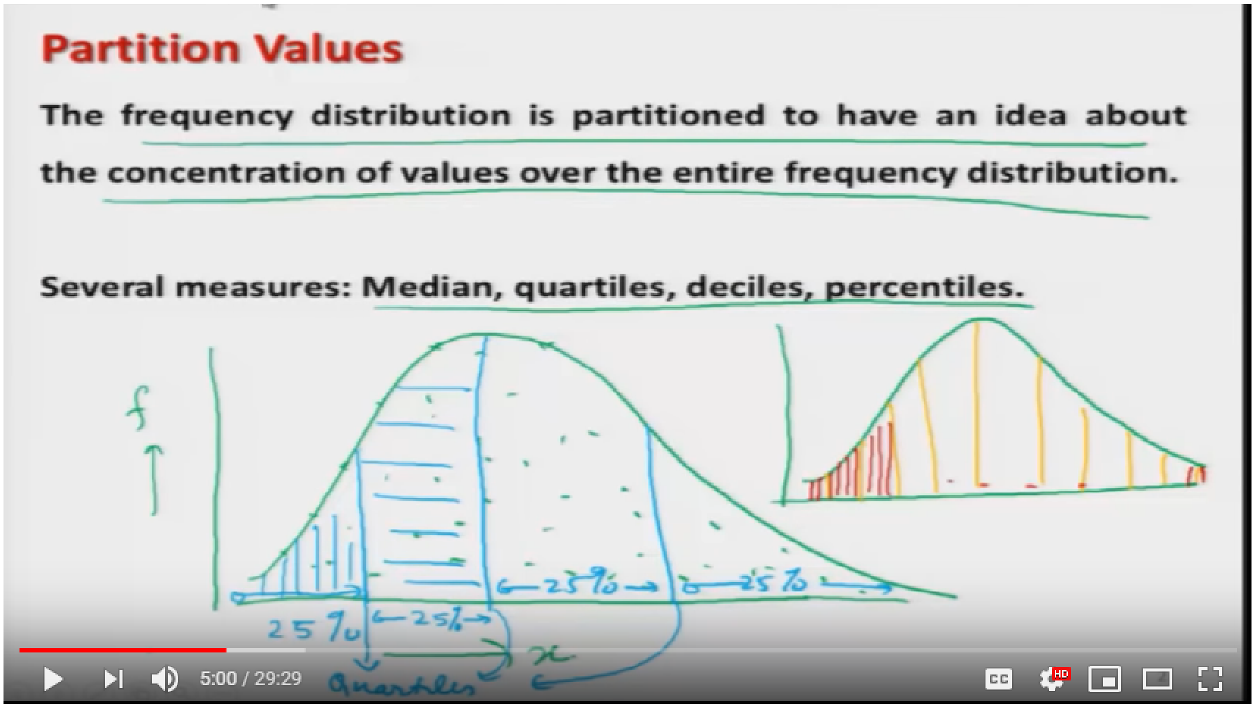
Lecture – 15
Central Tendency of Data - Median

Welcome to the next lecture on the, descriptive status it with R Software. You may recall that in the earlier lecture, we had discussed the idea of Central Tendency. And we had planned that, we will discuss, several measures of central tendency of the data. In the last lecture, we had explained the concept of arithmetic mean. Now, in this lecture, I am going to consider the aspect of partitioning values. And under that, topic I will try to consider the median. Okay? So, let us try to first understand what are these partition values.

Refer slide time :(1:02)



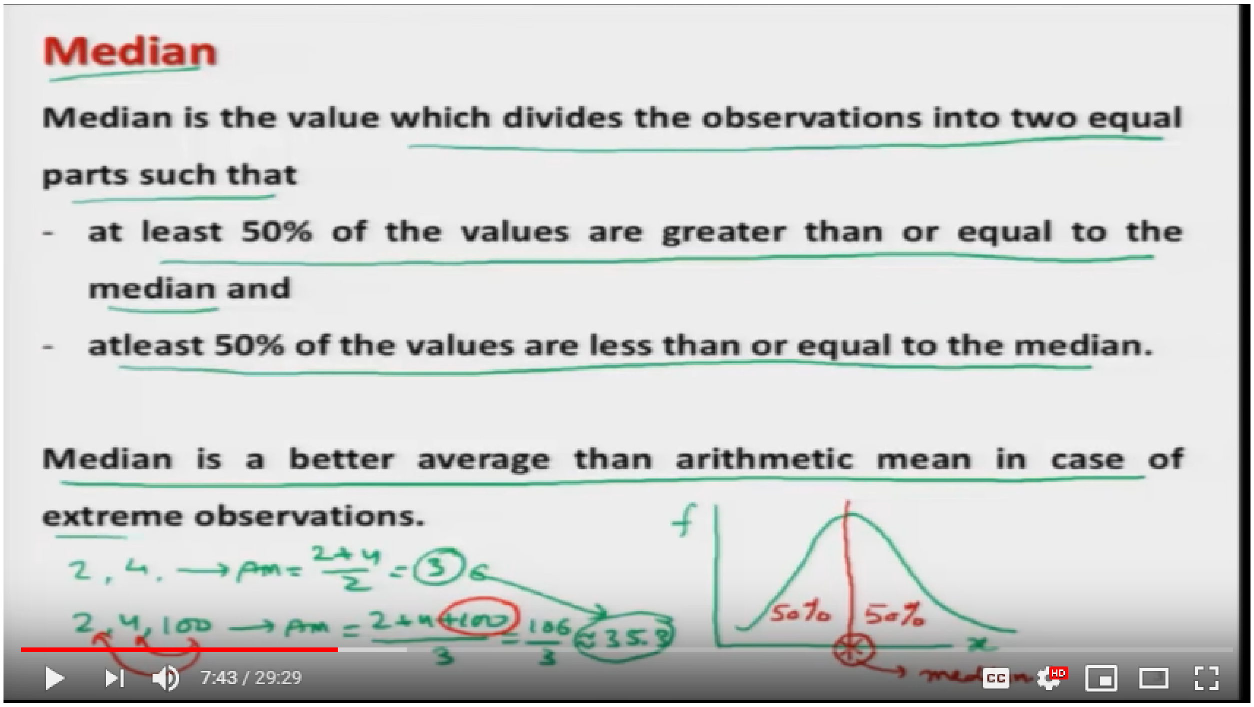
Partition Values

The frequency distribution is partitioned to have an idea about the concentration of values over the entire frequency distribution.

Several measures: Median, quartiles, deciles, percentiles.

If you try to see, means if I try to create here the frequency distribution means, on the x-axis we have values, say class width or the X values and on the y-axis we have frequency values and suppose we have got a, frequency distribution like, this one. Now, you can see here that the entire frequency is, being covered under this curve, you can see here, these are the frequency values and these are the different values here, of the frequency on the curve. Now, we would like to know that how these values are going to be partitioned, for example, if I say, suppose I want to divide this in four equal parts, so I can make here first, say here second, say here third and here fourth. So, you can see here, from here to here, this is indicating the area, which is containing nearly the 25% of the total frequency. And similarly from here to here, this is an area, which is trying to cover another 25% of the frequency. And similarly, this is also 25 percent of the total frequency and this is also 25 percent of the total frequency. So, these values here, they are called as, suppose here, partitioning values which are trying to divide, the total frequency into four equal parts, so I can call it here, as a quad tiles. So, this is quartile this is quartile value. And yeah, I mean, so the first value can be called as, ‘First Quartiles’. Second value can be called a, ‘Second Quartile’ and so on. So, essentially what is happening that we are trying to divide the entire frequency into ten parts? And similarly, in case if you want to define, divide it into more parts, for example, if this is the frequency curve, then possibly I can divide it into ten parts, one, two, three, four, five, six, seven, eight, nine, ten and so on. And similarly, we miss I can partition it in system other way also, for example, in case if I want to partition it into say hundred equal parts, 1, 2, 3, 4, 5, 6, 7, 8 and dot, dot, up to here, there will be hundred such partitions. And it is also not necessary that this partition, have to be of the same length, they can be of different learn, lengths. So, by looking at the partitioning value, we can have an idea that, how the frequency is distributed, over the entire range of the frequency distribution. And this will also give us an idea that, how the frequency is concentrated in different regions of the frequency curve. So, we will try to take up all these partition values, one by one, we will try to understand them and we will try to see, how to compute them on the R software. Now, I can say very simply that the frequency distribution is partitioned, to have an idea, about the concentration of the values, over the entire frequency distribution. And as I said there are several measures: median, quartile, deciles, percentile. So, you kind of try to start our discussion with median.

Refer slide time :(5:05)



Median

Median is the value which divides the observations into two equal parts such that

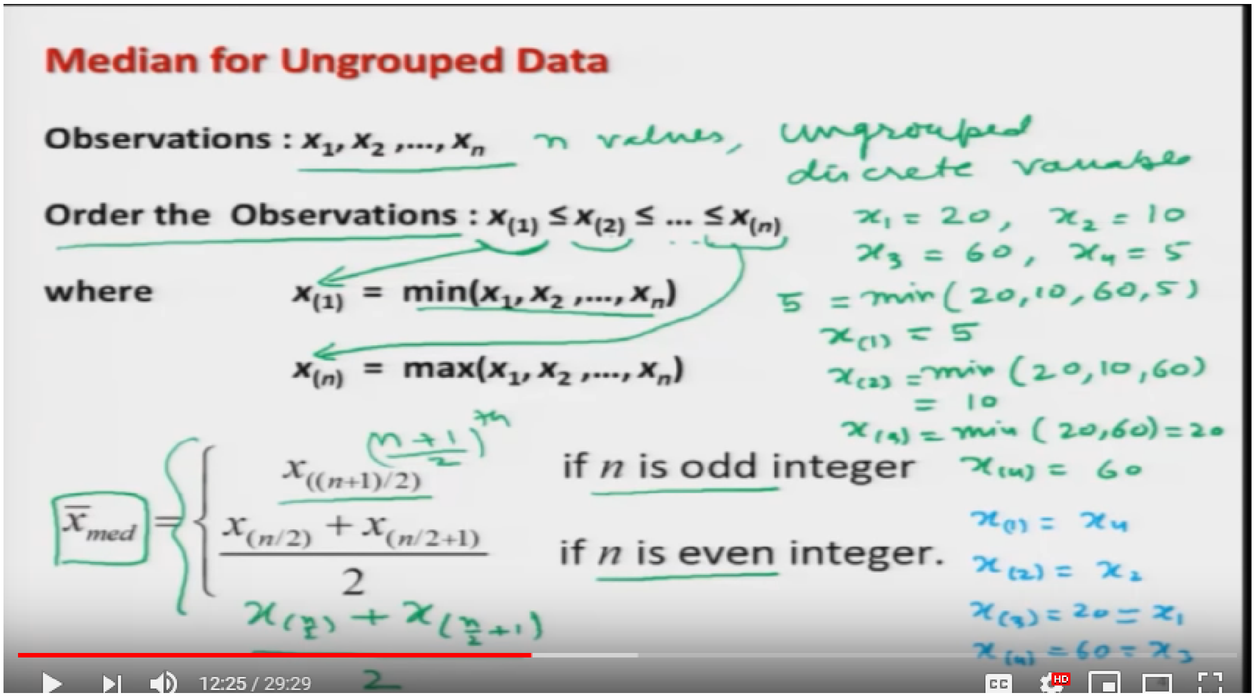
- at least 50% of the values are greater than or equal to the median and
- atleast 50% of the values are less than or equal to the median.

Median is a better average than arithmetic mean in case of extreme observations.

2, 4. → AM = $\frac{2+4}{2}$ = 3
2, 4, 100 → AM = $\frac{2+4+100}{3}$ = $\frac{106}{3}$ ≈ 35.3
f
50%
50%
x
median
7:43 / 29:29
CC
HD

Now, going through with that definition, suppose if I try to plot here, the frequency curve like this. And suppose I say, I would like to divide the entire frequency into two parts, two equal parts, such that, 50% of frequencies on the left side of this red vertical line and 50% of the frequency is on the P right-hand side of this red vertical line. So, now corresponding to which here is this value, this is called as, 'Median'. So, median is the value, which divides the observations into two equal parts, such that, at least 50% of the values are greater than or equal to the median and at least 50% of the values are less than or equal to the median. So, median is a measure, which is trying to divide the total frequency into two parts. So, if I say that the median of my frequency distribution is suppose say is 20. So, I can say here that, 50% values are less than 20 and there are 50% values which are more than 20. Okay? Now, if you try to compare median with arithmetic mean, then in all those situations, where we have got extreme observation that means, some observation which is taking very, very high value, then in those cases this median is preferred, why because if you try to see, suppose if I try to take here, two values. Two and four and then I try to find out it's automatic mean, automatic mean is going to be two plus four divided by two is equal to three. But, if I try to adhere to four and here 100, then this value becomes here, the automatic mean is equal to two plus four plus hundred divided by here three and this is hundred six by three and this is closely equal to thirty five point three. So, you can see here, there is a huge difference between three and thirty five point three and this difference is coming because, there is a new value, which is added here hundred and this hundred is very much different from two and four ,there is a huge difference between the two values. So, this medium is a better average, than automatic mean in case, if we have extreme observations. Right?

Refer slide time :(7:47)



Median for Ungrouped Data

Observations : $x_1, x_2, \ldots, x_n$ n values, ungrouped discrete variables

Order the Observations : $x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}$

where $x_{(1)} = \min(x_1, x_2, \ldots, x_n)$

$x_{(n)} = \max(x_1, x_2, \ldots, x_n)$

$$\boxed{\bar{x}_{med}} = \begin{cases} x_{((n+1)/2)} \quad \left(\frac{n+1}{2}\right)^{th} & \text{if } n \text{ is odd integer} \\ \dfrac{x_{(n/2)} + x_{(n/2+1)}}{2} & \text{if } n \text{ is even integer.} \\ \dfrac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2} & \end{cases}$$

$x_1 = 20, \quad x_2 = 10$

$x_3 = 60, \quad x_4 = 5$

$5 = \min(20, 10, 60, 5)$

$x_{(1)} = 5$

$x_{(2)} = \min(20, 10, 60)$

$= 10$

$x_{(3)} = \min(20, 60) = 20$

$x_{(4)} = 60$

$x_{(1)} = x_4$

$x_{(2)} = x_2$

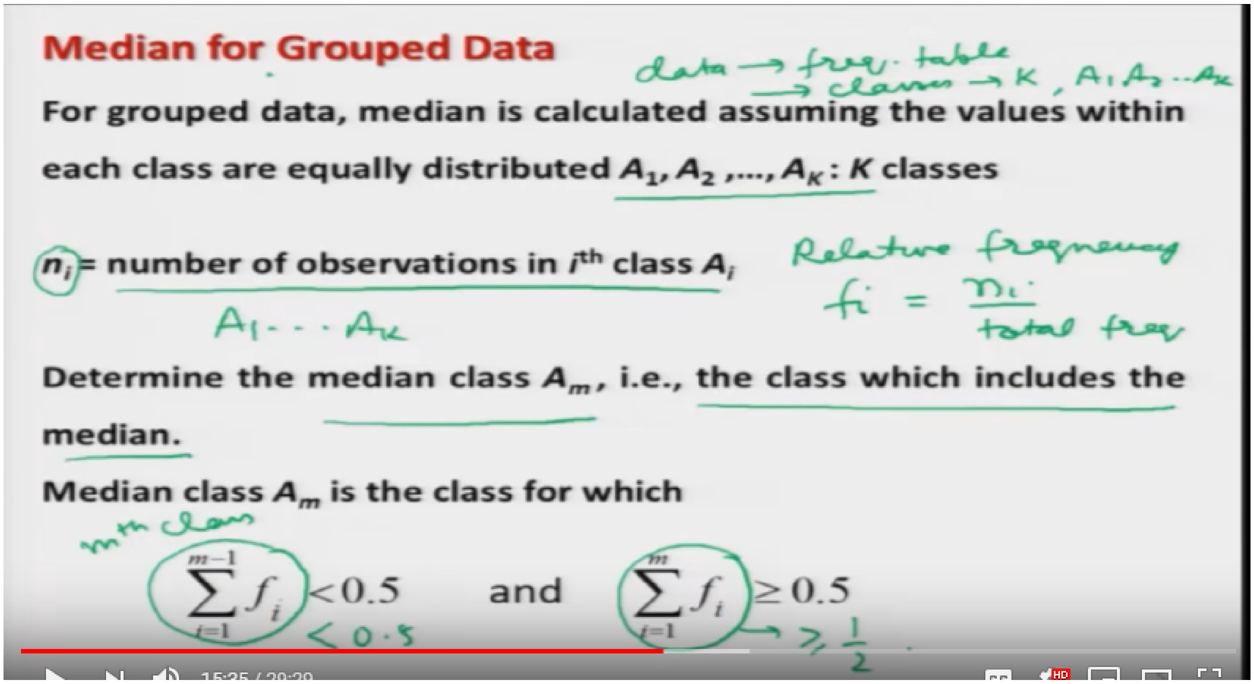
$x_{(3)} = 20 = x_1$

$x_{(4)} = 60 = x_3$

12:25 / 29:29

Now, I would try to give the definition and how to compute the median, in two cases. One is ungroup data and other is group data. So, first we try to understand the median and its computations, when we have a data that is ungrouped. So, we kind of try to say, we have observations X 1, X 2, X n, so there are n values and they are ungrouped. Ungrouped or you can call as, they are the values of some discrete variable. Now, what I do? I try to order the observation and I present the ordered values as, X and in the subscript inside the bracket, I am writing here one and the second ordered value will be, the value of x, inside the bracket, I am writing here two and so on. What does this mean? This means that, this value x 1 is the smallest value, this is the minimum value among X 1, X 2, xn and this X inside the bracket n, this is the highest value or the maximum value among X 1, X 2, X n. What does this mean, suppose if I try to take here 4 observation, say here, X 1 is equal to say here 20. X 2 is equal to here, see here, 10. X 3 is equal to here, 60 and X 4 is equal to 5. Now, these are the four values. So, I try to find out here the, the minimum value among, twenty, ten, sixteen and here five. So, this is here, five. So, the first ordered value, which I will denote as say, X and one inside the bracket in the subscript, this becomes here five. And after that, once again I try to find out the minimum value, among the remaining value, which is twenty, ten and 60. So, this gives me the second ordered value, which is equal to here, now you can see here this is a here ten. And similarly, X 3 is the minimum value among the remaining values, with between 20 and 60 and this is equal to 20. And obviously then the largest value is here, 60. So you can see here, the difference between, the simple observations and the ordered observations. So you can see here, what is the relationship? The relationship is this first order value X 1; this is same as you’re here, fourth unordered value. Similarly here, second ordered value is the same as, the second unordered value. And third ordered value is here, 20 which is the same as, X 1; this is the first unordered value. And fourth ordered value is 60, which is the same as, third unordered value. So, you can see here that how the ordered and unordered values are in two related. Okay? So now, the first step in finding the median of an ungroup data is to order it first. And once you order it, then there are two situation that, the number of observations, they can be odd or the number of observation, can be even. So, now in case if the number of observations are odd, then the median is going to be the, n plus 1 by 2 Th, ordered value. Which is here like this and in case if even that means the number of observations is, even then the median is going to be the average off and by 2 th ordered observation and n by 2 plus 1 th ordered observation like this. So, this gives you here the definition of the median, so you simply have to see that, whatever is the appropriate ordered value according to this rule that will give you the median. Now, we consider the median for the group data, so we know that whenever we have a group data or the data on any continuous variable, then the first step is that, we try to create the frequency table, then in frequency table, we will have classes.

Refer slide time :(12:58)



Median for Grouped Data

data → freq. table
→ classes → K, $A_1 A_2 \ldots A_K$

For grouped data, median is calculated assuming the values within each class are equally distributed $A_1, A_2, \ldots, A_K$: K classes

n_i = number of observations in i^{th} class A_i

$A_1 \ldots A_K$

Relative frequency

$f_i = \frac{n_i}{\text{total freq}}$

Determine the median class A_m, i.e., the class which includes the median.

Median class A_m is the class for which

m^{th} class

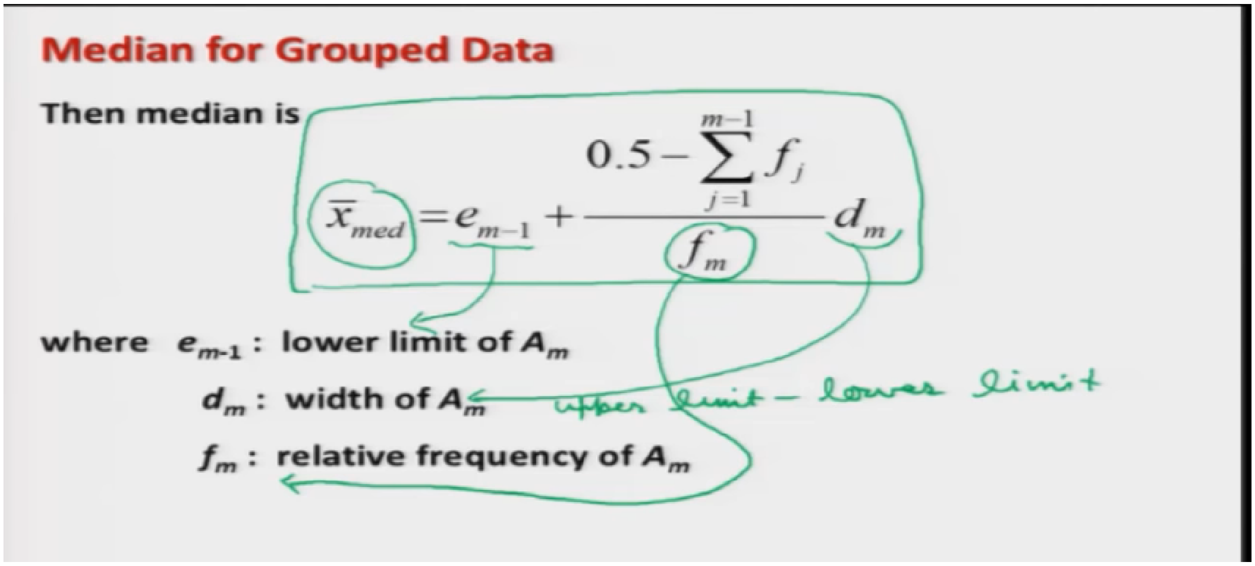
$$\sum_{i=1}^{m-1} f_i < 0.5 \quad \text{and} \quad \sum_{i=1}^{m} f_i \geq 0.5$$

< 0.5

$\geq \frac{1}{2}$

So now here, we start our discussion by assuming that we had that data and from the data, we have created the frequency table. And this frequency table has classes and there are K classes, denoted as a 1 a 2 see here, a K. So now, the entire frequency is distributed, equally among K classes and we assume here that, the absolute frequency of the I th class is n I. So, there are a number of observations in the I th Class A i. From this absolute frequency, I can compute the relative frequency. And we are denoting the relative frequency F I say here, Ni upon, say here, total frequency. One thing I would like to mention here and I would like to draw your attention that please notice the definition and symbol of fi, in the case of median I am trying to denote, fi the relative frequency. But, later on when we are trying to consider other type of measure, there is a possibility that I may define, this fi to be the absolute frequency, so be watchful. Now, after this what we have to do? Now, I have got here classes a 1, a 2, a K. And we know by definition, the median is the value where the total frequency is going to be divided into two equal parts, so there will be a class, where that half of the frequency will be lying. So, I would try to find out here, the class where half of the frequency is lying. And let this class be denoted by here, as see here Am. So, Am the interval or the class, which includes the median. So, I can define this median class, as the omit class where, in case if I try to sum all the frequencies, from one to M minus one and sum of all the frequencies one to M, then this sum is going to be smaller than half and this summation from I goes from 1 to M, this is going to be greater than or equal to half. So, this is going to be my median class.

Refer slide time :(15:40)



Median for Grouped Data

Then median is

$$\boxed{\bar{x}_{med} = e_{m-1} + \frac{0.5 - \sum_{j=1}^{m-1} f_j}{f_m} d_m}$$

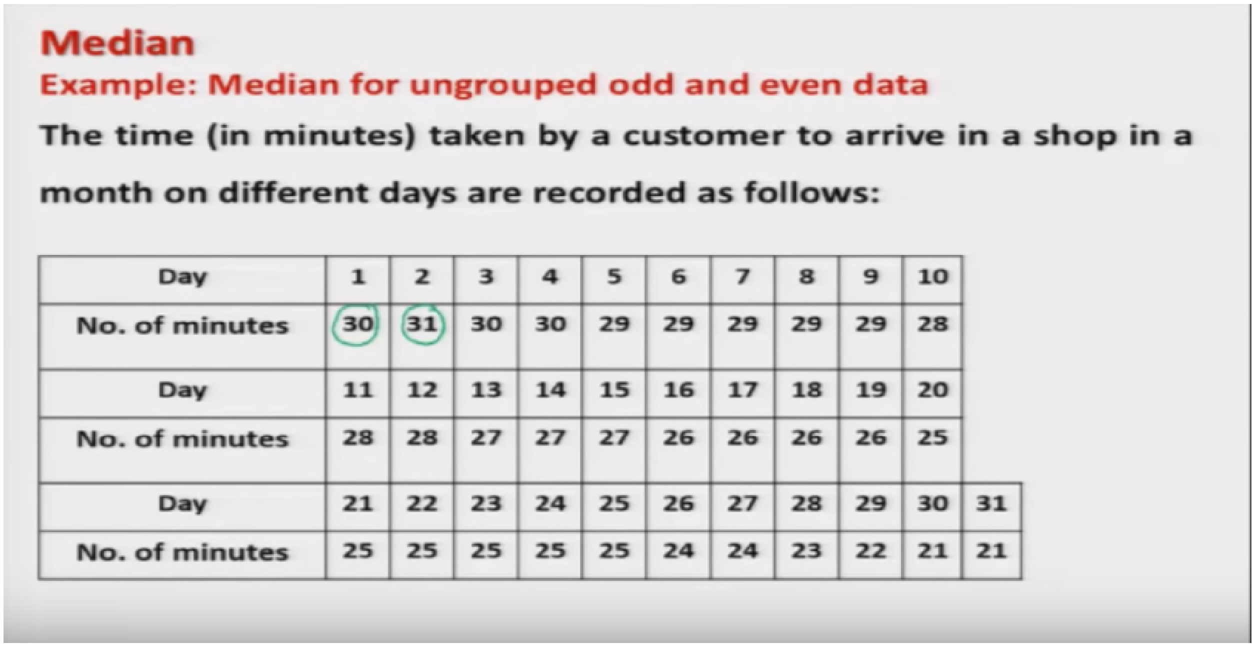
where e_{m-1} : lower limit of A_m

d_m : width of A_m upper limit – lower limit

f_m : relative frequency of A_m

Now, the expression for finding out the median, in case of group data is given by this. Here, you can see, there is a quantity here E m minus 1, this is denoting the lower limit of the median class. And similarly here, there is a quantity DM this is going to denote the, width of the median class, width means, upper limit minus lower limit. So, this is the class width, then there is a relative frequency here FM, which is going to be the relative frequency of the median class. And based on this, we try to compute the median of the given, grouped data and we denote it here say x-bar med, the short form of median. Now, let us try to take a example and try to see how to get it done. Here, I would like to, inform you that when we try to compute the median on the R Software, then at least to the best of my knowledge, there is only one command, which is available inside the RISC package, to compute the median. So, this R Software has no separate commands, for computing the median for group and ungroup data. So now, through this example, I will try to show you that how these different values, like as relative frequency of the median class and so on. How they are chosen and then, I will try to show you that, how the median is computed on the R Software. But, then I will not be able to show you that, how to specifically compute the median for the group data, well one can write a small program or a small function to compute, such thing but at least I will not be handling it here.

Refer slide time :(17:57)



Median

Example: Median for ungrouped odd and even data

The time (in minutes) taken by a customer to arrive in a shop in a month on different days are recorded as follows:

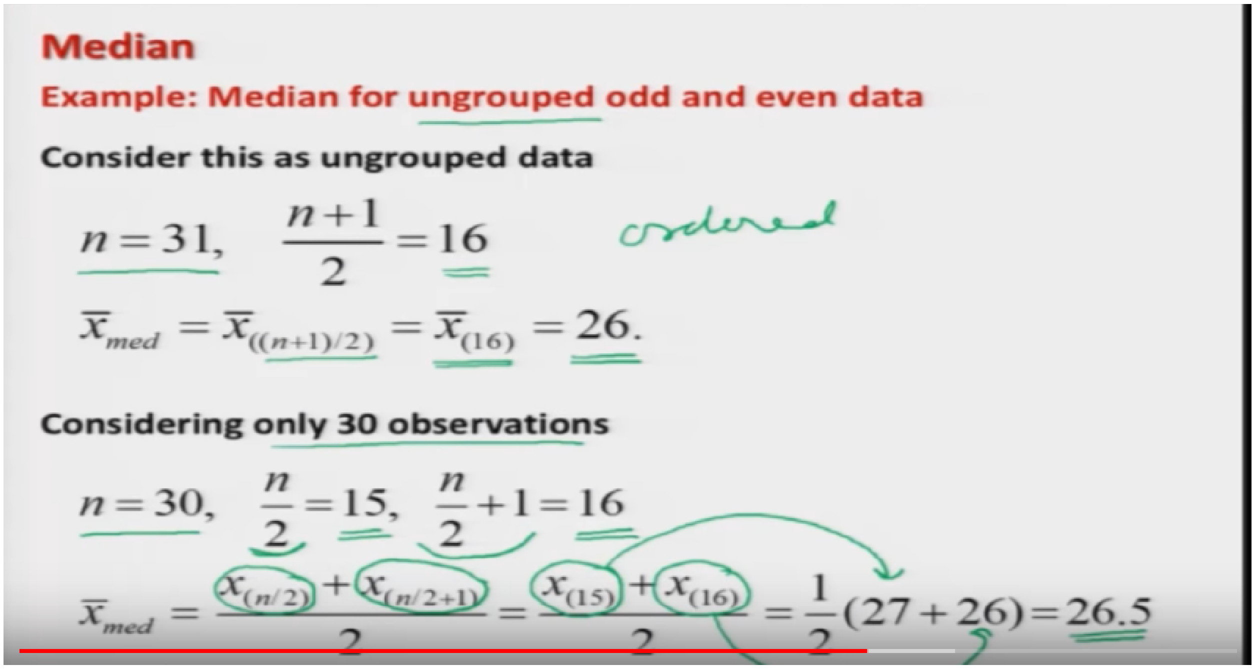
Day	1	2	3	4	5	6	7	8	9	10
No. of minutes	30	31	30	30	29	29	29	29	29	28

Day	11	12	13	14	15	16	17	18	19	20
No. of minutes	28	28	27	27	27	26	26	26	26	25

Day	21	22	23	24	25	26	27	28	29	30	31
No. of minutes	25	25	25	25	25	24	24	23	22	21	21

So, you can try to consider this example, in which the data is collected, on the time taken by a customer to arrive in a shop, in our insider’s shopping mall and this time is recorded on different days, of the month. So, assuming that, there are 31 days in the month, this data has been collected for example, on the day one he takes 30 minutes, on the day to the customer take 31 minutes and so on. So now, I will try to find out the median, from this data first considering it as, ungrouped data and then I will try to group it and then once again I will try to find out the median. Okay?

Refer slide time :(18:43)



Median

Example: Median for ungrouped odd and even data

Consider this as ungrouped data

$$\underline{n=31}, \quad \frac{n+1}{2}=\underline{16}$$

ordered

$$\bar{x}_{med}=\bar{x}_{\underline{((n+1)/2)}}=\underline{\bar{x}_{(16)}}=\underline{26}.$$

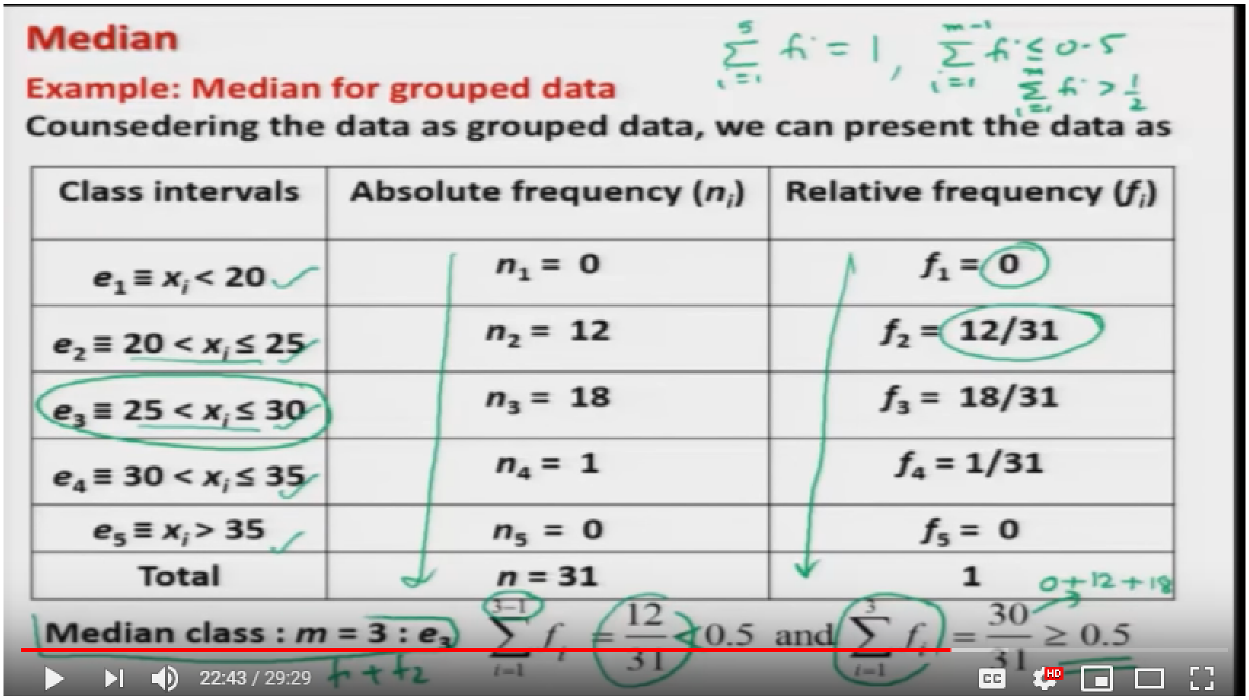
Considering only 30 observations

$$\underline{n=30}, \quad \frac{n}{2}=\underline{15}, \quad \frac{n}{2}+1=\underline{16}$$

$$\bar{x}_{med}=\frac{x_{(n/2)}+x_{(n/2+1)}}{2}=\frac{x_{(15)}+x_{(16)}}{2}=\frac{1}{2}(27+26)=\underline{26.5}$$

So, now here in this case, when I try to consider this data, as an ungrouped data. So, the number of observations here is 31, so n is equal to 31, so the value of n plus 1 is here, Two this is equal to 16. So, now what we have done that, we have ordered the data, this data has been ordered ,well I'm not showing you here that you can do and then I am trying to find out the, n plus 1 by 2 th ordered value and this is the 16 the value in the ordered data. And this value comes out to be here 26, so 26 minutes is the median time. And now, in case if I try to convert the same data into n, even number of observations, so I can drop the last observation and I try to consider only here the, 30 observation ,so in that case, the number of observation becomes a 30 and then, n by 2 here is 15 and n by 2 plus 1 is here 16. Now, according to the definition of the median, the median is going to be the, mean off and by 2 th ordered value. And, and by 2 plus 1 th ordered value. So, essentially this is going to be the automatic mean of the fifteenth ordered value and sixteenth order value and from the data, we find that the fifteenth ordered value is 27 and 16th ordered value is 26. So, this median comes out to be twenty six point five. So, this is how we compute the median in case of ungrouped data, considering the data to be or an even numbers.

Refer slide time :(20:35)



Median

Example: Median for grouped data

Counsedering the data as grouped data, we can present the data as

$\sum_{i=1}^{5} f_i = 1$, $\sum_{i=1}^{m-1} f_i \le 0.5$ $\sum_{i=1}^{m} f_i > \frac{1}{2}$

Class intervals	Absolute frequency (n_i)	Relative frequency (f_i)
$e_1 \equiv x_i < 20$	$n_1 = 0$	$f_1 = 0$
$e_2 \equiv 20 < x_i \le 25$	$n_2 = 12$	$f_2 = 12/31$
$e_3 \equiv 25 < x_i \le 30$	$n_3 = 18$	$f_3 = 18/31$
$e_4 \equiv 30 < x_i \le 35$	$n_4 = 1$	$f_4 = 1/31$
$e_5 \equiv x_i > 35$	$n_5 = 0$	$f_5 = 0$
Total	$n = 31$	1

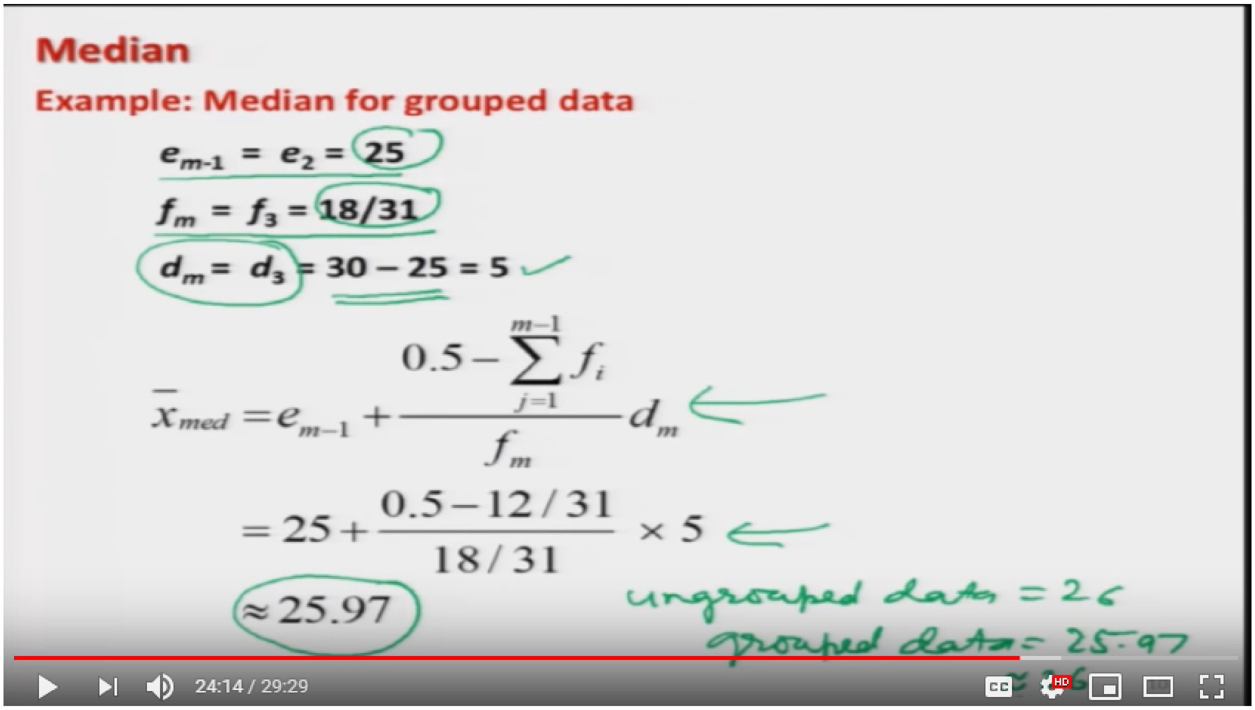
0+12+18

Median class : $m = 3 : e_3$ $\sum_{i=1}^{3-1} f_i = \frac{12}{31} < 0.5$ and $\sum_{i=1}^{3} f_i = \frac{30}{31} \ge 0.5$

$f_1 + f_2$

And similarly in case if I try to consider this data as a group data, then you try to create here the frequency table, so you can see here, I have already created the frequency table here, these are my class intervals, of the width five units, like as 22, to 25, 25 to 30 and so on. And then in the second column, I have found the absolute frequency and in the third column, I have computed the relative frequencies, of all the classes. Now, you can see the advantage of working with the total relative frequency. Total relative frequency is always going to be, there are five pluses, so I goes from here one two five, this is going to be here one. so, I simply need to find out here, what is the class say here M minus one, fi which is smaller than 0.5 and for what value of here M, the sum of frequency is greater than half. So, once I try to do it here, I observe that there is a third class, this is E3for which the sum of the relative frequency of class 1, class 2 and class 3, this comes out to be 12 on 31. So, this is going to be 3-1, so essentially I am trying to find out, f1 plus, f2. So, you can see here this F 1 is 0 and F 2 here is, 12 on 31. So, this comes out to be smaller than 1/2 and if I try to find out the sum of F 1, F 2 and F 3, this comes out to be 30 upon 31. How this 30 is coming into picture? This is coming out to be, the absolute frequency of class one, this is zero plus absolute frequency of class two this is 12 and absolute frequency of class three, this is 18. So, this is essentially 12 plus 18, which is equal to 30 and this comes out to be greater than half. So now, I can say here, my median class is third class E3 and so here, M is equal to 3.

Refer slide time :(22:47)



Median

Example: Median for grouped data

$$\underline{e_{m-1} = e_2 = \boxed{25}}$$

$$\underline{f_m = f_3 = \boxed{18/31}}$$

$$\boxed{d_m = d_3} = \underline{30 - 25} = 5 \checkmark$$

$$\bar{x}_{med} = e_{m-1} + \frac{0.5 - \sum_{j=1}^{m-1} f_i}{f_m} d_m \leftarrow$$

$$= 25 + \frac{0.5 - 12/31}{18/31} \times 5 \leftarrow$$

$$\boxed{\approx 25.97}$$

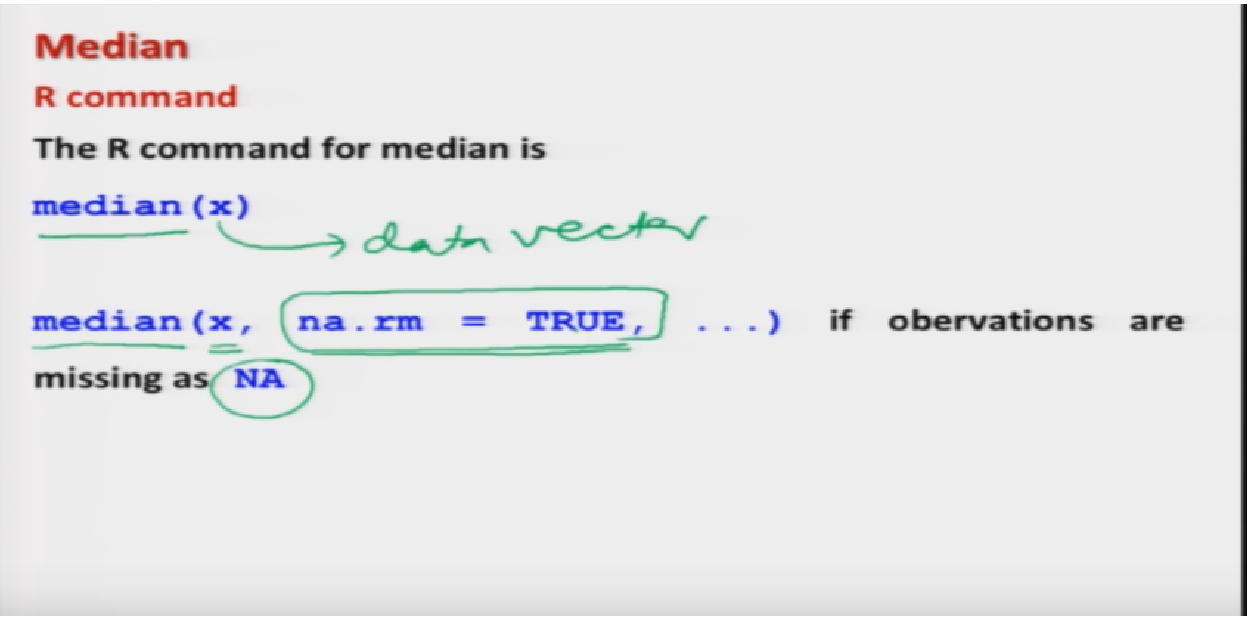
ungrouped data = 26

grouped data = 25.97

≈ 26

Now, I try to find out the lower limit of the median class, which is here 25, the relative frequency of the median class, which is here 18 upon 31. And then, the width of the interval, of the median class, this is here 30 minus 25 it is equal to 5. Now, once I try to substitute all these values over here, in this expression and I try to simplify it, I get here 25.97. So, you can see here, there is not much difference in the value of the median, when we are trying to compute it ,as a group data or say ungroup data . Right? For the ungroup data, you may recall that, this value was coming out to be 26, you can see here. And for the group data, this is coming out to be 25.97. So here, you can see means, if your data is proper then, practically there is no, difference either you try to compute the median, say by this formula or by that formula and possibly this is the reason that R has not implemented it. And then if you try to see this 25.97 is also close to 26. So, for all practical purposes there is not much difference.

Refer slide time :(24:18)



Median

R command

The R command for median is

```
median(x)
```

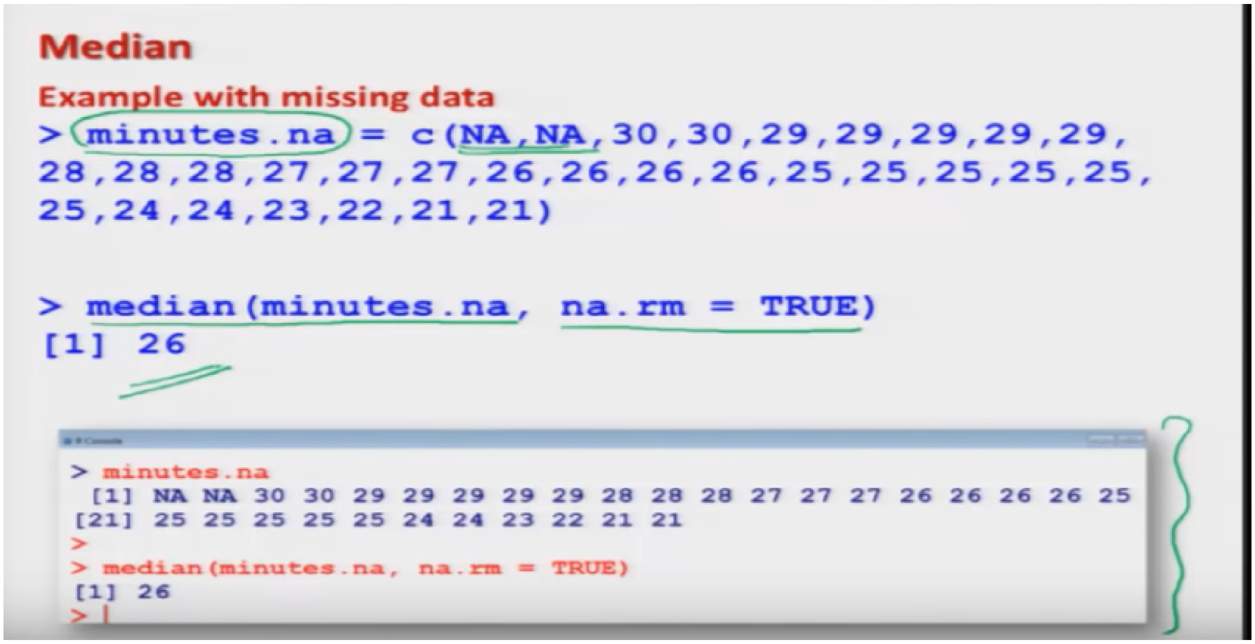
data vector

```
median(x, na.rm = TRUE, ...)
```

if obervations are missing as NA

Now, I try to come on the aspect of R Software. Inside the R Software, to compute the median, the command is MEDIAN and median and this X is my here, data vector. And then in median also there are several option, so but I would once again we trust you that you try to look into the health menu and try to see, what are the different possible parameters that can be given inside the arguments. But, here I would certainly like to address that how would you compute the median, in case if some data is missing and that is represented, as say here NA. So, in that case, use the same command median and give the data vector and use the option here and a dot rm is equal to true. So, this will give you, the value of the median.

Refer slide time :(25:12)



Median

Example with missing data

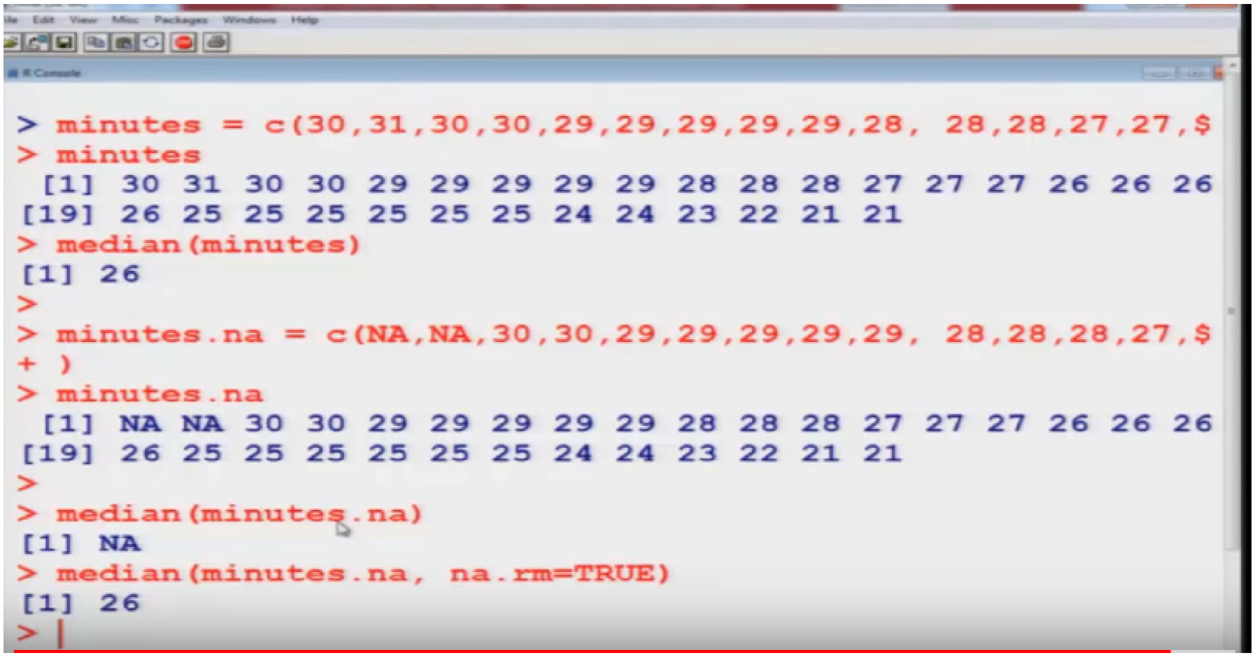
```
> minutes.na = c(NA,NA,30,30,29,29,29,29,29,
28,28,28,27,27,27,26,26,26,26,25,25,25,25,25,
25,24,24,23,22,21,21)

> median(minutes.na, na.rm = TRUE)
[1] 26
```

```
> minutes.na
 [1] NA NA 30 30 29 29 29 29 29 28 28 28 27 27 27 26 26 26 26 25
[21] 25 25 25 25 25 24 24 23 22 21 21
>
> median(minutes.na, na.rm = TRUE)
[1] 26
> |
```

So, I try to now collect all the data on the minutes, inside this data vector here minutes and then I simply try to find out the median of minutes here, this comes out of here, 26. So you can see here, this is matching with the value that you had obtained earlier. And this is the screenshot. So, I would like to also show you here, inside the same data that in case if the data is missing, then how you are going to handle it. So, inside the same data set, I try to make the first two values, to be not available and I replace them by NA. So now, in this case, I try to create or I try to store the data inside a new data vector, minutes dot NA, well I would like to address here, one thing in R there is an option to name the variables using the dot sign or C full stop sign. So that is why minutes dot n/a I am writing ,it is not a say built-in function or there is a rule, it is simply trying to denote, for, just for the sake of convenience that this is the same data of minutes. But, now I am using the missing values. So now, using this data set, I try to find out here the median, using the same command, median mean is dot-na and I'm using here a na dot rm equal to true. And you can see here that this is the screenshot and this value comes out to be again here, 26. So, before I try to do something more, let me try to show you, how to compute these things on the R Software? So, first I try to store this data on the R console,

Refer slide time :(26:54)



```
> minutes = c(30,31,30,30,29,29,29,29,29,28, 28,28,27,27,$
> minutes
 [1] 30 31 30 30 29 29 29 29 29 28 28 28 27 27 27 26 26 26
[19] 26 25 25 25 25 25 25 24 24 23 22 21 21
> median(minutes)
[1] 26
>
> minutes.na = c(NA,NA,30,30,29,29,29,29,29, 28,28,28,27,$
+ )
> minutes.na
 [1] NA NA 30 30 29 29 29 29 29 28 28 28 27 27 27 26 26 26
[19] 26 25 25 25 25 25 25 24 24 23 22 21 21
>
> median(minutes.na)
[1] NA
> median(minutes.na, na.rm=TRUE)
[1] 26
> |
```

see here minutes, so you can see here these are the values of minutes and then I try to find out the median, of your minutes. So, you can see here, this comes out to be like this. And similarly and I try to consider here the, missing values, I tried to once again, store the data inside a new vector median dot n/a . And you can see here, this, this is the data meet minutes dot n/a and you can I try to find out the median of this data. Right? And if you try to see you have not used the option, n a dot R M is equal to true, so that's a very common mistake. So, now let me use it here na dot R M is equal to true. So, you can see here, this value is once again coming out to be 26. Now, I would like to stop here, I have given you a detailed, overview of median, how to compute it, what is the concept and how to compute it in R. And you please try to practice it, take some data and try to calculate the median manually and then, try to do the same thing, with the software. And try to see, what is the difference usually I expect that unless and until the data is extremely high true genius, this difference will be very, very small. And for all practical purposes, the value of the median that you, compute from the R command either for the group data or the ungroup data, they will not differ much. So, for all practical purposes, you can accept them. So you practice and I will see you in the next lecture. Till then, Goodbye.