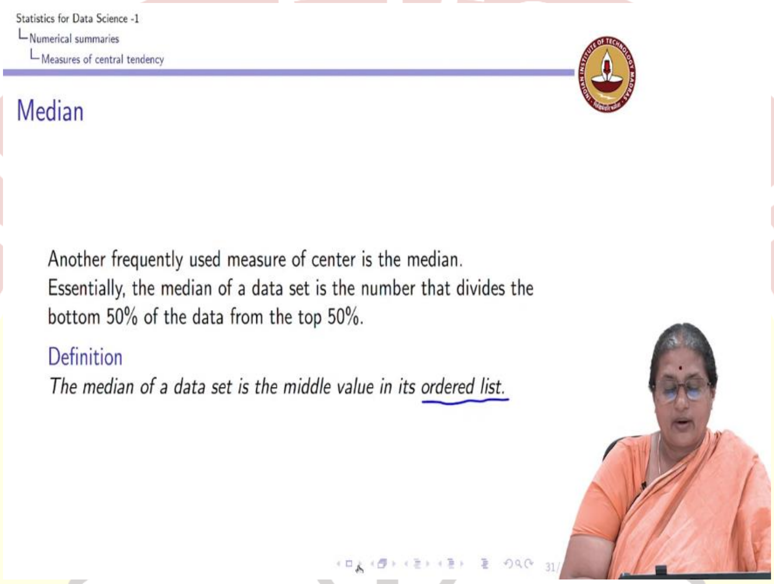
IIT Madras

ONLINE DEGREE

Statistics for Data Science - 1
Prof. Usha Mohan
Department of Management Studies
Indian Institute of Technology, Madras

Lecture – 3.3
Describing Numerical Data – Median and Mode

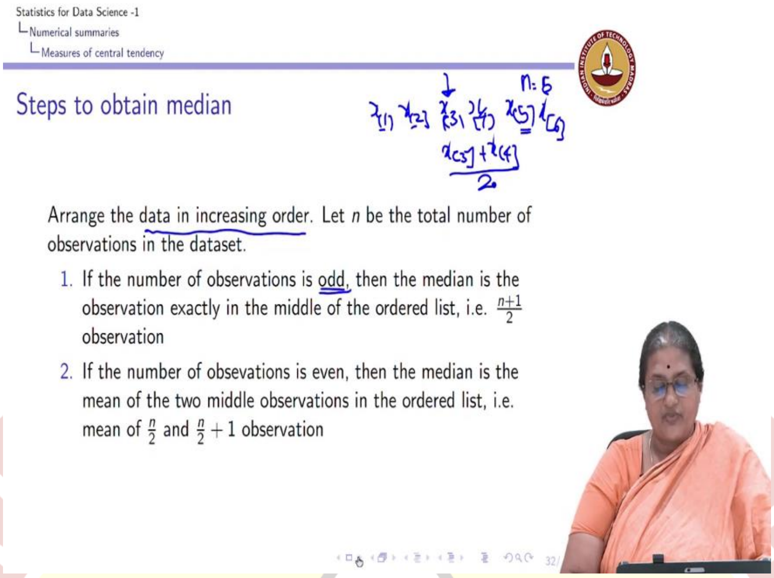
(Refer Slide Time: 00:15)



This is what we another popularly or frequently used measure is what we refer to as a median. Essentially the median of a data set is the number that divides the data set into a bottom 50% and a top 50%.

The minute we say top 50% and a bottom 50%, we quickly realize that we are talking about an ordered data set. So, I formally define a median of a data set is the middle value in the ordered list. The ordered list is extremely important here.

(Refer Slide Time: 00:57)

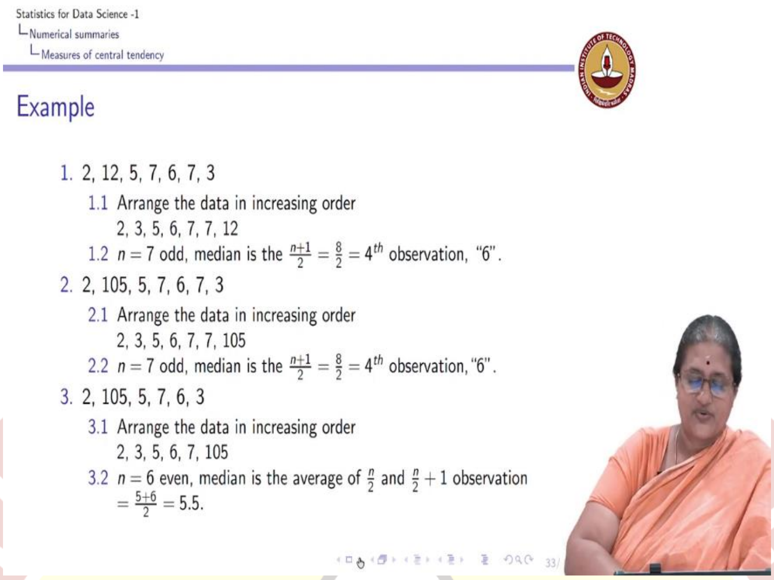


So, now how do I compute the median of a data set? The computing data, so I have an ordered list. So, I arrange the data in increasing order. Let n denote the number of observations in the data set. Now, this is important, if the number of observations is odd, then the median of is exactly in the middle of the ordered list.

For example, if my observations $x_1, x_2, x_3, x_4, x_5, 5$; n equal to $5, 5$ is odd, then this is assume it is ordered. So, for order data, let me introduce a notation x_1, x_2, x_3, x_4, and x_5, x_1 is the order that is the first data, x_2 is the second, x_3, this is my ordered data, then the data in the $\frac{n+1}{2}$, $\frac{n+1}{2}$is $5+1$, $\frac{6}{2}$, third. So, this would be my median. Remember x_1, x_2, x_3, x_4, x_5 are is my data arranged in increasing order. And x_3 which is the third rank data would be my median.

If the number of observations is even, for example, I have x_6 also in this case, my n equal to 6, then the median is going to be your x_3 that is my ($\frac{n}{2}$ observation) $+(\frac{n}{2}+1)$, $x_4$divided by 2, that is what this means. So, if at the median depends on whether the number of observation is an odd number or a even number.

(Refer Slide Time: 03:03)



So, now let us apply this definition and steps to compute the median for the data sets we have already seen before. So, when I have this data $2, 12, 5, 7, 6$ and 3, the first step says arrange the data in increasing order. So, the arrangement of data is going to be $2, 3, 2, 3, 5, 6, 7, 7,$ and 12. I have arranged my data in ascending order.

Now, what is my n in this case? I have $1, 2, 3, 4, 5, 6, 7$: n equal to 7 which is odd. So, if n equal to odd, then I apply my, first n equal to odd, the median is $\frac{n+1}{2}$. So, I have n equal to 7 which is odd. So, median is 8 by 2 which is the $4th$ observation which is equal to 6. So, my median of this data set is 6.

Now, let us look at another example, again the same example. Remember when we are looking at the same examples which we computed the mean for. The difference between the second data set and the first data set is in only one observation which is the second observation here which is 105 for the second data set, and 12 for the first data set.

Remember when we computed the mean, we saw that this one observation actually influenced the mean, and the mean of the first data set and the second data set were very different from each other. Now, let us see what happens to the median of these two data sets.

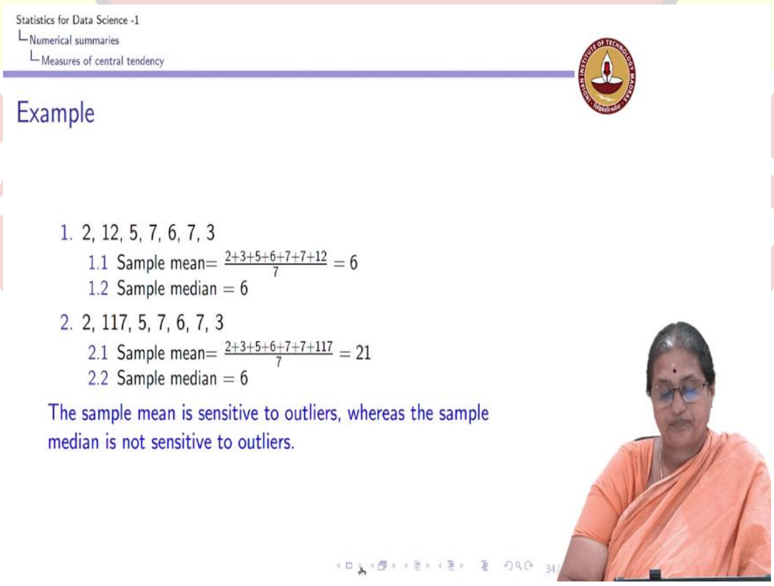
The number of observations again is the same. I arrange the data in ascending order. So, when I arrange this data in ascending order, I have 2, I have 3, I have $5, 6, 7, 7,$ and I have a 105. The number of data is 7 which is again odd. The median is again the 4th observation and which is again 6; it does not change. The median is the $4th$ observation which does not change which is equal to 6.

So, what you can immediately notice here is while the mean was very different for these two data sets, the median is the same for both the data sets even though it differs very drastically in an outlier. So, the median is not very sensitive to the outliers the way the mean was.

Now, let us look at the third data set which had only 6 observations. Again I arrange this data in ascending order. So, I have a 2, I have a 3, I have a $5, 6, 7$ and 105. Again I have n which is equal to 6. So, my median is going to be the mean of the $3rd$ observation and the $4th$ observation which is $\frac{5+6}{2}$which would give me 5.5.

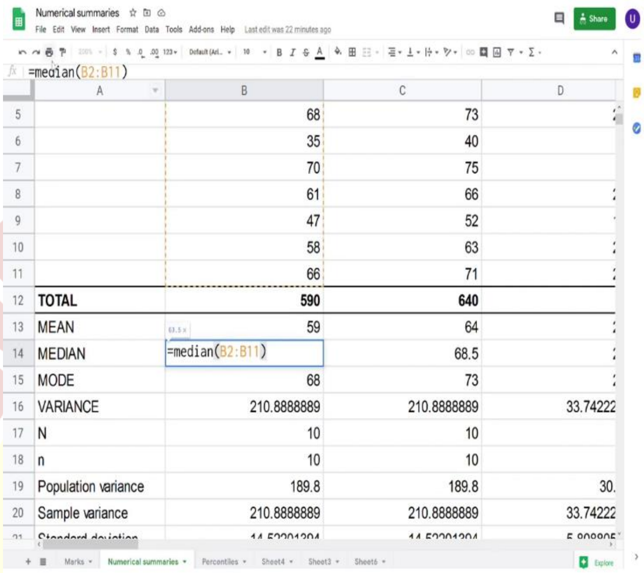
Notice that this 5.5 is not a member of the data set. So, the median need not belong to the data set. Whereas, for the first two data, the date median was a member because we are stay looking at a particular observation; whereas, here I am not looking at a particular observation.

(Refer Slide Time: 06:55)



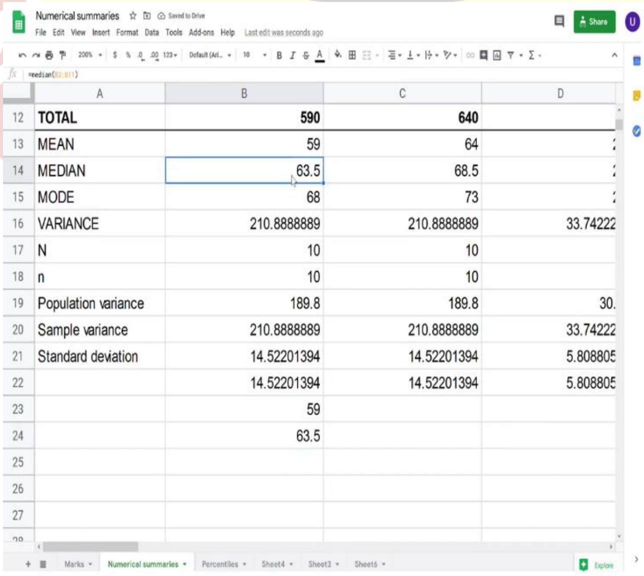
So, let us go back to our example here. So, for here the sample means were the same. In the second one, the sample mean is 21, the sample median is 6. The first data set and second data set only differ in the second observation which is 117 here and 12 here. And we already see that there is a significant difference in the mean, but the median remains the same. So, the sample mean is sensitive to outliers, whereas the sample median is not sensitive to outliers.

(Refer Slide Time: 07:33)



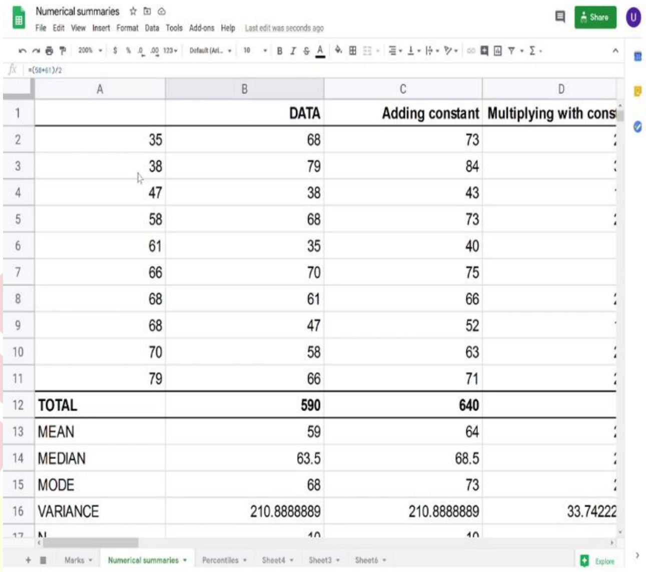
Now, let us compute in our Google sheet. What happens in our Google sheet? So, what is the median? So, you can see that in the Google sheet, the median of data is obtained by the function median $B2$ to $B11$ which will give me the data.

(Refer Slide Time: 08:03)



So, now I can see that the median here is 63.5. How did we obtain 63.5? I can arrange this data.

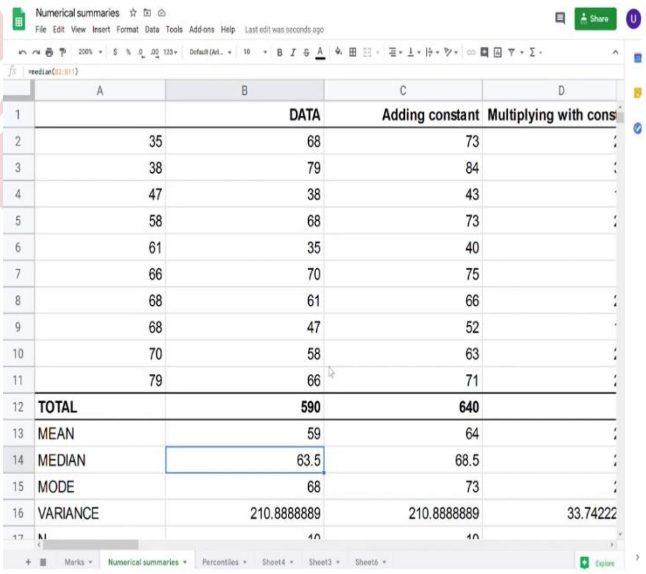
(Refer Slide Time: 08:13)



	A	B	C	D
1		DATA	Adding constant	Multiplying with cons
2	35	68	73	
3	38	79	84	
4	47	38	43	
5	58	68	73	
6	61	35	40	
7	66	70	75	
8	68	61	66	
9	68	47	52	
10	70	58	63	
11	79	66	71	
12	TOTAL	590	640	
13	MEAN	59	64	
14	MEDIAN	63.5	68.5	
15	MODE	68	73	
16	VARIANCE	210.8888889	210.8888889	33.74222

So, if I arrange this data, n equal to even; n equal to 10, which is equal to even. Sorry, 1, 2, 3, 4, 61+ 66, it is which is 63.5. And that is what I have here which is 63.5, because 61 is my $5th$ ranked observation here.

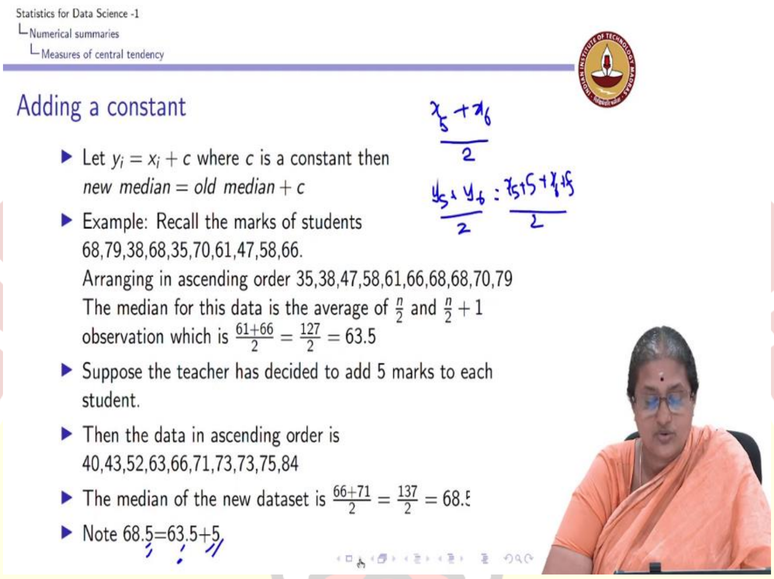
(Refer Slide Time: 08:43)



	A	B	C	D
1		DATA	Adding constant	Multiplying with cons
2	35	68	73	
3	38	79	84	
4	47	38	43	
5	58	68	73	
6	61	35	40	
7	66	70	75	
8	68	61	66	
9	68	47	52	
10	70	58	63	
11	79	66	71	
12	TOTAL	590	640	
13	MEAN	59	64	
14	MEDIAN	63.5	68.5	
15	MODE	68	73	
16	VARIANCE	210.8888889	210.8888889	33.74222

So, this is $5th$ $61 + 66$, I have these observations here. So, you can see the $61 + 66$ which is $\frac{137}{2}$ which will give me 63.5 which is the median. I can get this through the command median of the array in Google sheets.

(Refer Slide Time: 09:07)



So, what happens when we add a constant to the data set? When I add a constant to the data set, again let y_i be x_i plus a constant, where c is a constant, then what happens to my new median? So, let us go back to the example $68, 79$, these are the marks of my students. Again I arrange them in ascending order. When I arrange them in ascending order, we saw that the mean is $\frac{61+66}{2}$ which is 63.5.

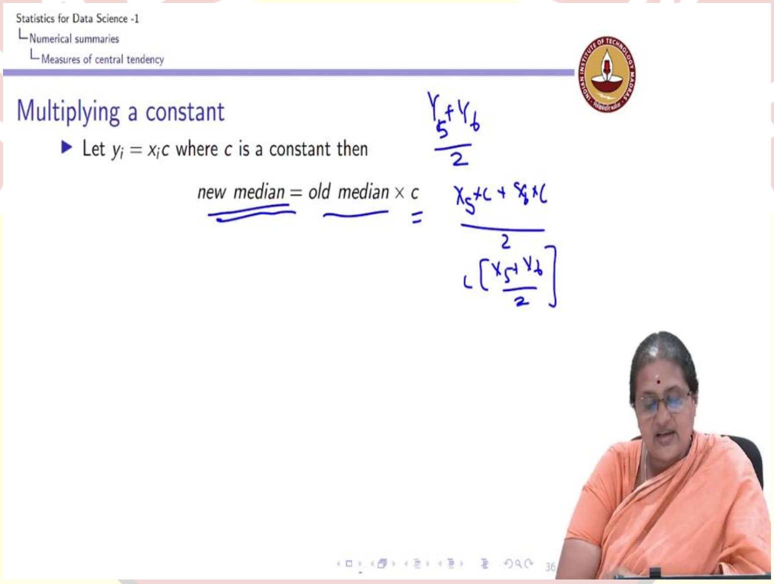
Again if I decide and the teacher decides to add 5 marks to every student, then the data becomes $40, 43, 47 + 5$ which is $52, 63, 66, 73, 75,$ and 84; I am adding 5 to each point of the data set. Now, again you notice that by adding a constant to the data set does not change the order of the observation.

So, here this $\frac{n}{2}$ observation was 61, and $\frac{n}{2} + 1$ is 66. So, corresponding to 61, I have 66; corresponding to 66 I have 71. The n does not change; the number of observations does not change. So, the median in this case is $\frac{66+71}{2}$, and you can see that it is 68.5. Whereas, 66 is was $61 + 5$ – 66 was 71, $66 + 5$. So, the new median is nothing but your old median plus a constant, because the values are the same.

If I have x, so here it is $\frac{x_5+x_6}{2}$ was my old median. My y_5 is $\frac{y_5+y_6}{2}$ is my new median. But y_5 was x_5 plus my constant, y_6 is x_6 plus my constant which is 5. So, I have the new median is $\frac{x_5+x_6}{2}$ which is my old median $+ 5$. So, 5 is the constant.

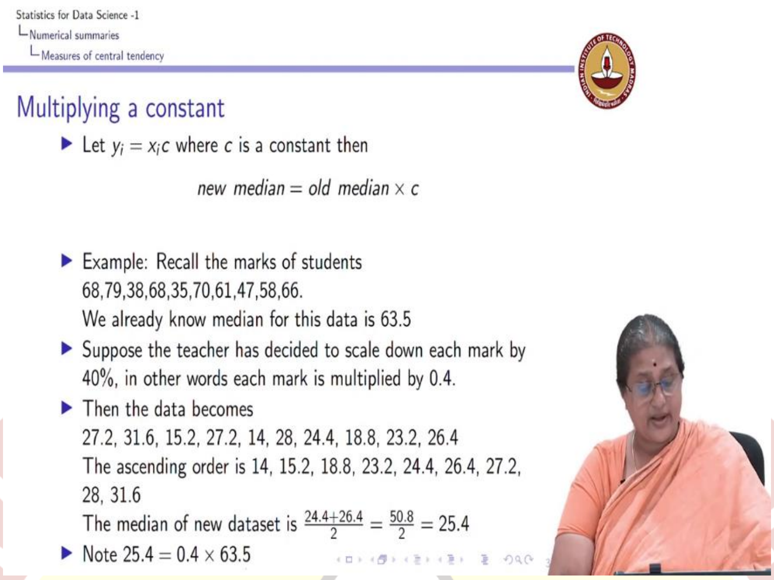
Old median $+5$ is my new median. So, whenever I am adding a constant, the new median is your old median plus the constant. It does not it you are adding that constant to the new median.

(Refer Slide Time: 11:59)



What happens when you multiply the entire data set with a constant? When I multiply the data set with an entire constant, again my old data set, so y_1, so I had $\frac{y_5+y_6}{2}$ which is my new median so, but y_5 is $x_5 +, \times\, c, y_6$ is going to be $x_6 \times c$, I can remove the c. So, I have $\frac{c \times (x_5+x_6)}{2}$, $\frac{x_5+x_6}{2}$, was my old median. So, my old median into the constant will give me the new median. And this we can see in our example.

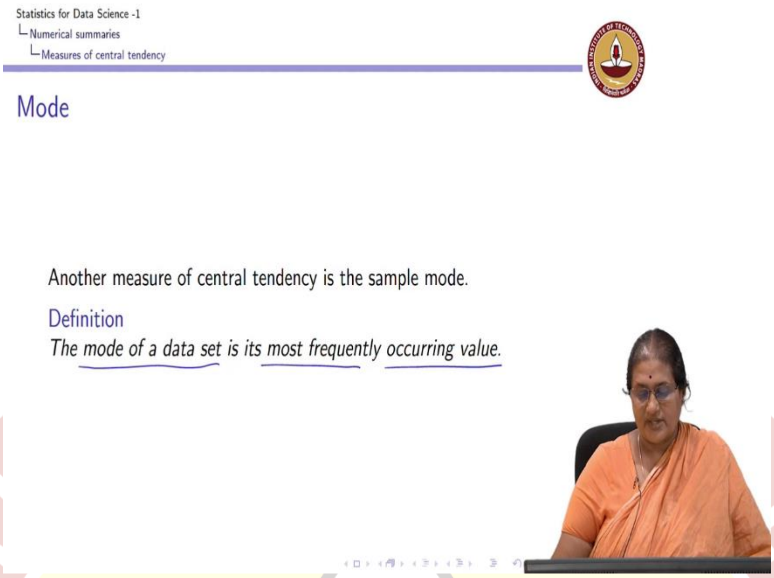
(Refer Slide Time: 12:47)



So, again recall we know the median is 63.5. If I scale down or each mark is multiplied by 0.4, I saw that this is my what is happening to my data set. Again I arrange the data set in ascending order, my $5th$ observation here is 24.4, $6th$ observation is 26.4, the median of the new data set hence is 25.4, which I can verify is 0.4×63.5; 25.4 is the new median which is the old median $\times$ 63.5.

So, when I go back and see that in my Google sheets, so when I am add a constant, so you can see that the old median is 63.5, when I add a constant of 5, $63.5 + 5$ is 68.5, whereas 63.5×0.4 gives me 25.4. So, this is how we can obtain our new median.

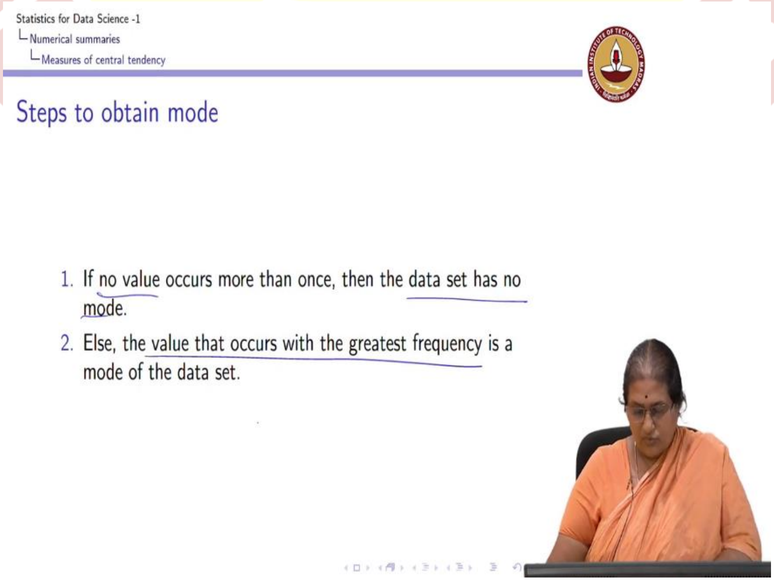
(Refer Slide Time: 14:03)



Now, we move on to the third measure of central tendency which we refer to as a mode. We have already seen what is a mode while describing categorical data. We see that the mode as we defined when we talked about categorical data is that observation which has the highest frequency of occurrence.

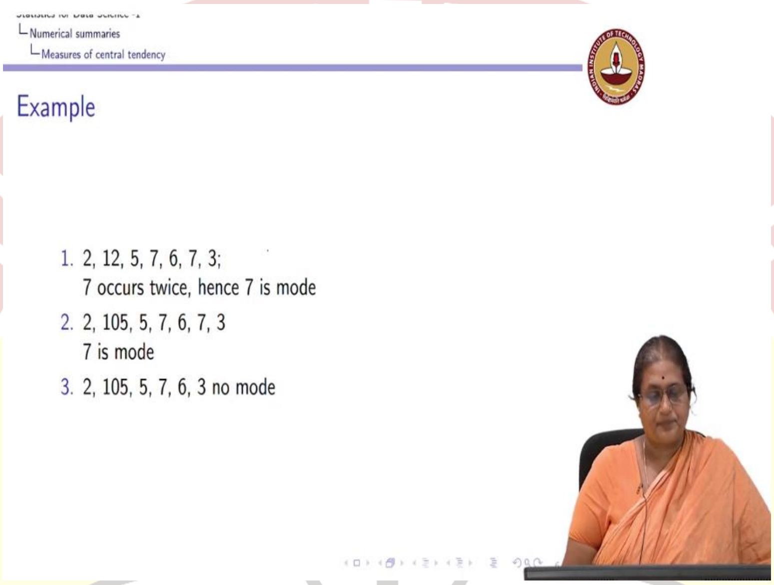
So, that is the same way we define even for numerical data. So, the mode of a data set as it is given here is the most frequency frequently occurring value, so that is what we refer to as a mode.

(Refer Slide Time: 14:37)



So, now how do we obtain a mode? Just as we did in the case of the categorical data, even in the numerical data, what we do is we check for the mode by computing or calculating that observation which appears the most number of ties. If a value occurs more than once, if no value occurs more than one, the data set has no mode; otherwise that value which occurs with the greatest frequency is the mode of a data set.

(Refer Slide Time: 15:15)

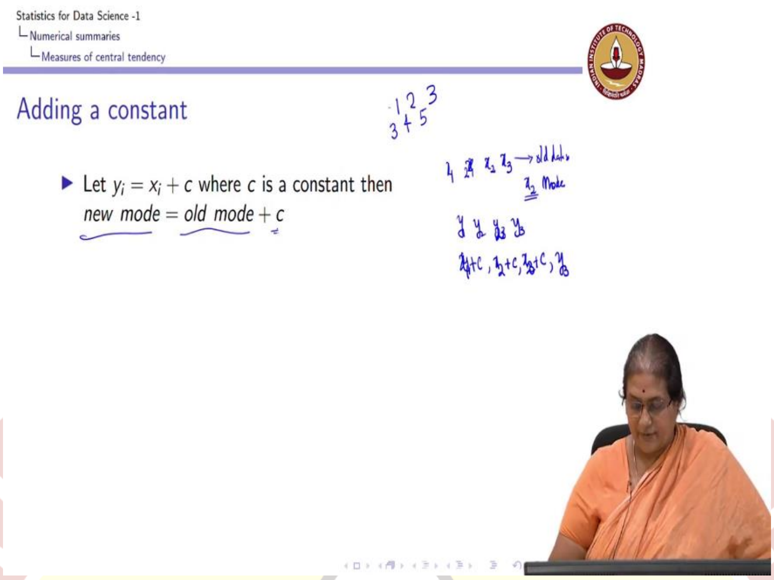


So, now moving forward we find again we go back to the same data sets which you have considered so far. In this data set I have 2, 12, 5, 7, 7, 6, 3. We can see that 7 appears twice, hence the mode of this data set is 7.

The second data set also 7 occurs twice, again you can see that the difference between the first data set and the second data set is only one observation, namely 12 appears in the first data set, 105 appears in the second data set. The mode again is 7 for this data set.

The third data set has all 6 values that are distinct; hence there is no mode for the third data set. Now, again if you look at the first and second data set, recall when we computed the mean, the mean was very different for both these data sets, the median was the same, the mode is also the same for both the data sets.

(Refer Slide Time: 16:27)

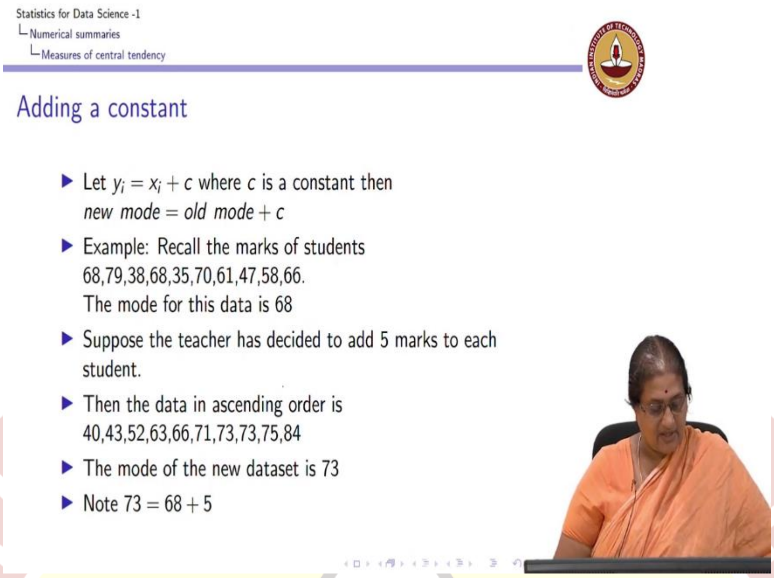


Now, as we did in the case of the mean and the median, let us see what happens when we manipulate a data set namely when we add a constant and when we multiply with a constant. When we add a constant to each of the observations of the data set, for example, I have $1, 2, 3$, and I am adding a constant to each one of them, this becomes 3, this becomes 4, and this becomes 5, I can see that nothing the characteristic of the data set remains the same.

So, the mode of the data set the new mode is just the old mode $+c$. So, the new mode is that. So, if I have x_1, x_2, x_3 which is my old data set, and suppose x_2 is the mode of my old data set, I add to get y_1, y_2, y_3 which is my x_1 plus a constant, x_2 plus a constant, and x_3 plus a constant. If x_2 is that which appears the most number of time, so it would be $x_1, x_2, x_1, x_2, x_2, x_3$; x_2 is the mode.

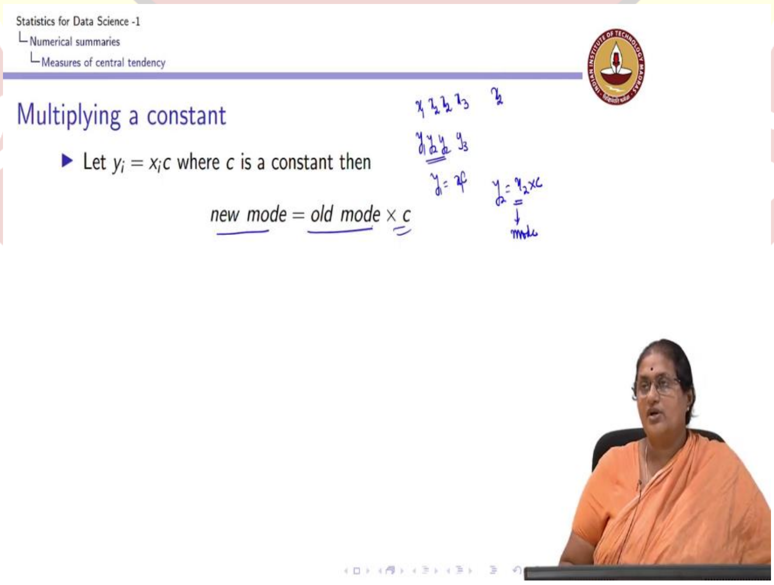
So, it becomes y_1, y_2, y_2, y_3 where y_2 appears the most number of times which so you can see that the new mode is $y, x_2 + c$, hence the new mode of my data set is old mode $+$ the constant.

(Refer Slide Time: 18:07)



Recall again going back to our marks example the data set is 68, 79. I have 10 students, and I can see that the mode for this data set is 68. I add 5 marks just as earlier. And you can see that the data set in ascending order becomes 40, 43, I do not need it in ascending order now, but nevertheless I can see that the mode now is 73 which corresponds to 68 +5. Hence the new mode is nothing but the old mode+ the constant.

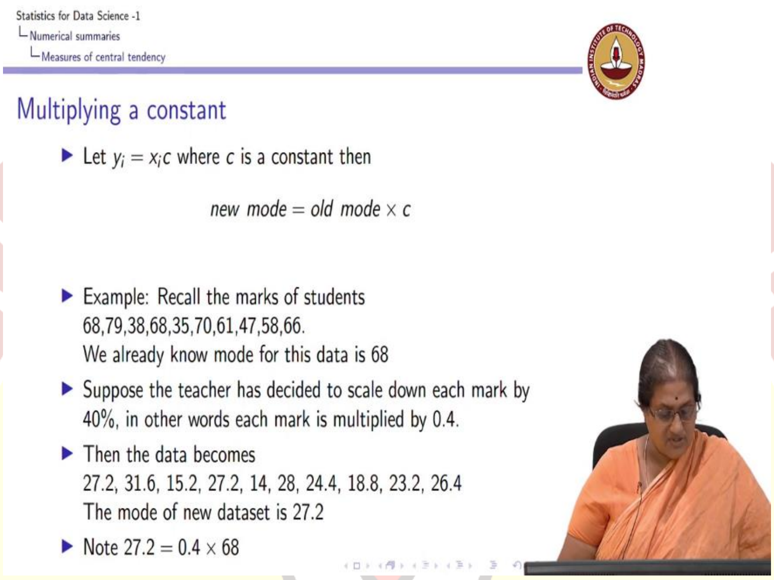
(Refer Slide Time: 18:47)



What happens when we multiply a constant? When we multiply a constant again the new mode is nothing but the old mode times the constant. Again the reasoning is very simple.

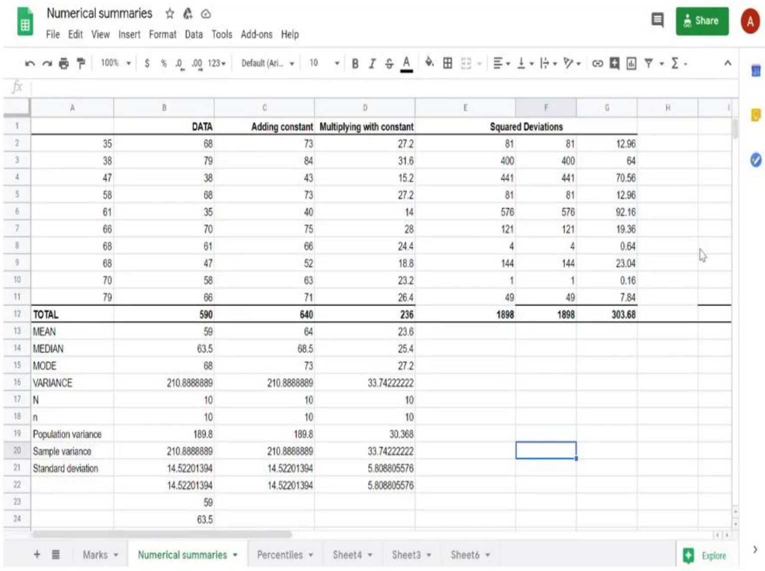
Suppose, I have a data set x_1, x_2, x_2, x_3; x_2 being the mode here; I have y_1, y_2, y_2, y_3 where y_2 is $x_1 \times$ a constant, y_2 is the mode here. And I know y_2 is nothing but x_2 times the constant; x_2 is the mode for my earlier data set. So, the new mode is old mode $\times$ the constant.

(Refer Slide Time: 19:29)



So, again recall the example the mode is 68, this is what we saw from the earlier table example. Now, if the teacher decides to scale down each mark by 40 % and each mark is multiplied by 0.4, the data set becomes the following 27.2, 31.2. The new mode is 27.2, this appears twice. And we can verify that this 27.2 is 0.4 times 68.

(Refer Slide Time: 20:01)

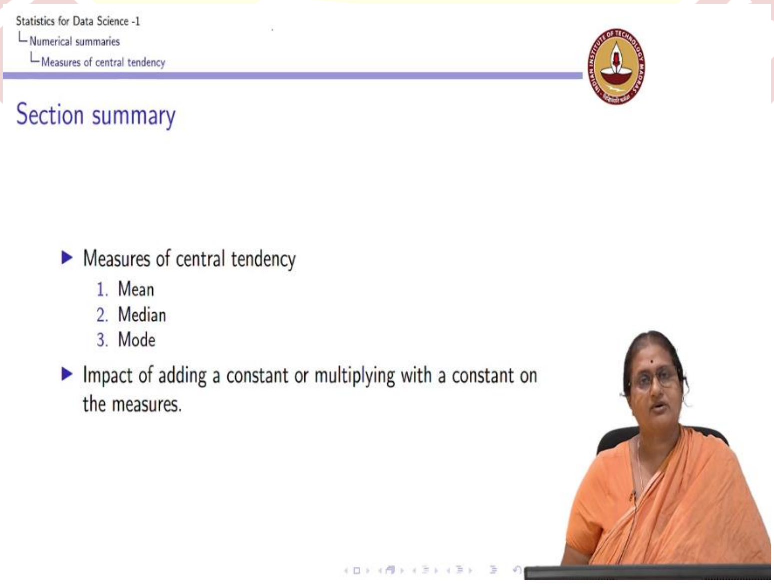


	DATA	Adding constant	Multiplying with constant	Squared Deviations		
35	68	73	27.2	81	81	12.96
38	79	84	31.6	400	400	64
47	38	43	15.2	441	441	70.56
58	68	73	27.2	81	81	12.96
61	35	40	14	576	576	92.16
66	70	75	28	121	121	19.36
68	61	66	24.4	4	4	0.64
68	47	52	18.8	144	144	23.04
70	58	63	23.2	1	1	0.16
79	66	71	26.4	49	49	7.84
TOTAL	590	640	236	1898	1898	303.68
MEAN	59	64	23.6			
MEDIAN	63.5	68.5	25.4			
MODE	68	73	27.2			
VARIANCE	210.8888889	210.8888889	33.74222222			
N	10	10	10			
n	10	10	10			
Population variance	189.8	189.8	30.368			
Sample variance	210.8888889	210.8888889	33.74222222			
Standard deviation	14.52201394	14.52201394	5.808805576			
	14.52201394	14.52201394	5.808805576			
	59					
	63.5					

So, if we look at this, we can go back to our numerical summaries. So, you can see that this is my data set. This I arrange my data set in ascending order here. This is the data set. The highlighted portion is the data set in my Google sheets.

So, you can see from this 68 is that value that appears twice, and that is given by the function mode – mode times the data returns the value 68. When I add a constant, the mode of the new data set is 73; 73 is 68 + 5. And 27.2 is when I multiply it with a constant 27.2 is 68 times 0.4, so that is what we have seen.

(Refer Slide Time: 21:05)



So, moving forward what we have seen so far is we have studied about the measures of central tendency, namely we looked at what is the mean, we define both the population mean and the sample mean.

But then our discussion centered mostly around the sample mean. Then we moved on to define what is a median of a data set, and then what is a mode of a data set. For each one of these operations or measures, we saw what was the impact of adding a constant or multiplying with a constant on the measures.

That is what we have.