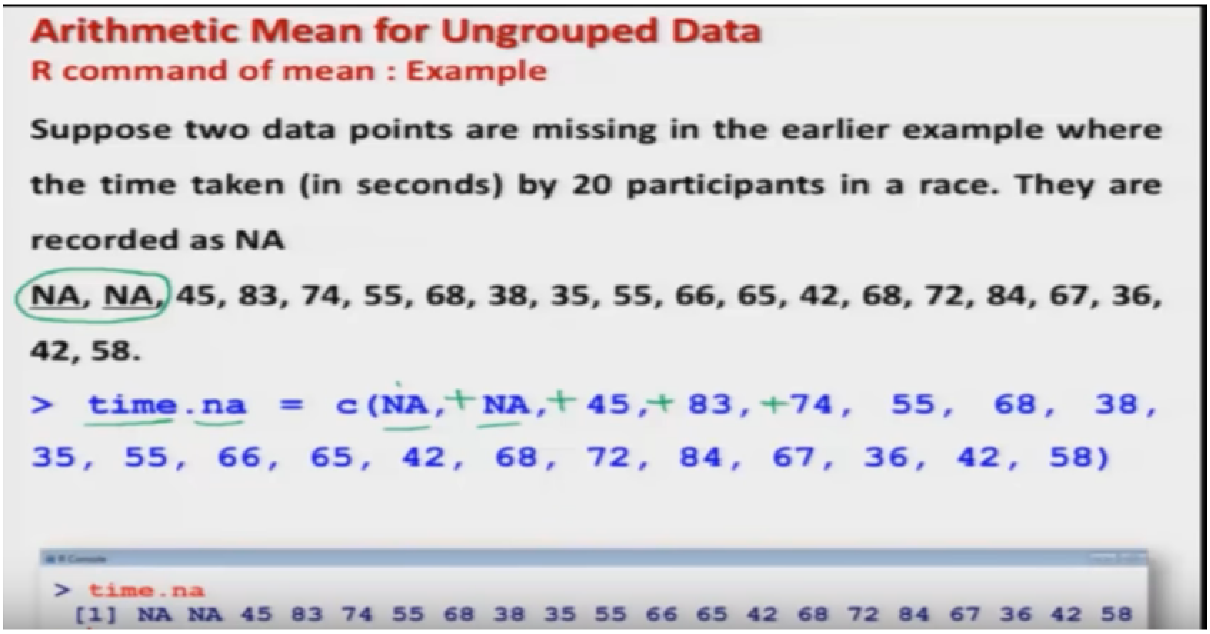
But,before that, let me try to show you here:that in this case, which I just discussed actually:that mean of this time, which is containing here, 20 values, is computed like this, sum of all the 20 values divided by 20 and whose value is coming out to be 56, whereas when you are trying to use, the data with missing values, then it is actually, based on only 18 values and this value is coming out to be 58.5. Now, look at me, first come to R console and try to show you here, was really happening. So, I try to copy this data,

Refer Slide Time :(17: 14)



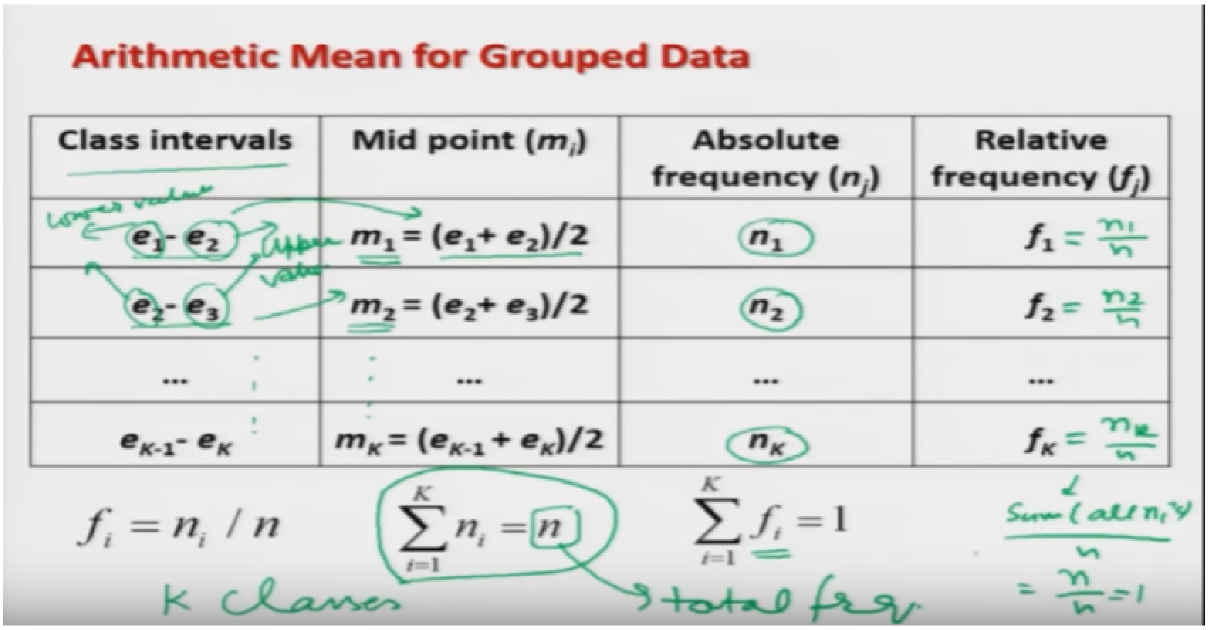
data here, so you can see here, time dot n aconsole has, this thing. And now, if I try to find out the mean of here, time dot n a, this comes out to be here n a, but in case if I try to add here, one more parameter and a dot RM, is equal to here true, then you can see here, this is coming out to be 58.5, this is the same outcome that we have, discussed in the slides. Okay?

Refer Slide Time :(17: 50)



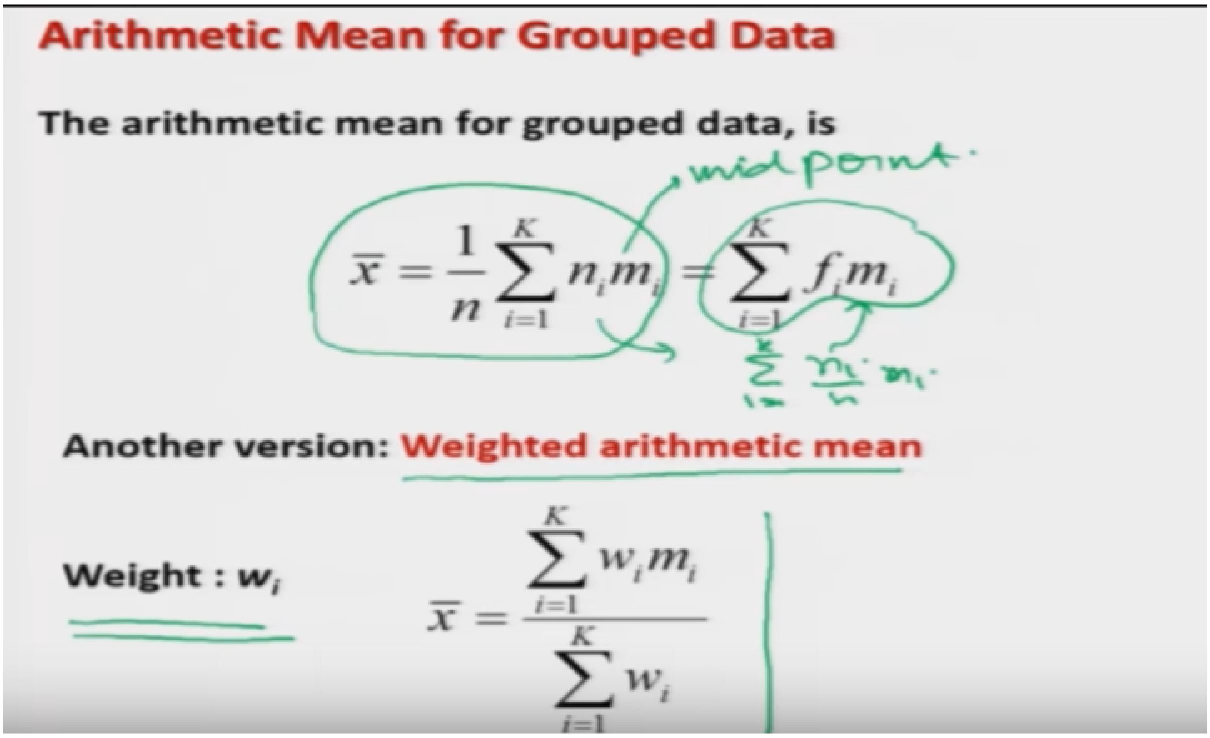
So, let me come back to our slides and quick, let us try to have here another aspect. So, up to now we have discussed the automatic mean, for L and group data.Now, I'm going to discuss how to compute the automatic mean, in case of group data,you remember: that in the case of group data, first you need to construct a frequency table. So, now we will learn, how to compute the arithmetic mean, from the frequency table.

Refer Slide Time :(18: 22)



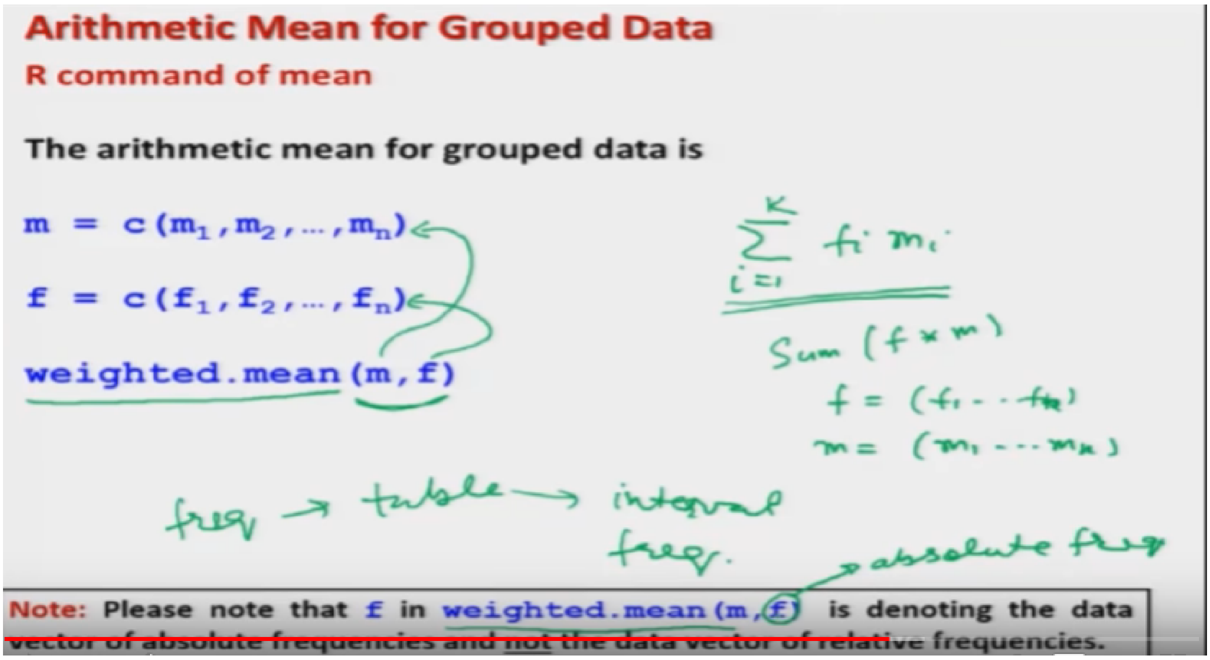
So, you may recall: that while constructing the frequency table,we had constructed the class intervaland the class intervals were constructedon the basis of given set of data andthey were divided in suitable number ofintervals, of suitable widths. And theseintervals I am, denoting as e1 to, e2 to e3 and so on. And this part here, the first value e 1 and here, this here e 2, in the second case, they are called the,'Lower Values' and similarly this e 2, in the first interval and III in the second interval, they are con the upper values, of the interval and so on,so similarly i have created here k classes. So, i have created here, case such classes and then I'm trying to find out the midpoint, of this interval, so midpoints of the first interval, i am denoting by here m1, which is simply here,e1 plus, e2 divided by 2 and similarly,the midpoint of the second interval is denoted by m2, which is the lower limit,plus upper limit, divided by 2 and so on and similarly i try to find out the weight values, of all the intervals and based on the given data, set i try to find out the absolute frequency. So, there are n 1 values in the first interval and two values in the second interval and so on, NK values in the Keith interval. And we also know: that the sum of all this n 1, n 2, n k we are trying to denote by here, n and the relative frequency, has been obtained say, SI here F 1, which is a n 1 upon n, F 2 is here N 2 upon and so on here, FK is, NK upon n and in that total frequency and is here the, total frequency. And means obviously, in case if you, try to sum all the relative frequencies over here.So, this will come out to be sum of all n a is divided by n.So, sum of all n is here's n,so this become and upon n, which is equal to here 1, which is written here.Right?

Refer Slide Time :(20: 31)



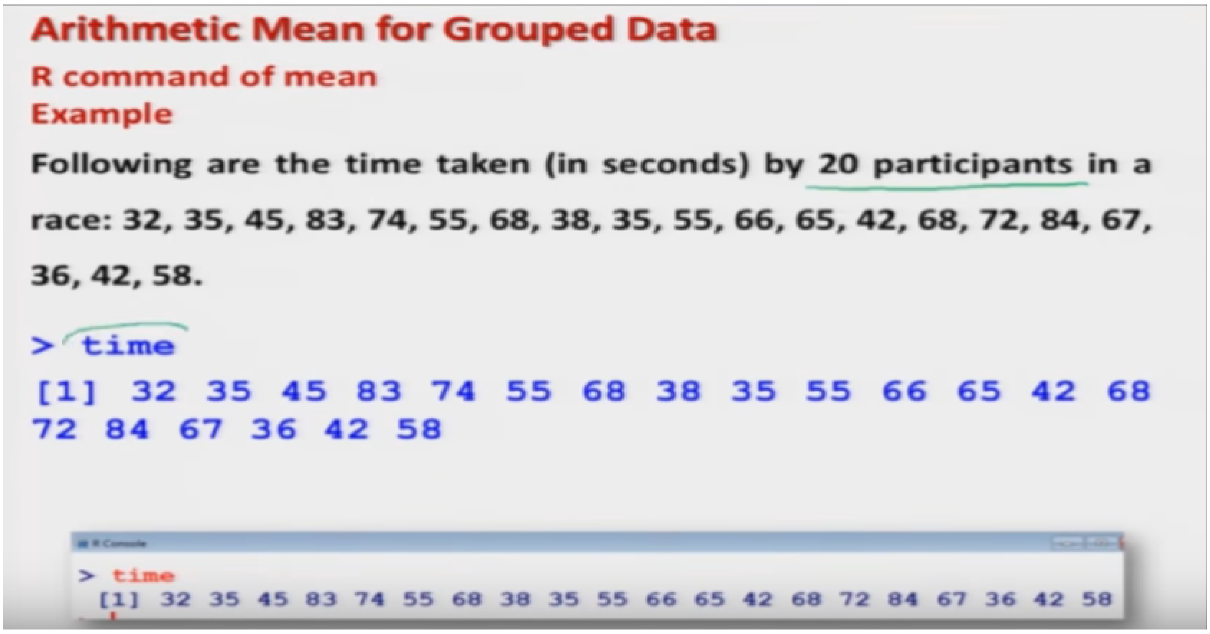
Now, I would try to define the automatic mean for this group data. And it is defined here as say X bar is equal to, 1 upon n summation I goes from 1 to K and IM I, so MI is here the, midpoint of the interval. So, now in case if you try to, simplify it, so this can be written here they say I goes from 1 to K and I upon n, into M I and so N I upon n is here, fi frequency, so other alternative is that, I can simply find it out here as the sum, of the product of m IL, fi. And based on this, there is another version, of the this type of mean in case of group data, which is called as see here,'Weighted Arithmetic Mean' and weighted arithmetic mean is defined, as say some of say WI mi upon, say divided by sum of WI where,w i's are the weights, which are assigned to the values, right. So, this is a more generalized, function which is useful in many, many statical, datasets.

Refer Slide Time :(21: 38)



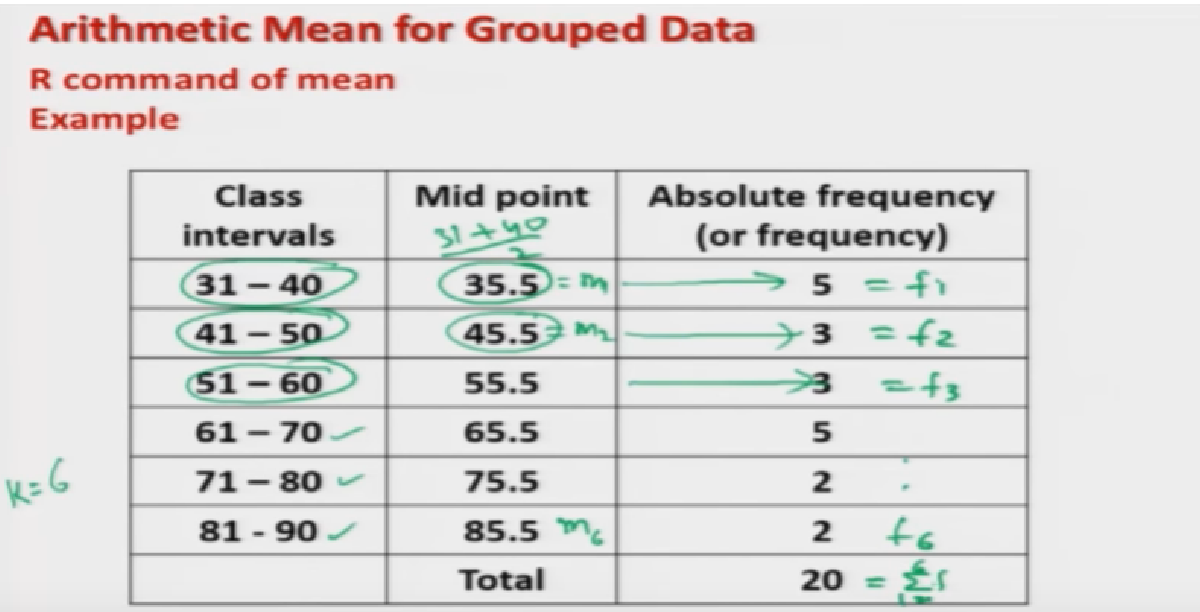
Now, in case if you want to find out here,the arithmetic mean of the group data set.So,we know this is now going to be simply here, I goes from here, 1 to K,which is here, see here, fi, mi. So, if you remember, when we were discussing, different types of mathematical operation using the R software, then we had discussed that, this type of thing,can be obtained by, say sum, of two data vector, F in to hear m. Right? So, we had F is going to be the data, on say F 1, F 2, F K and M is going to be the data, on say mid values, M 1, M 2, M K. So, in order to do it, this R has already a built-in function, which is called as,Weighted dot Mean', WEEI GSTED dot MEA N and inside the arguments, I have to give that two vectors here, m and f,where m is containing all the midpoints and f is containing, all the frequencies.So obviously, when you want to compute,the arithmeticmean of this group data,first objective will be to find out the frequencies and in order to find out the frequencies, you may recall: that we had,used the command, table and table will have two types of components, when the first one will be intervals and second will be frequency. So, we need to, extract the frequencies from the outcome of a table command. Now, you have to be,watchful here that in this function,weighted dot mean, I have used, this symbol F, to indicate the absolute frequencies, whereas, if you observe, in this slide, in this formula, I have used here F, to indicate the, relative frequency. So now, in this example and in order to compute the weighted mean I will be using the, the indicator F, to indicate the data vector, of absolute frequencies and not the data vector of relative frequencies.

Refer Slide Time :(23: 54)



So, let me take here an example and show you that, how the things are going to what? So, once again this is the same example: that I discussed earlier: that there are 20 participant, who participated in a race and their time is recorded in seconds.And this data has been recorded here,under the variable here time. Right?

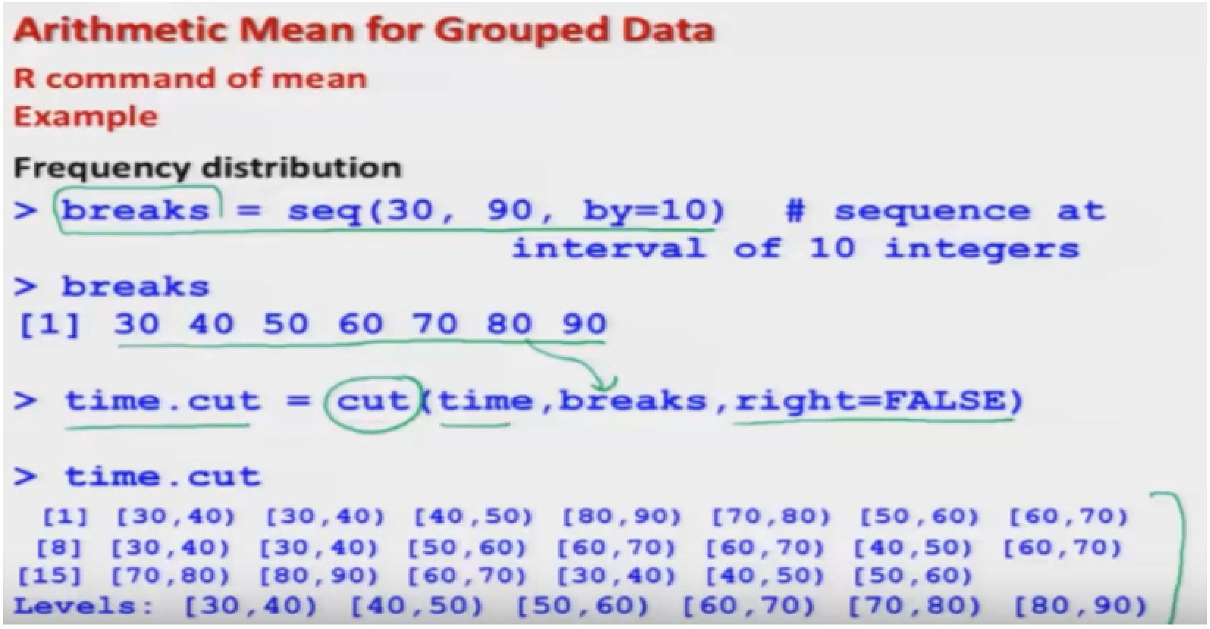
Refer Slide Time :(24: 17)



Class intervals	Mid point	Absolute frequency (or frequency)
31 – 40	35.5	5
41 – 50	45.5	3
51 – 60	55.5	3
61 – 70	65.5	5
71 – 80	75.5	2
81 - 90	85.5	2
	Total	20

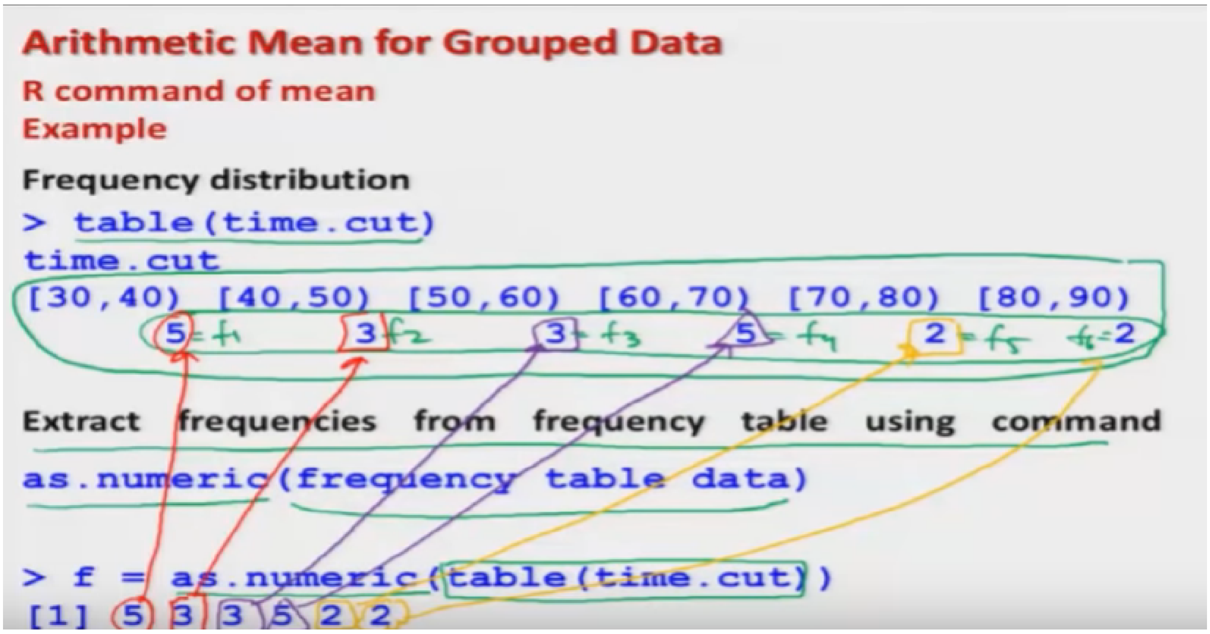
And now, we had converted this data earlier,in the form of a frequency table and youcan see here: that I have created herethe class intervals, like this 31 to 40,41 to 50, 51 to 60 and so on. So, there arealtogether, six class intervals, so K here is equal to, 6. And then, I have found the midpoint, which is here 31 plus 40 divided by 2 and so on, so this is the value of here m 1, this is the value of here M 2 and so on, so we have here, m6. And their absolute frequency have been obtained as here five, for the first class, 3 for the second class, 3 for the third class and so on, so these are the value of here f1,f2,f3 and here see here, f6 and this is here the value of here sum of all the frequencies, which is equal to here, n.Right?

Refer Slide Time :(25: 09)



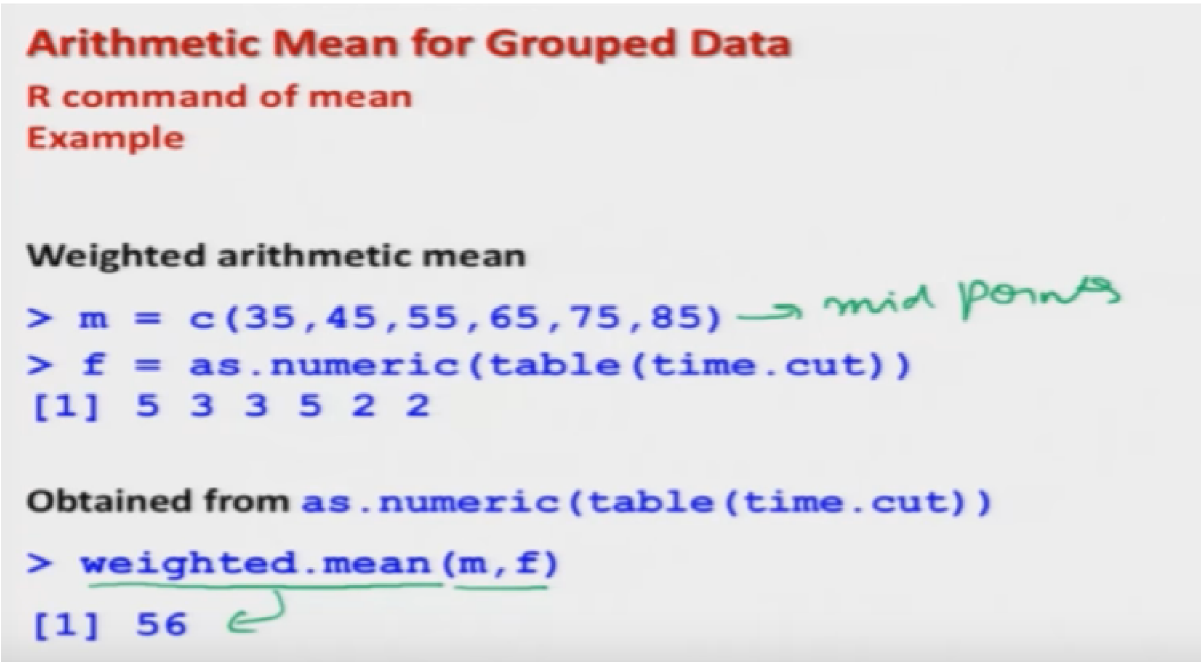
So now, just to give you a brief recall: that how we had found the frequency distribution, I mean I have taken some slide from the earlier,lecture and you may recall that first we had defined sequence, between 30 to 90, at an interval of 10, by using the command,seq and we had stored this data, inside a variable breaks. So, breaks was 30, 40, 50,60, 70 ,80,90 values and then, we had using the data time and using this data vector,here breaks and putting the right hand side, interval to be open, we had used the R command cut, C UT, to convert the data into factors. So and this data was, stored in time dot cut. So, this outcome was, like this, this all we had discussed earlier, right.

Refer Slide Time :(26: 03)



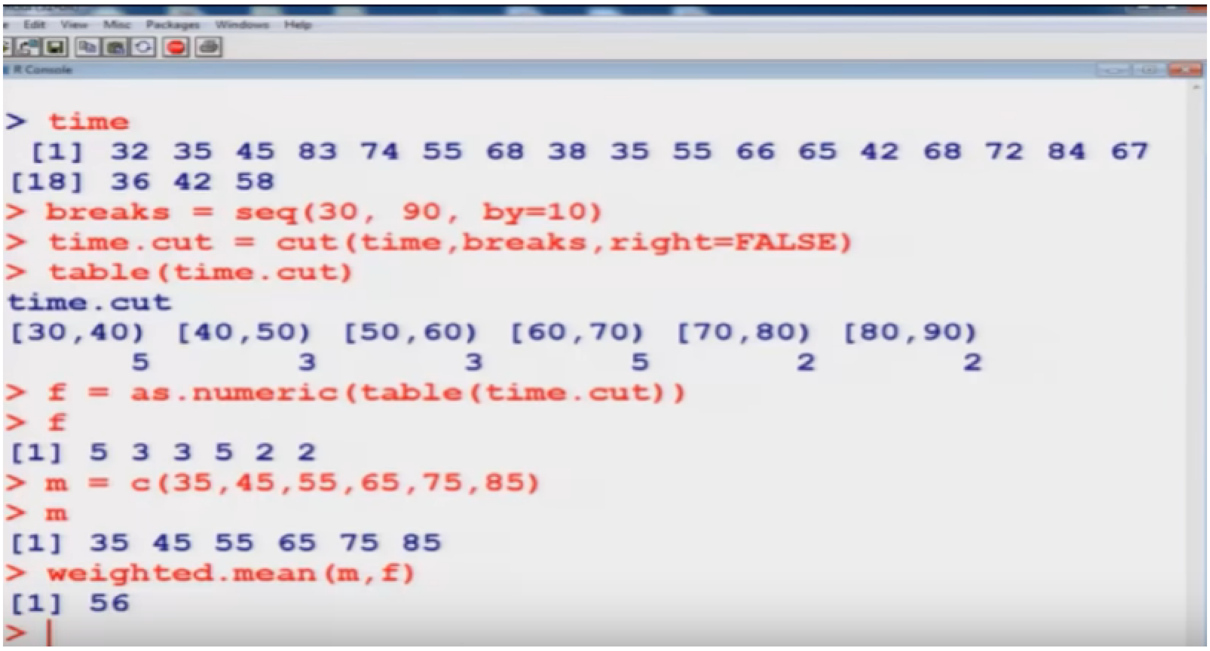
And based on this time dot cut data, we had found the frequency table of the data, in time vector by using, table inside the argument time dot cut and this was the frequency distribution that we had obtained.Right? Now, what we have to do? We need to find out the weighted arithmetic mean or the arithmetic mean for this group data. So, this can be done by the following ways, first step is that, we need to extract the frequencies, from this frequency table. So this is, here the frequency table and we want to extract only this data vector 5 3, 3 5, 2 2,because this is the value of F 1, this is the value of F 2, this is the value of F 3 ,this is the value of F 4,this is f 5 and this is here f6. So, now in order to do it, we try to operate, a command, as dot numeric on this frequency table data.So that will be a s, dot and um er IC or any small, alphabets and inside the argument, I have to give the data vector and the data vector is going to be the outcome of the frequency table.So, in this case, you can see here, your frequency table data, is given by table and inside the argument time dot cut. So,I try to operate this command over here,as dot numeric and inside the alchemists table, time dot cut and this gives me here, this outcome. So, you can see here:that this data is the same data that you have obtained, say this 5 is the same,as here this 5,this 3 is the same here, this 3 and this 3, this is the same here, this 3 and this 5,this is here the same 5 and then this 2,this is here the same 2 and this 2, this is here the same 2. Right?

Refer Slide Time :(28: 16)



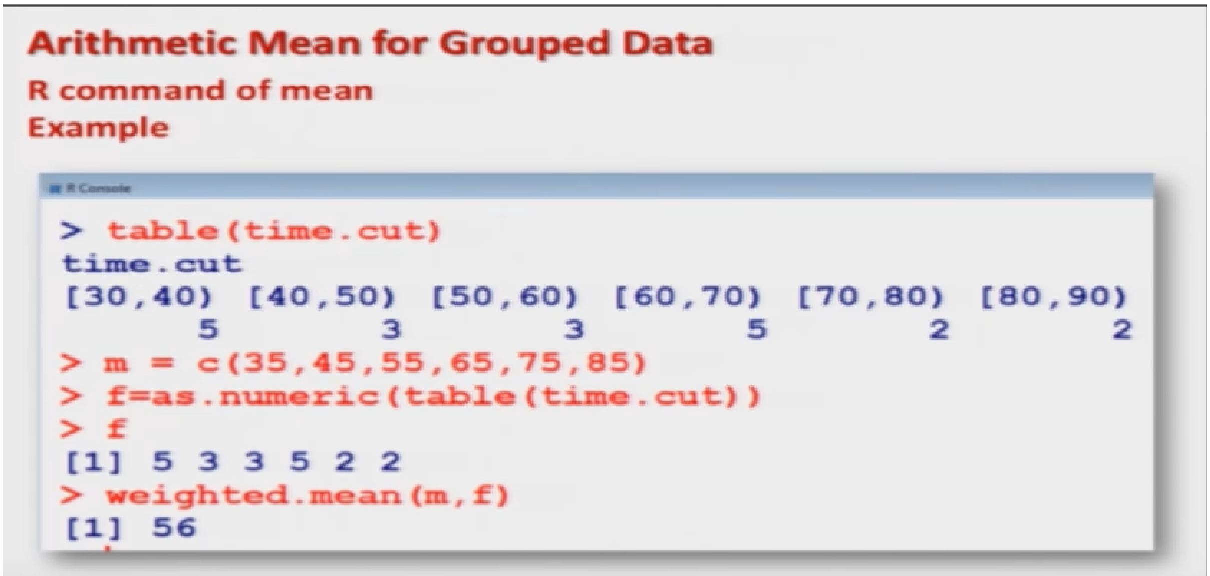
So now, once we have obtained, this here, vector here F,which is the vector of the frequencies,similarly we can find out the vector of the data, on midpoints and then, I simply have to use here the command, weighted dot mean and with M and F and this will come out to be 56, right. Okay? So, let me now, first try to operate, this thing on the R console to show you.

Refer Slide Time :(28: 45)



So now, if you try to see, we already had entered the data on time, which is here, now I first need to create a frequency table.So, I will simply try to, copy and paste the commands: that we had used earlier. So, this is about the breaks and then, I will try to execute the command, to get the data, time dot cut and then I will try to find out the frequency table, using the time dot cut and you can see here,this is the same letter set. And now, I will try to extract here, the data, from this table, using the command as numeric.So, you can see here,this data is shall had the same, if you try to see here, this line, which I am highlighting and this line, which I'm highlighting, they are the same, right. And now I need to define here the, the vectors of midpoints. So, this is here M,so you can see is here M and now, once you try to use here, the function or the command, weighted mean, this weighted mean come out to be 56. So, this is how you can obtain this weighted mean, in case of this group data.

Refer Slide Time :(30: 16)



And now here is, the screenshot of the same operation that we have just done. Now, I would like to stop here in this lecture and if you try to see, what we have done? We have simply learned, the concept of arithmetic mean and we have learnt how to execute, it on the R software. And this arithmetic mean, is found for the group data, for theungrouped data. So that is pretty simple,but it is more important, to learn that,what are the different other parametersthat can be used, in the command mean:that can be looked, through the help menu.So, I would request all of you: that you try to take, a small data set, say only few, numbers and try to compute the arithmetic mean by your hand, manually and try to compute it, using the R software.And once you see that, both the things are matching, then it will give you more confidence: that yes, the software is also doing the same thing, what we wanted to do? So, you practice and we will see you in the next lecture. Till then, Goodbye.