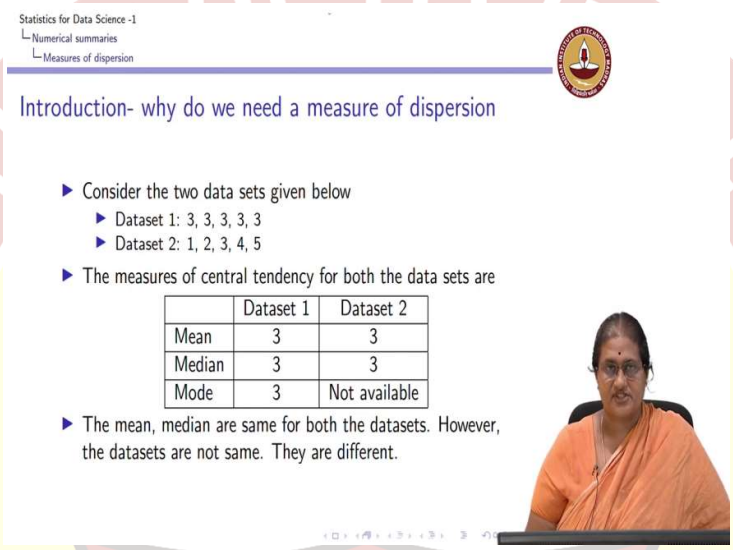
IIT Madras

ONLINE DEGREE

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

Lecture - 3.4
Describing Numerical Data - Measures of dispersion: Range, Variance, and Standard deviation

(Refer Slide Time: 00:14)



The next thing which we are going to look at is Measures of Dispersion. So, why do we need to know about measures of dispersion? So, measures of central tendency actually captures what we call the center or the typicalness of the dataset. So, what is measure of dispersion capture? Why first of all even before going to define what is a measure of dispersion, let us understand why do we need a measure of dispersion.

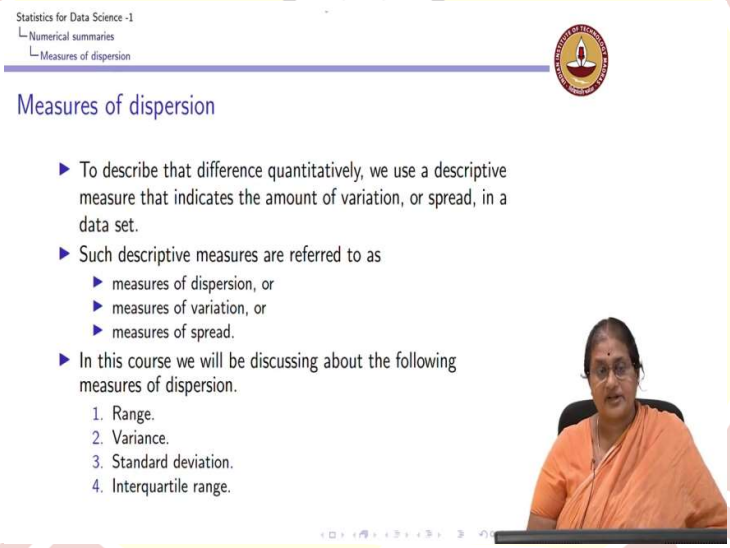
Towards this, let us look at two datasets. The 1st dataset is or both the datasets have 5 observations each, the 1st dataset has observations 3, 3, 3, 3, and 3. The 2nd dataset has observations 1, 2, 3, 4, 5.

Let us work out the measures of central tendency for this dataset. So, if we work out the measures of central tendency for this dataset, we observe the following: the mean of the 1st dataset is 3, the mean of the 2nd dataset is also 3 because the mean of the 1st dataset is 3 + 3 + 3 + 3 + 3 which is 15 divided by 5 which is 3. The 2nd dataset is 1 + 2 + 3 + 4 + 5 which is again another 15 divided by 5 which is equal to 3.

We look at the median, the number of observations is odd which is 5 so, 5 + 1 is 6, 6 by 2 the third observation, the third observation and dataset 1 is 3 again the median for the second observation also is 3. The mode, the mode for the 1st dataset 3 appears there is only one value and 3 which appears 5 times so, the mode is 3 whereas, for the 2nd dataset, there is no mode.

However, when you look at this dataset and only if the numerical summaries and the measures of central tendency are given to you, you somehow tend to believe that both the datasets are very similar in nature because the mean and the median of both these datasets are the same. However, we see that both these datasets are very different from each other.

(Refer Slide Time: 02:52)

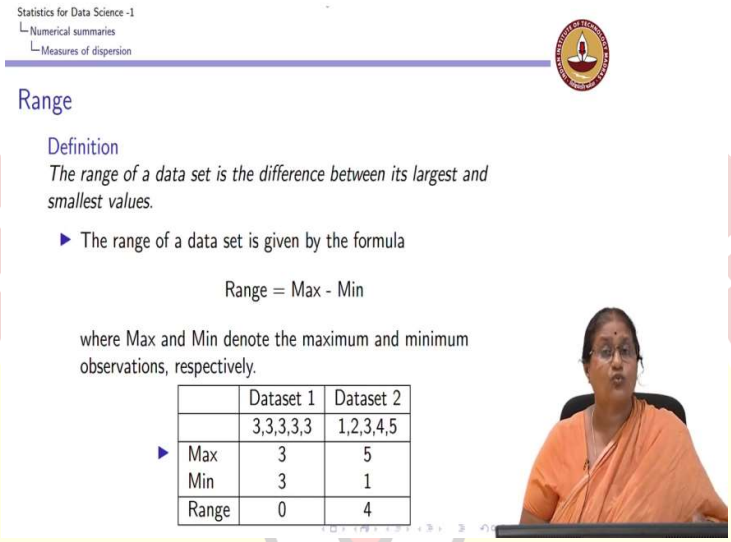


So, they when they are different, we want to see that is there any other measure that can capture this difference and hence, we need to understand what is a measure of spread or dispersion. The first understanding we need to have is I have to describe this difference quantitatively which will actually tell me what is the amount of variation, what is the amount of spread of a dataset.

So, the descriptive measures are popularly referred to as measures of dispersion or measures of variance or measures of spread. What are the key measures of variation we are going to understand or dispersion we are going to understand in this course?

We start with defining what is a range, then we go on to define what is variance, once we establish what is a variance, variance is the most frequently used measure of dispersion, we again we define what is a standard deviation and then, after introducing what are percentiles, we will introduce a notion of a interquartile range.

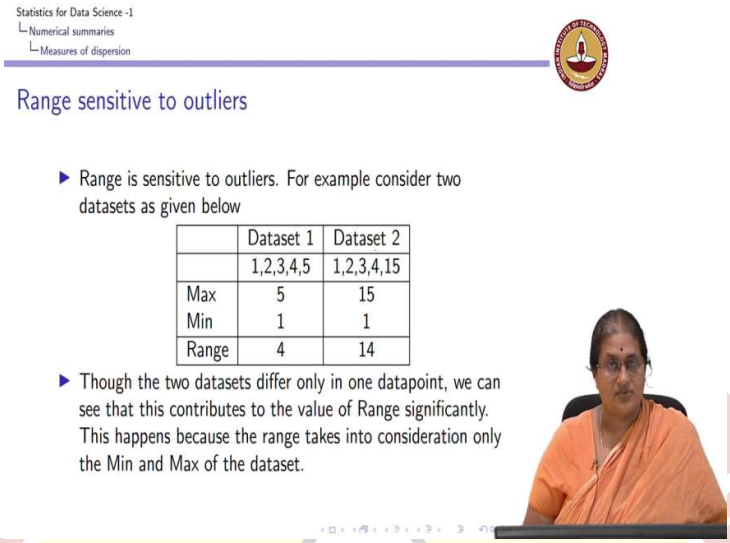
(Refer Slide Time: 03:55)



So, let us go ahead and understand what is a range. A range is defined range of a dataset is defined as the difference between its largest and smallest value. So, the range as the name suggest is a difference between the largest and the smallest value. So, the range is basically maximum - minimum where maximum is the largest and minimum is the smallest value.

So, let us go back to the two datasets we have. The 1st dataset was 3, 3, 3, 3, 3, the 2nd dataset was 1, 2, 3, 4, 5 the maximum of the 1st dataset is 3 because I have only one data value in that, the maximum of the 2nd dataset is 5, the minimum of the 1st dataset is again 3, the minimum of the 2nd dataset is 1. Hence, the range of my 1st dataset is 0 whereas; the range of my 2nd dataset is 4.

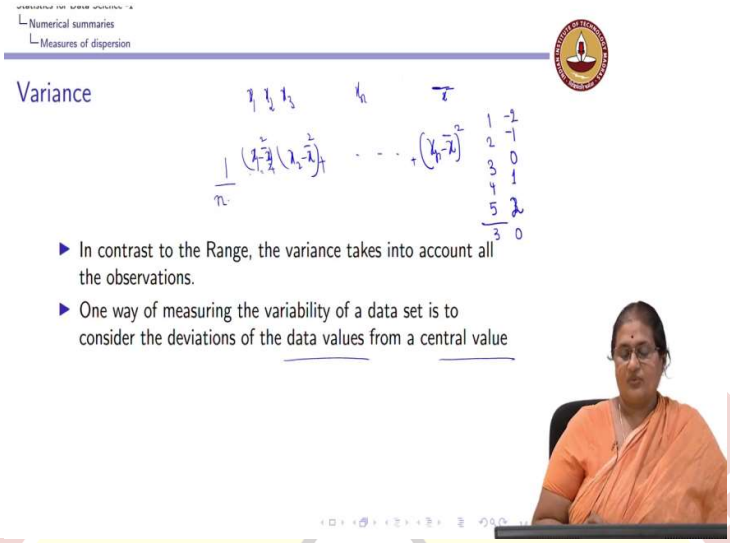
(Refer Slide Time: 05:12)



So, we can see that the range in a sense tells us more about the datasets, then what the measures of central tendency actually revealed. But; however, what is the problem with the range? Now, again consider the two datasets. The 1st dataset is just 1, 2, 3, 4, 5 and the 2nd dataset is 1, 2, 3, 4 and 15. These two datasets differ with each other only in one observation namely I have a 5 here whereas, I have a 15 here. So, this is the key thing where they are different.

So, the maximum of the 1st dataset is a 5, the maximum of the 2nd dataset is a 15, the minimum for both the datasets is 1. So, the range of the 1st dataset is 5 - 1 which is a 4 whereas, the range of the 2nd dataset is 14. So, as you can see very similar to what we observed when we discussed about the mean, we see that the range is extremely sensitive to outliers. So, the range is an extremely sensitive measure because the range takes into consideration only the extreme values to compute it the formula.

(Refer Slide Time: 06:20)



Now, the next measure of dispersion which we are going to talk about and this is the most frequently used measure of dispersion is what we refer to as the variance of a dataset. In contrast to the range, the variance takes into account all the observations. What do we mean by this? Again the range takes into account only the minimum and the maximum namely the extreme observations are taken into consideration when you actually compute the range whereas, the variance takes into account all the observations.

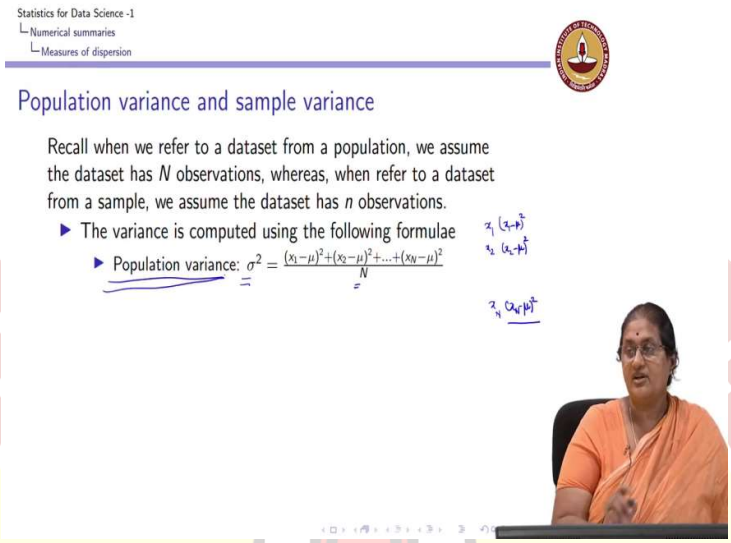
So, the one way of measuring the variability is to consider the deviations of the data value from a central value. What do we mean by this? Suppose, I have a data x_1, x_2, x_3, x_4, x_5, I have a measure of central tendency for now let me call that $\bar{x}$, I have this measure of central tendency which I have defined. So, I have a measure of central tendency which I have defined as $\bar{x}$, okay.

So, the deviation of the data values from this central value is x_1 - $\bar{x}$, x_2 - $\bar{x}$ and so for x_n - $\bar{x}$ this is what I mean by the deviation or the difference of each of the data points from its central value and the central value I have chosen here is the mean, okay.

Now, one thing is in a given dataset for example, let us again take a 1, 2, 3, 4, 5 I know $\bar{x}$ for this dataset is 3, the deviation of 1 from 3 is a - 2, 2 from 3 is - 1, 3 from itself is 0, 4 from 3 is a + 1 and 5 from 3 is 2. We can see that these are the deviations of the dataset, but then after so, I need an aggregate measure, I just cannot give the deviation. One possibility is if I sum up all the deviations, I see it goes to 0. Hence, again I see that

summing up the deviations is not a very good measure of the variability even though I have taken all the data points into consideration.

(Refer Slide Time: 08:59)

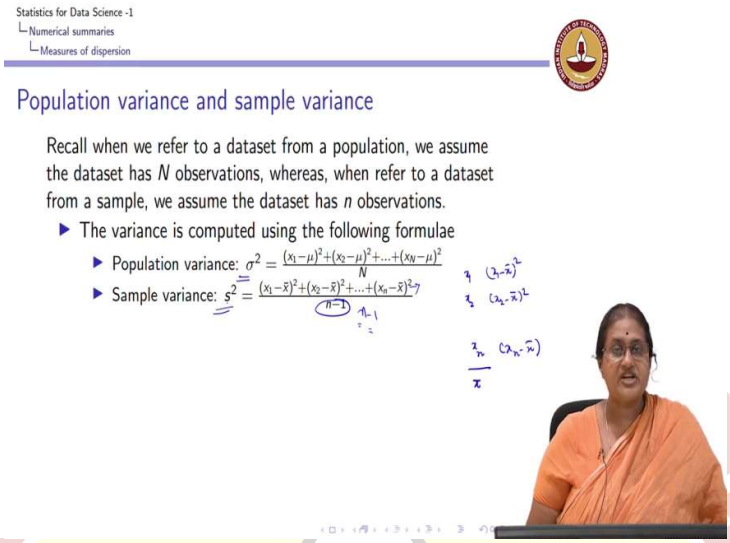


So, what variance the population variance and the sample variance does is the following. It look at the deviation of each of my dataset from data point, from the central value and it squares it up. It adds the squares of the deviation. So, sum of squared deviations from the central value and it averages it. Again I repeat, it takes the deviation of every data's point from its central value, it squares the deviation and adds up all the deviations and divides it by a number.

Now, if it we are talking about a population variance, then the variance of a population variance or the population variance given by σ^2 remember population mean is μ so, I have x_1, I have x_2, I have x_N which are my population units.

x_1 - μ is the deviation of the first unit from the mean, x_2 - μ is the deviation of the second unit from the mean, x_N - μ is the deviation of the nth unit from the mean I square each of these deviations, I add them up, I get the numerator. If I divide the total sum of square deviations with the total number of observations, I refer it to the population variance.

(Refer Slide Time: 10:57)

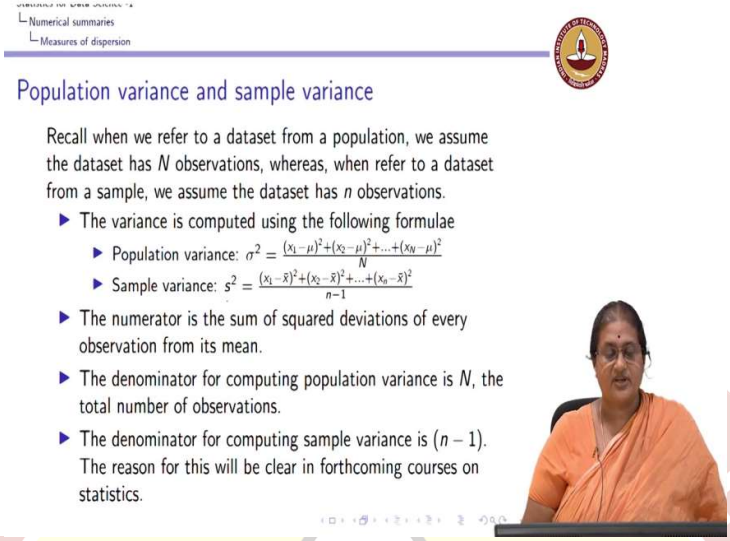


However, when I talk about the sample variance again I have x_1, x_2,…, x_n, n is my sample size, the mean is given by $\bar{x}$. I look at the deviations of each data point from the mean and I square it up, the sum of the square deviations divided by n - 1. Notice the difference between the population variance and the sample variances.

The population variance I divide the sum of square deviations by the total number of observations. When I refer to the sample variance, I divide the total number of sample squared deviations by the number of total observations - 1. There is a reason to do so. The explanation is out of the scope of this particular course and you will learn about this as you go forward.

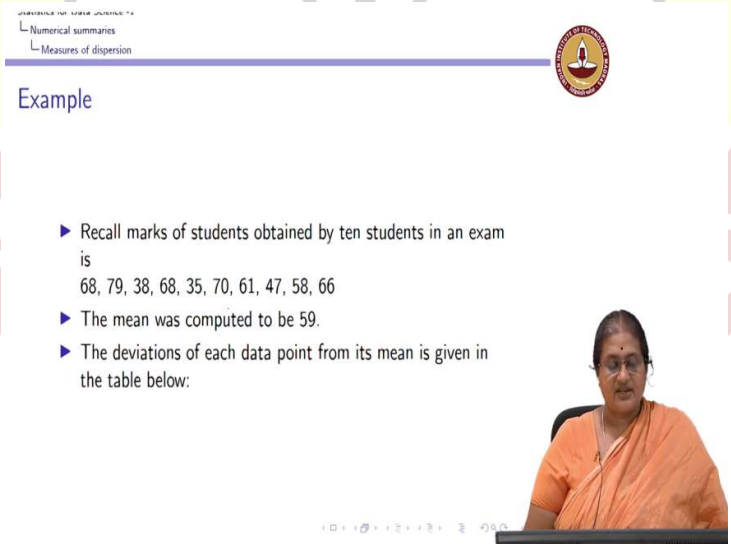
The notation for a population variance I refer as σ^2, sample variance I refer as s^2, okay. So, I repeat the numerator be it the population variance or the sample variance is the sum of squared deviations of a data point from its mean value.

(Refer Slide Time: 12:25)



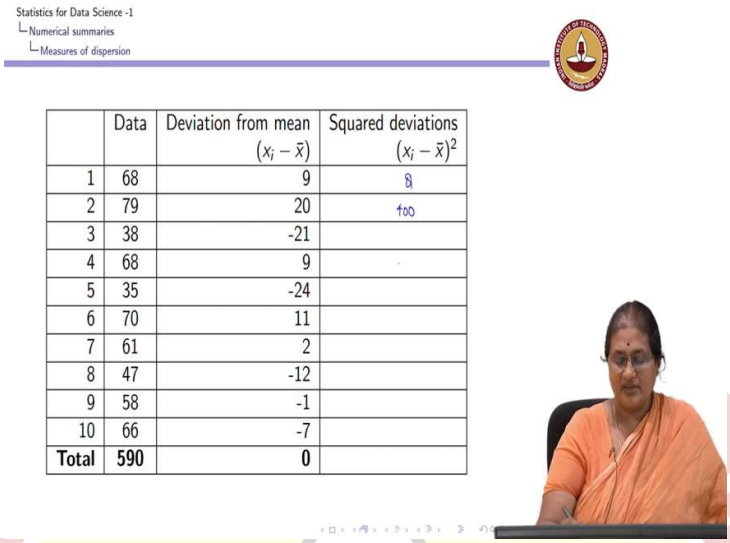
So, the numerator is a sum of square deviation, the denominator for population variance is N whereas, it is n - 1. The reason will become very clear in the forthcoming courses, but for now we are going to restrict ourselves to sample variance.

(Refer Slide Time: 12:43)



So, again go back to looking at the same example that is the marks obtained by the 10 students in an exam. So, the mean remember, we computed the mean to be 59. So, let us compute this deviations.

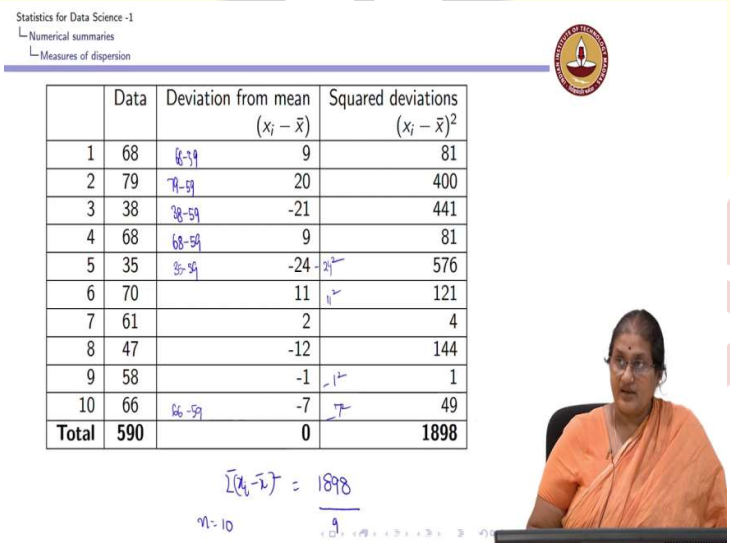
(Refer Slide Time: 13:03)



	Data	Deviation from mean $(x_i - \bar{x})$	Squared deviations $(x_i - \bar{x})^2$
1	68	9	81
2	79	20	400
3	38	-21	
4	68	9	
5	35	-24	
6	70	11	
7	61	2	
8	47	-12	
9	58	-1	
10	66	-7	
Total	**590**	**0**	

So, you can see that the deviation is given here 68 - 59 is 9, 79 - 59 is 20, 38 - 59 is - 21 and the square deviations here I have a 81, I have a 400.

(Refer Slide Time: 13:27)



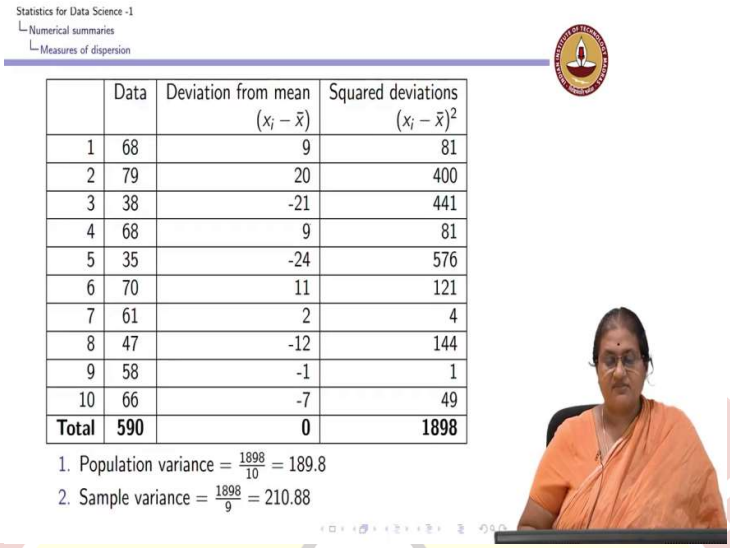
	Data	Deviation from mean $(x_i - \bar{x})$	Squared deviations $(x_i - \bar{x})^2$
1	68	9	81
2	79	20	400
3	38	-21	441
4	68	9	81
5	35	-24	576
6	70	11	121
7	61	2	4
8	47	-12	144
9	58	-1	1
10	66	-7	49
Total	**590**	**0**	**1898**

$\sum(x_i - \bar{x})^2 = 1898$

$n = 10$

So, fourth I keep writing this and I can see that I can find out what are the sum of the squared deviations for each one of the dataset. So, this is 68 - 59, this is 79 - 59, this is 38 - 59, 68 - 59, 35 - 59. So, you can see that and this is 66 - 59, this is 7^2, 1^2, 11^2, 24^2, the sum of the square deviation -24^2, -1^2, -7^2. So, the numerator $\sum(x_i - \bar{x})^2$ is 1898. I have n equal to 10 to compute the sample variance I divided by 10 - 1 which is 9.

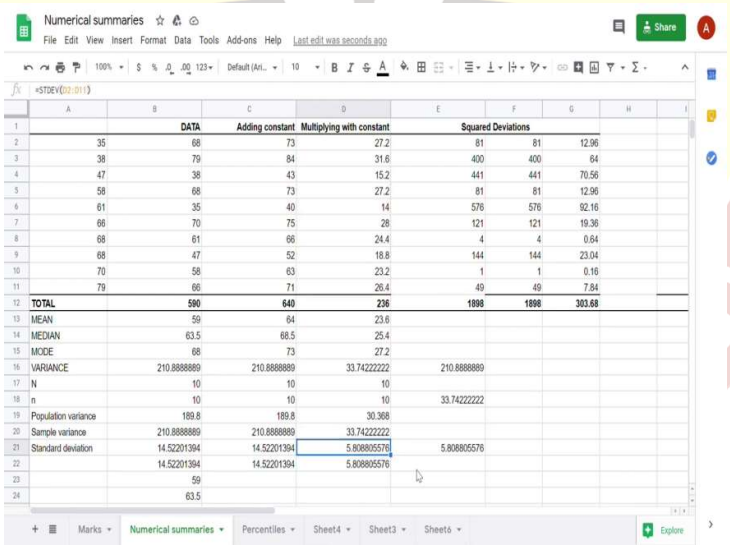
(Refer Slide Time: 14:26)



	Data	Deviation from mean $(x_i - \bar{x})$	Squared deviations $(x_i - \bar{x})^2$
1	68	9	81
2	79	20	400
3	38	-21	441
4	68	9	81
5	35	-24	576
6	70	11	121
7	61	2	4
8	47	-12	144
9	58	-1	1
10	66	-7	49
Total	**590**	**0**	**1898**

And you can see that, my population variance is 189.8 whereas, my sample variance is 210.88, okay.

(Refer Slide Time: 14:42)



	A	B	C	D	E	F	G
1		DATA	Adding constant	Multiplying with constant	Squared Deviations		
2	35	68	73	27.2	81	81	12.96
3	38	79	84	31.6	400	400	64
4	47	38	43	15.2	441	441	70.56
5	58	68	73	27.2	81	81	12.96
6	61	35	40	14	576	576	92.16
7	66	70	75	28	121	121	19.36
8	68	61	66	24.4	4	4	0.64
9	68	47	52	18.8	144	144	23.04
10	70	58	63	23.2	1	1	0.16
11	79	66	71	26.4	49	49	7.84
12	TOTAL	590	640	236	1898	1898	303.68
13	MEAN	59	64	23.6			
14	MEDIAN	63.5	68.5	25.4			
15	MODE	68	73	27.2			
16	VARIANCE	210.8888889	210.8888889	33.74222222	210.8888889		
17	N	10	10	10			
18	n	10	10	10	33.74222222		
19	Population variance	189.8	189.8	30.368			
20	Sample variance	210.8888889	210.8888889	33.74222222			
21	Standard deviation	14.52201394	14.52201394	5.808805576	5.808805576		
22		14.52201394	14.52201394	5.808805576			
23		59					
24		63.5					

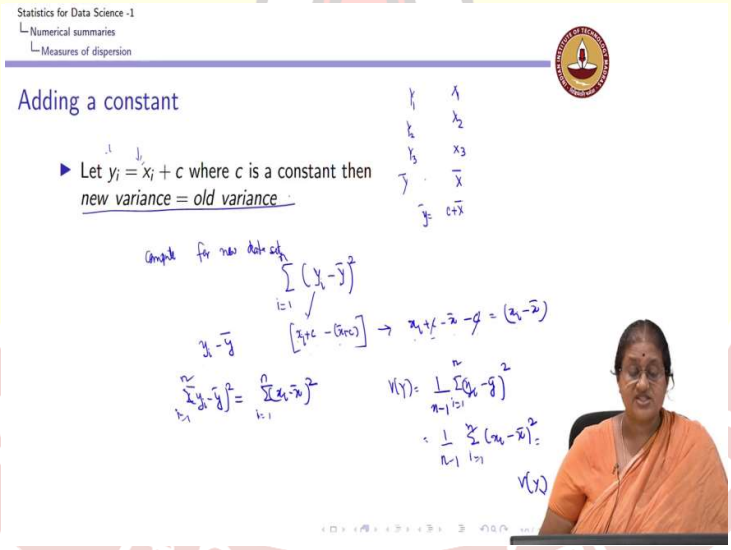
So, now let us look at how to compute this using our Google sheet. So, in our Google sheets, this is again my data is what I have highlighted here, okay. So, now, if you look at this portion here, this is the squared deviation. So, you can see that this square deviation is B2, B2 is my data point, B13 is my mean whole square. So, this corresponds

to the 81 I have here okay. Similarly, I have a 400 the second data point, I have 441 the third and so forth my total sum of square deviations as highlighted here is 1898.

So, this divided by my 10 would be 189.8, but when I am dividing it by 9 I get 210.88. Now, the same thing if you use the function VAR.S; VAR.S, S to represent the sample statistic VAR.S of the array returns the sample variance. I reply I repeat VAR.S returns the sample variance and we can see that this is equal to the sum of the square deviation divided by 9 and I get the same value.

Now, the population variance is nothing but now as in the earlier case, let us see what happens to these variance when we manipulate the dataset. What we mean by this is what would happen to the dataset if I add a constant or I multiply each one of the values with the constant.

(Refer Slide Time: 17:10)



So, what would happen if I add a constant? So, I have y_i is $x_i + c$ where c is a constant. So, again let me just look at 3 numbers y_1, y_2, y_3. I have x_1, x_2, x_3. $\bar{x}$ is the mean of these three numbers, $\bar{y}$ is the mean of these three numbers. We have already seen $\bar{y}$ is c + $\bar{x}$ this is already what we have seen.

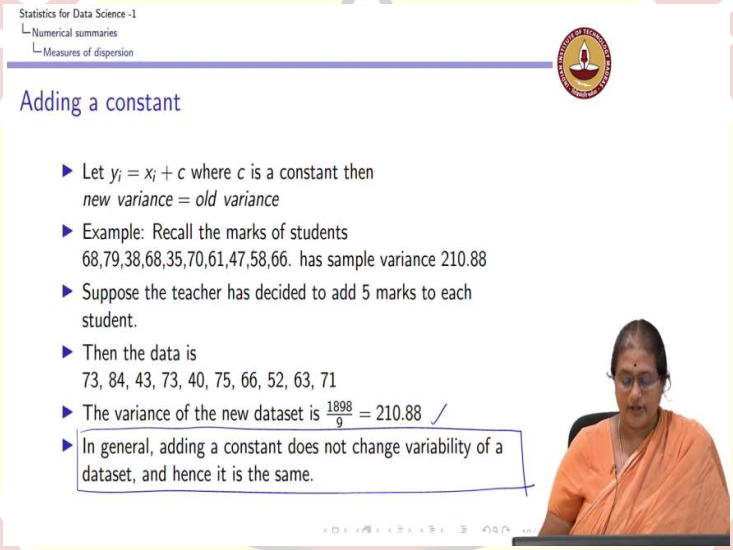
So, now, when I compute my variance for the new dataset, I get a $\sum_{i=1}^{n}(y_i - \bar{y})^2$. My y_i so, it is $x_i + c$ that is each y_i, $\bar{y}$ is $\bar{x} + c$ so, you can see that this is $x_i + c - \bar{x} - c$ which

is same as $x_i - \bar{x}$, is that clear. So, I have $y_i - \bar{y}$ is $x_i + c$ because each y_i is $x_i + c$, $\bar{y}$ is $\bar{x} + c$. So, it is $x_i + c - \bar{x} - c$ which is these two get cancelled out and I get $x_i - \bar{x}$.

So, I have the numerator$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2$. Hence, v(y) = $\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ = v(x).

Hence, if I add a constant to every value of my dataset, the variance of the dataset does not change the new variance is equal to the old variance and that is what we have just seen.

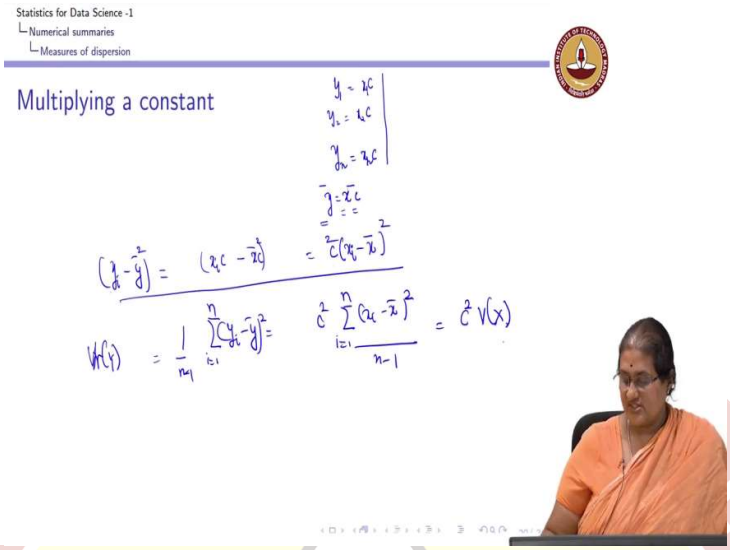
(Refer Slide Time: 19:49)



So, let us look at a example to see what is happening. Recall we just computed the marks of the students and the sample variance is 210.88, we add find marks the new dataset is this and you can see that the new variance of this dataset is also 210.88. In general, we have adding a constant does not change the variability of a dataset.

Let us go back to this example. Again this is my original data the one that is highlighted, the variance was 210.88, I add a constant and I get this as my dataset and see that the variance for both these datasets is the same.

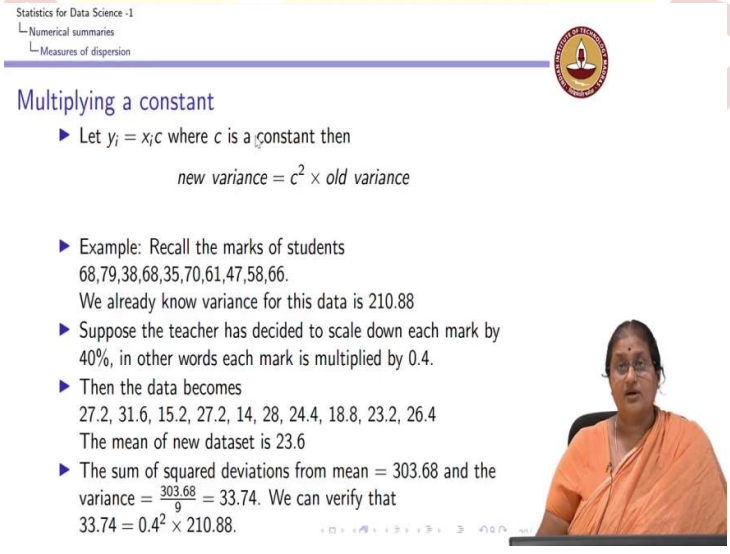
(Refer Slide Time: 20:58)



So, adding a constant does not change variability of a dataset. What happens when we multiply the dataset with a constant? Again, I have my $y_1 = x_1c$, $y_2 = x_2c$, $y_n = x_nc$, I know $\bar{y} = \bar{x}c$. So, my $y_i - \bar{y} = x_ic - \bar{x}c = c(x_i - \bar{x})$, okay.

I repeat, for every $y_i - \bar{y} = c(x_i - \bar{x})$. So, $(y_i - \bar{y})^2 = c^2(x_i - \bar{x})^2$. So, $\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 = c^2 \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$. Hence, v(y) = c^2v(x), is that clear. So, this is my dataset I already know $\bar{y} = \bar{x}c$. I just substitute the values and I get the v(y) = c^2v(x).

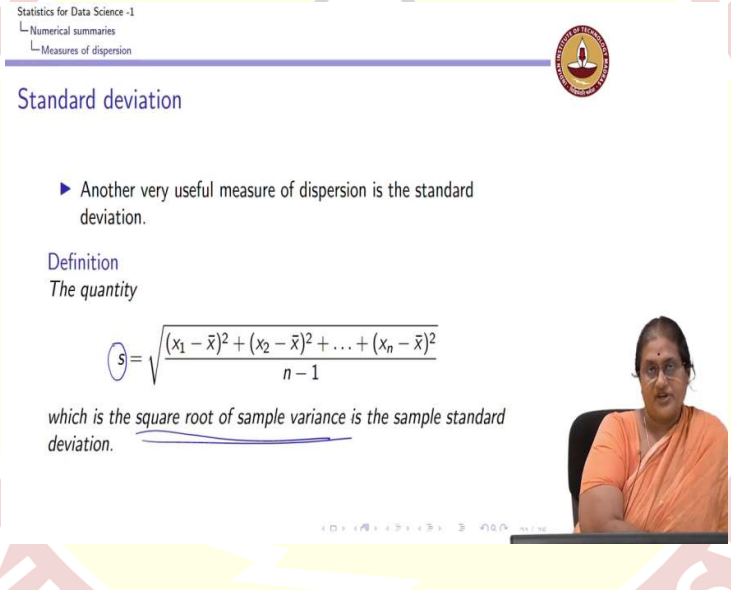
(Refer Slide Time: 22:53)



So, we have the following that if $y_i = x_i c$, the new variance = c^2 × old variance, we have just established that relation. We can verify that using our dataset. I already know the data the variance for the dataset is 210.88, I multiply it with 0.4, I know this is my dataset the mean of this new dataset is 23.6 and I can compute that the variance which is 33.74 is 0.4 square 210.88.

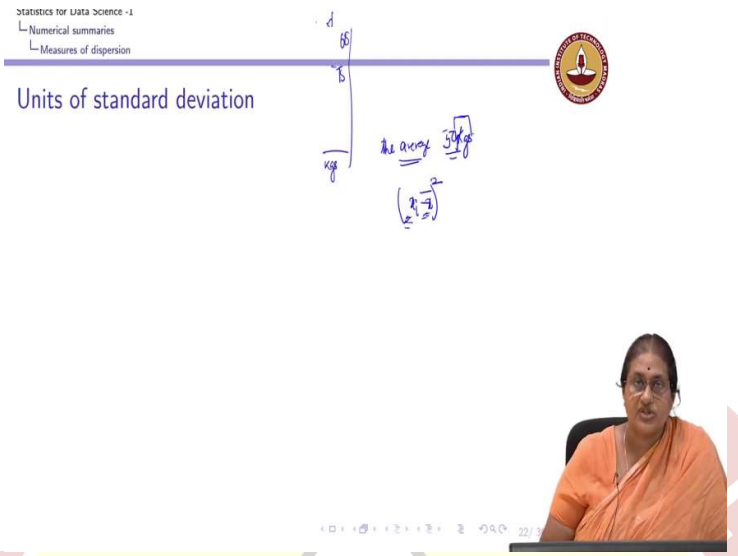
So, you can see that when I multiply with a constant, I get 33.74 is my variance and I can verify that this is 0.4 times the old variance. So, this 0.4 times the old variance is my new variance. So, this is how we compute the dataset with adding a constant and multiplying with a constant.

(Refer Slide Time: 24:08)



Another useful measure of dispersion is the standard deviation. The standard deviation is nothing, but the square root of the variance. Now, why are we interested in the standard deviation? The standard deviation is referred to by the small lower-case letter s. So, the standard deviation is the square root of the sample variance. Similarly, I can define this population standard deviation also.

(Refer Slide Time: 24:40)

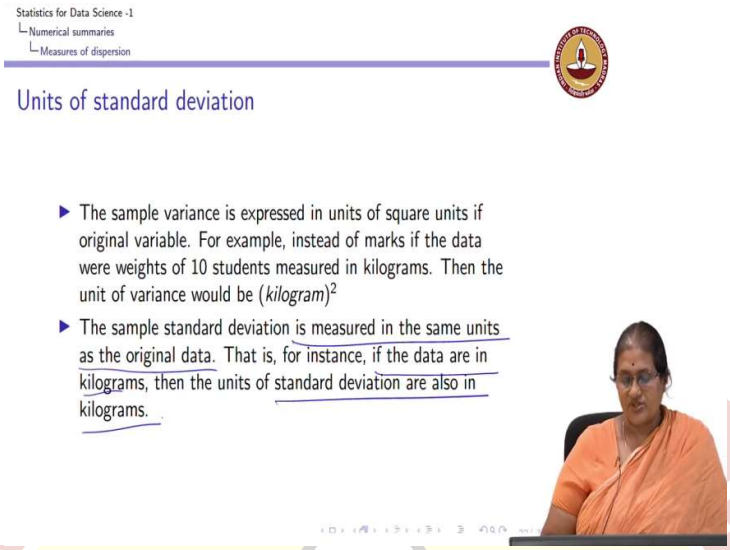


So, why do I require a standard deviation? Remember whenever I talk about a numerical measure, there are units associated with the numerical measures. For example, if I were looking at were ages of people or I was looking at heights of students or I was looking at the weights of students 68, suppose 75 instead of marks they were weights of students measured in kilograms, then the average which was 59 which I did not given any units, if it were weights would have been 59 kilograms.

In other words, the average of a dataset has units which is same as the original measurement. I repeat, the average of a dataset have the same units as that of the original dataset. But when I am computing the variance, what I do is I will take the dataset. So, average has the same units as the dataset.

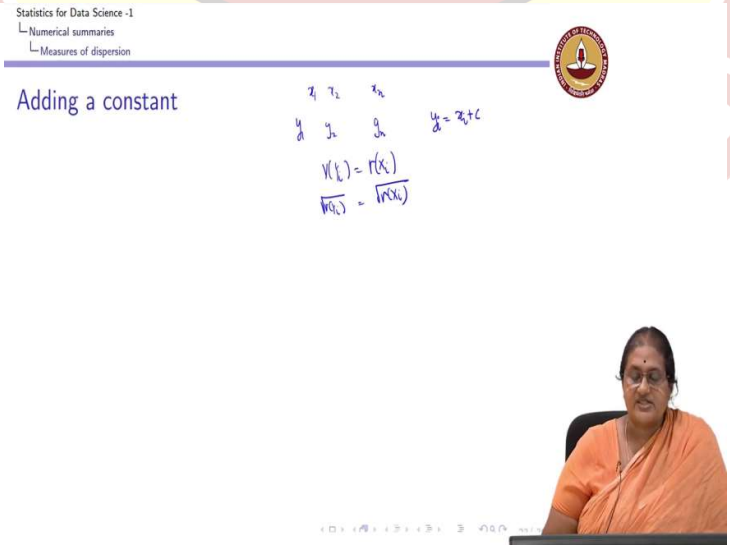
So, if it is kilogram this is a kilogram, this is a kilogram I am squaring it up so, difference will also be kilogram so, the square is not going to be kilogram, it would be kilogram square I add units of kilogram square so, my units of the variance is square of the units of the original variable.

(Refer Slide Time: 26:23)



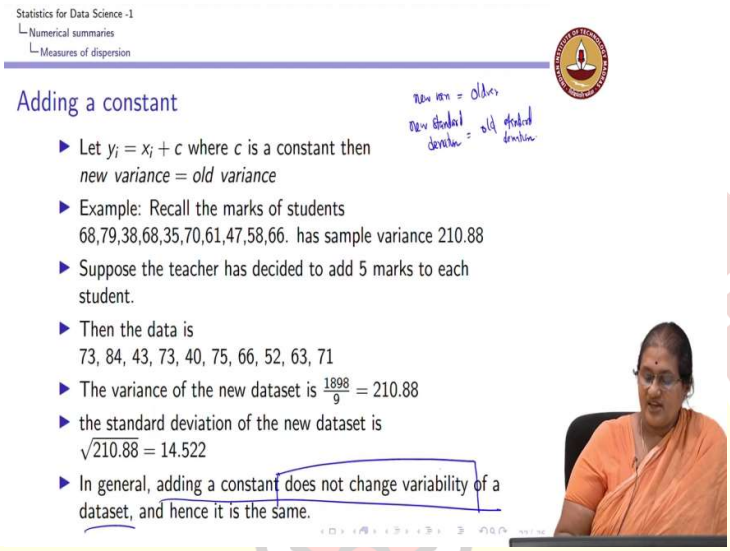
So, to overcome this, what we do is we define what is called the standard deviation so, that the units of measurement of the standard deviation and the original units of measurement are the same. So, that if I am looking at variance, I have a unit which is $(\text{kilogram})^2$, I bring it back to kilograms. So, the sample standard deviation is measured in the same units as the original data. If the data is in kilograms, the units are also in kilograms.

(Refer Slide Time: 27:03)



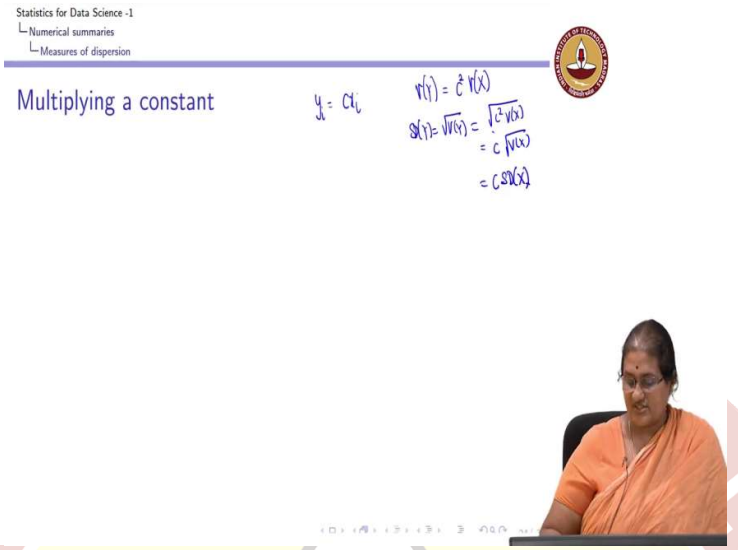
So, what happens when we add a constant to the data? So, again I have $x_1, x_2, .., x_n$, I am getting $y_1, y_2, .., y_n$ where each $y_i = x_i + c$. I know $v(y_i) = v(x_i)$. Hence, $\sqrt{v(y_i)} = \sqrt{v(x_i)}$.

(Refer Slide Time: 27:37)



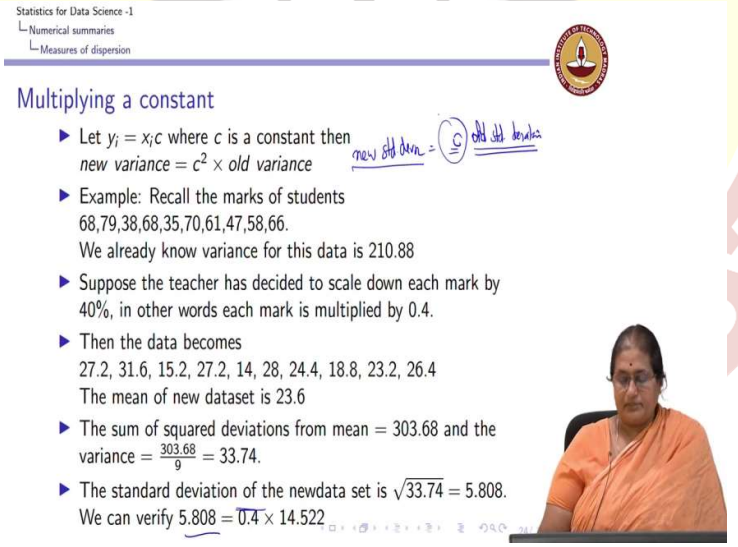
And hence, the standard deviation is also the same, the new variance is equal to old variance. Recall this, the sample variance was the same, the sample standard deviation is also the same. Hence, the constant does not change the variability. Both new variance is equal to old variance, new standard deviation is equal to old standard deviation. In other words, adding a constant does not change the variability of a dataset.

(Refer Slide Time: 28:25)



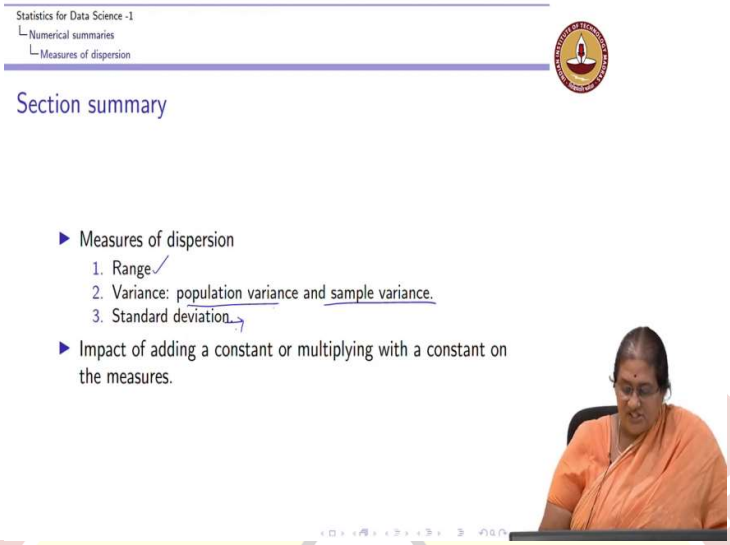
What happens when we multiply a constant? Again, you know that if $y_i = cx_i$, we saw the v(y) = c^2v(x). Hence, SD(y) = $\sqrt{v(y)} = \sqrt{c^2 v(x)} = c\sqrt{v(x)} = c$ SD(x).

(Refer Slide Time: 28:56)



So, when I multiply with a constant, I know new variance = c^2 × old variance. So, when I look at the dataset, you can see that the standard deviation is 0.4 × 14.522. So, new standard deviation = c × old standard deviation this is my c and I can see that, that is the c times the old standard deviation which is 14.522.

(Refer Slide Time: 29:33)



So, what let us look at this in our. So, you can see that, when I look multiply it with a constant, the standard deviation of my original dataset is 14.522. This you can verify is nothing, but my square root of the sample variance. So, square root of the sample variance is 14.522. In the Google sheets, the command is STDEV standard deviation of B2 to B11 gives the standard deviation.

I can see that when I add a constant, the standard deviation remains the same. You can see that the standard deviation for these two, the highlighted columns are the same whereas, when I multiply it with a constant, you can see that the standard deviation is 5.808. You can verify this is 0.4 times my standard deviation is 5.808 and you can see that is equal to my standard deviation of multiplying with the constant.

Hence, when we multiply it with a constant, you can see that the standard deviation, the new standard deviation is the constant times your old standard deviation, okay. So, adding a constant does not change the variability of a set, multiplying with a constant changes the variability of a set by a scalar multiple.

So, what we have seen so far is we started with range, we saw the range is very sensitive to outliers. We defined what was population variance and sample variance. The definitions are very important because just by talking about variance, we need to know whether we are referring to population or sample variance. The numerator captures the sum of square deviations; the denominator is a number of observations if you are

considering population variance and number of observations - 1 if you are considering the sample variance.

Then, we introduce the notion of a standard deviation which has the same units as the original data. We also saw that when you add a constant variability of a dataset does not change whereas, when you multiply with a constant the variability changes.