Mine Automation and Data Analytics

Prof. Radhakanta Koner

Department of Mining Engineering

IIT (ISM) Dhanbad

Week - 07

# Lecture 34: Introduction - II

Welcome back to my course, Mine Automation and Data Analytics. In the earlier lesson, we introduced the basic concept of statistics and the central tendency measures. Then we started with measures of the dispersion, and we completed up to the range. We will start from what we left in the last lesson in this lesson. So, in this lesson, we will cover the following. We will compute and interpret numerical summaries of data. In that, we will measure dispersion and calculate.

The range is done in the earlier lesson. Now, we will compute the variance and standard deviation. We will calculate the inter-percentile and quartile range IQR. We will compute and interpret five number summaries and try to find the association between two variables. This is very important in finding the relationship between the two variables.

So, this is to be understood from the scatter plot, and then, based on that, we will calculate and compute the covariance and correlation. So, what is variance? So, in contrast to range, we have seen that range is nothing but the difference between the maximum and minimum. So you can see that the range was max minus min. So, this range is susceptible to outliers, so we have to navigate to another measure. So, variance is one measure that is not sensitive to the outlier.

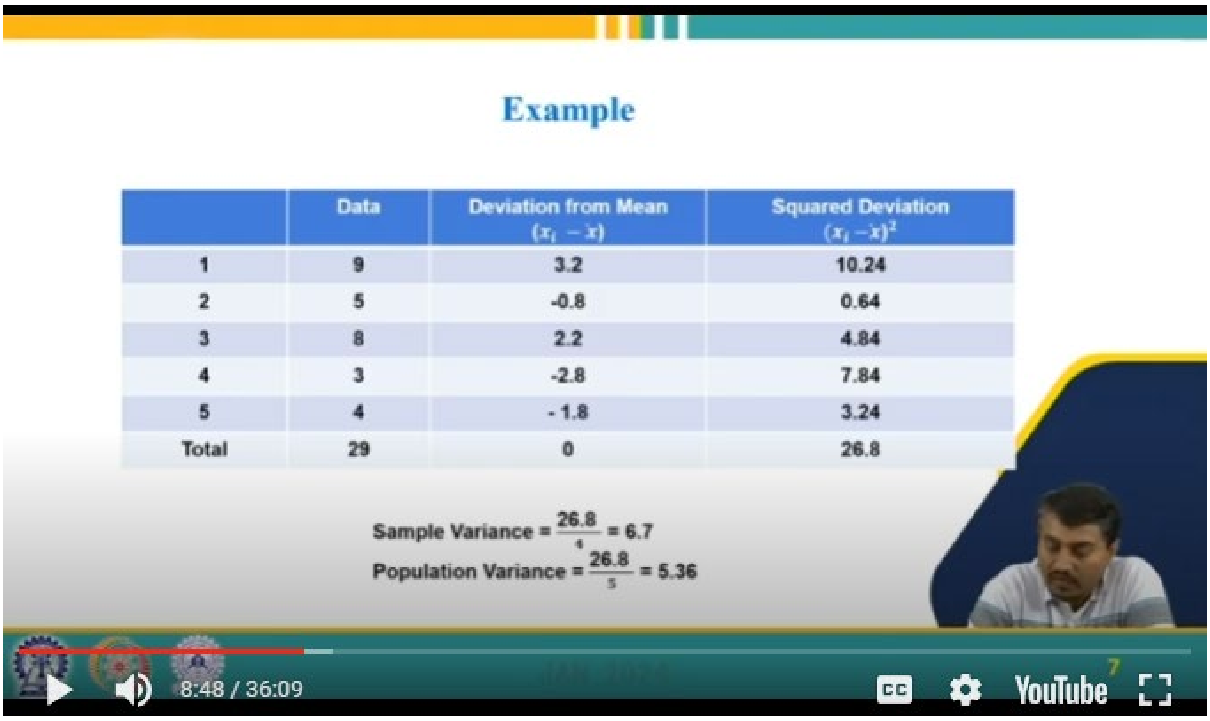
So variance is considered in all these observations, and here, the range is only considered the max and minimum values where variance is considered from X1 to Xn values all these values. So, one way of measuring the variability of a data set is to evaluate the deviation of the data X1, X2, X3, and Xn from a central value, and the central value we assume here is the mean. So, population variance and sample variance. So, when we refer to a population data set, the data set has a capital N number for the observation population. So, in contrast, when we refer to a data set from a sample, we assume the data set has a small n number of observations.

**Population Variance** $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$

**Sample Variance** $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$

So we compute the population by this for sample variance s square is nothing but X1 minus X bar whole square plus X2 minus X bar whole square dot dot dot dot Xn minus X bar whole square divided by n minus 1. So that is summarized as i equals 1 to n, that is, Xi minus X bar whole square divided by n minus 1. Here, one difference is there, so for the sample variance, it is n minus 1, and for the population variance, it is capital N that we must remember. In these and subsequent lessons, we will maintain that the sample variance is a square s and the population is a sigma square. So this is two nomenclature we will use. This is for the population, and this is for a sample; for this population, we do not have to minus 1, and here, we have to negate one value from the N. This is the only difference. So, what is the unit of variance you have seen here? It is nothing but the value observation X minus the mean. Also, the same unit of the whole square is the unit of this particular value divided by N, which is unrelated to the unit.

So, the unit of this variance is the square of this unit of the observation. So, if it is a meter height of different persons, the variance unit is a meter square. If we measure the time taken for the athlete to complete 100 meters, suppose the seconds 10 seconds, 12 seconds, 13 seconds, so the different values for different observations, so the variance unit is the second square. So, that is a problem because the unit square is challenging to imagine, but in reality, it is the square of the original value. So this is one example we have cited, and you see the 9, 5, 8, 3, 4, the total is 29, so the average is 29 divided by five observations, so 5.



Example

|  | Data | Deviation from Mean<br>$(x_i - \bar{x})$ | Squared Deviation<br>$(x_i - \bar{x})^2$ |
|---|---|---|---|
| 1 | 9 | 3.2 | 10.24 |
| 2 | 5 | -0.8 | 0.64 |
| 3 | 8 | 2.2 | 4.84 |
| 4 | 3 | -2.8 | 7.84 |
| 5 | 4 | - 1.8 | 3.24 |
| Total | 29 | 0 | 26.8 |

Sample Variance = $\frac{26.8}{4}$ = 6.7

Population Variance = $\frac{26.8}{5}$ = 5.36

8. for the X bar to be equal to 5.8, we have to calculate X 1 minus X bar. We have to calculate X 2 minus X bar like here, so this is for observation two, observation one, and observation 3. Hence, this is X 1 minus X bar, so nine minus 5.8 is 3.2; similarly, all is calculated. Now, we have to square the square deviation to 3.2 square is 10.24, 0.8 square is 0.64,

and you see the unit is changed because it is square value, so you are summing over all because summing over Xi minus X bar, we have to sum it over. Hence, the summation is 26.8 divided by n minus 1 for the sample variance, so the sample number 5 should be 4. So that is shown in the next slide. Here is the sample: the sample variance is divided by 4, 26.8 divided by 4 is the sample variance, and the population variance is 26.8, divided by 5, so the sample variance is 6.7, and the population variance is 5.36.

So just for the earlier observation we have seen for the mean median mode whether it is change adding and multiplying a constant value with the set so here for the addition of some constant value to the data space variance is not changed why it is so let us see that so our data was X 1 X 2 X n and X bar that is mean is equal to X 1 plus X 2 X n divided by n so the variance for is this is X 1 minus X bar square plus X 2 minus X bar square dot dot X n minus X bar square divided by n minus 1 so what will happen now in a new data space when we add some value so X 1 plus C X 2 plus C X n plus C what is the variance because we have seen that when you add some value the mean is also change that same value so X bar plus C so this particular expression will be X 1 plus C minus X bar minus C because the new mean value is X bar plus C so this will cancel out for all cases so the same expression will remain when you add C constant with each sample so in the observation you have seen that when you add some constant value the expression remains the same so variance remains the same so the old variance is the new variance here for the multiplication for the multiplication let us see so for the multiplication it is basically C X 1 minus C X bar because in the multiple case we have seen the mean will also be multiplied by the old mean so this is the square plus X 2 C minus X bar C square so and so so basically what will happen C is the common factor so C square of X 1 minus X bar square C square X 2 minus X bar square same so this expression remains the same whereas C 2 is basically the common factor and multiply the old variance so when we multiply in a with with some constant value in the sample space then the variance is the new variance is basically a C square multiplication of the old value so variance is affected due to the multiplication of some constant value in the sample space

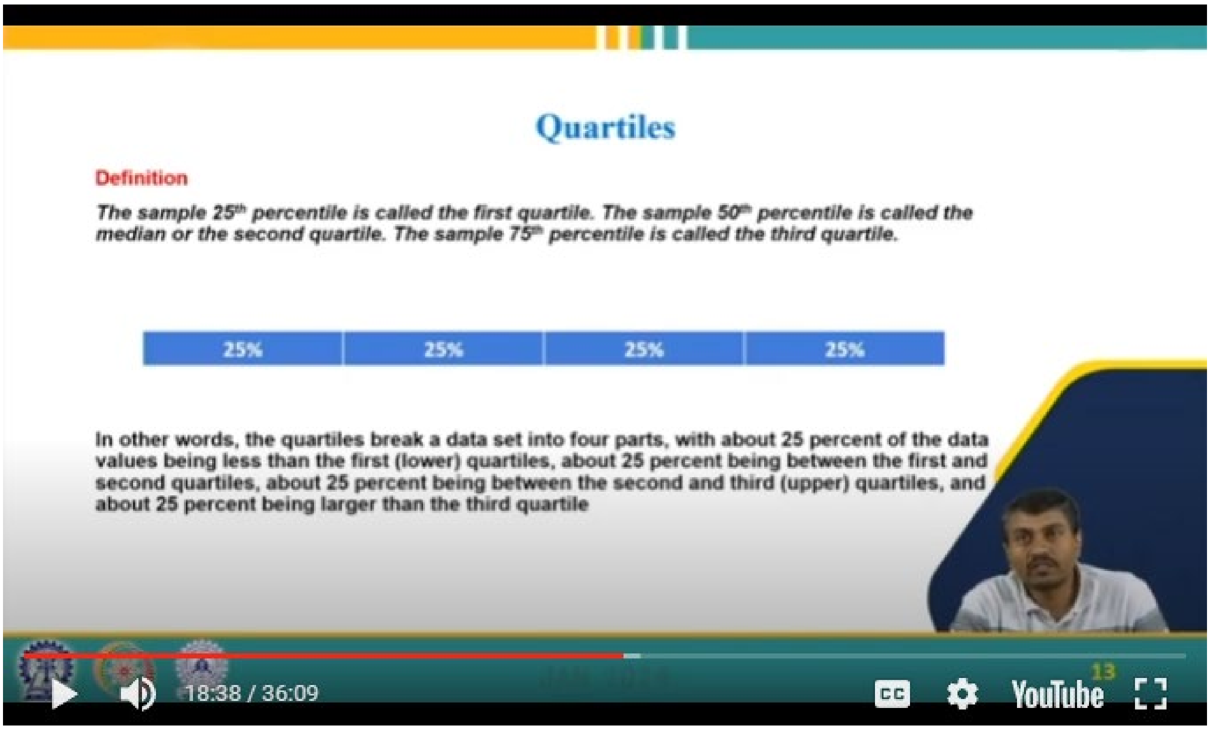
now let us discuss the another concept about the dispersion that is standard deviations so this is a this is an another measures for measuring the dispersion so the quantity of the square root of the sample variance is the sample standard deviation so what we have discussed earlier just the slide before the variance is a square root of that value is the standard deviation so here one thing is sorted out the observation of the height unit in the in the variance was meter square whereas the original unit is meter so in standard deviation when you again root it square root it so again it is down to the meter or in a time taken by the athlete to cover the hundred meter so that is second so again it is second square square root is second so unit problem is resolved and this is a very handy tool to measure the dispersion of the data

$$\mathbf{S} = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2}$$

so the for the sample you see for the sample variance was n minus 1 so here we keep and for the representation of the standard deviation for the sample is s and for the population it is basically Sigma so here the Sigma would be Sigma would be a square root of i is equal to 1 to n xi minus x bar divided by n so this is the Sigma so we will use this is the population population standard deviation and this is sample standard deviation so nomenclature we have to follow for the sample it is s and for the population Sigma example so here the same observation so total was 29 and mean was x bar was equal to 5.8 okay so mean was 5.8 so this is the deviation this is the square of deviation then this is the sum so for the sample standard deviation we have to divide by n minus 1 that is 4 in the square root of that is basically the sample standard deviation so this has been done so 26 divided by 4 is basically 6.7 square root of that is equal to 2.58 so this is the sample standard deviation this is the population variance square root is 2.31 is the population standard deviation,

the difference between the variance and the standard deviation. Hence, nothing but the square root of the variance value is the standard deviation for the respectively for the sample and the population. Unit of standard deviation: as I said, the unit of the standard deviation is the original unit of the variable, so here, for the height, it is a meter for the distance covered by the athlete to cover 100 meters. The second standard deviation unit is also second. Hence, it is a more handy representation, so we always assign or describe that value or data set in terms of standard deviations for all the standard protocols or standard data sets. It is a more widely used measure of dispersion. Now, adding multiplying a constant we have seen earlier during the calculation of the variance when you add the constant value, you multiply the continuous value, so in the case of an addition, you have seen that variance is not affected old variance is equal to the new variance after adding a continual factor C in the sample space and for the multiplication C is multiplied the new variance is the C square multiplication of the old value. Hence, the standard deviation is the square root, so the old value is the new value when the addition is the same, so the square root of the same value is identical.

The standard deviation is not changed when you add some value, but when you multiply some value that was a C square, multiply the variance so C square root is C, so the new standard deviation for this data set when you multiply C with all the sample data the new standard deviation will be the old standard deviation multiplied by the C this is the only difference quartile. So, the quartile is a concept that subdivides the data set into four percentile orders. The 25th percentile is the first quartile, the last 25%, and the second up to the 20th 50th percentile is the second quartile. The sample 27th percentile is the third quartile, and the last is the fourth Q4. In other words, the quartile breaks the data set into four parts, so with about 25% of data value being less than the first, that is the lower quartile, about 25 percentile being between the first and the second quartile, about 25% being between the second and the third quartile. About 25% is more significant than the third quartile.



## Quartiles

Definition

*The sample 25th percentile is called the first quartile. The sample 50th percentile is called the median or the second quartile. The sample 75th percentile is called the third quartile.*

In other words, the quartiles break a data set into four parts, with about 25 percent of the data values being less than the first (lower) quartiles, about 25 percent being between the first and second quartiles, about 25 percent being between the second and third (upper) quartiles, and about 25 percent being larger than the third quartile

So these are the divisions Q1, Q2, and Q3; as you know, a journal paper has quartile numbers Q1, Q2 journal, and Q3 Q4 journals like that. So, in that case, for the journal, Q1 Q2 is the top 25%, then the medium 25% between the top 25 and the medium 50%, so that this finance and the last is last 50 to 75%, the last 75 to 100% that order is reversed. So, the IQR interquartile range or IQR is the difference between the first and third quartile. So IQR is equal to Q3- Q1, so this value is essential for box plots that we might use from time to time for statistical analysis of the data set. Five-number summary: so in a five-number summary, there are five components. The first one is the minimum value for the range calculation. We have seen minimum and maximum values. These are the two components of the five numbers.

The other three components are Q1, the first quartile; Q2, the second quartile; and Q3, the third quartile, or the upper quartile. So here is the data range: 11, 22, 15, 29, 33, 15, 17, 22, 19, 25, and 27. So you have to arrange the data in an orderly, increasing order. This is the box plot, the maximum and minimum, the upper two, and the IQR. The association is a significant interpretation of the relationship between the variables, which is essential in analysis. This association of one of these represents how these variables are associated. There are different ways to measure this association, so we will discuss how these associations can be calculated in subsequent slides.

So, the data pattern in one variable occurs in a particular manner related to the data pattern in one or several other variables. So, for example, X is the variable, and Y is a variable, so how the variable plot of the X and Y is interrelated so that measures the association. So, the scatter plot is like this: a graphical representation of the relationship between two numerical variables. It allows you to visually inspect the pattern of data points and understand the association or correlation between the variables. So here you see the X value is the X value on an increasing order variable 1, and Y is the Y value axis variable two. It is also in increasing order. You see, when X is increasing, Y is also increasing. X is expanding in this direction, and Y is also growing. Hence, they found a relationship: X is also increasing, and Y is also growing.

This is called a positive relationship. If the points on the scatter plot generally form an upward-sloping pattern from left to right, it indicates a positive correlation. As one variable increases, another also tends to grow, which is a critical observation. This is because variable one is rising, but variable two value is decreasing, so it is reversed. What we saw in the last slide, which was earlier, was a positive correlation. Here, it is a negative relation. So conversely, the point on the scatter plot from a downward-sloping pattern from left to right indicates a negative correlation, which means that one variable increases and the other decreases. Now, this is another scatter plot from the scatter plot. It isn't easy to understand how X and Y, variable one and variable 2, are related because this is distributed randomly.

No significant relationship was found, particularly in 2D space. Remember, particularly in 2D space, we could not find any ties from this scatter plot. This can be categorized as no relationship under 2D, so if the point appears randomly scattered without any apparent pattern, it suggests no solid linear relationship between the variables. However, other relationships, such as non-linear or complex associations, might still exist when the dimension is increased. Measures of association: how do we measure the strength of association between two variables from the earlier plot? We have seen positive correlation, negative correlation, and no relation. So, one way is the covariance. Another way is to find the correlation, which lets us see the concepts of covariance and correlation.

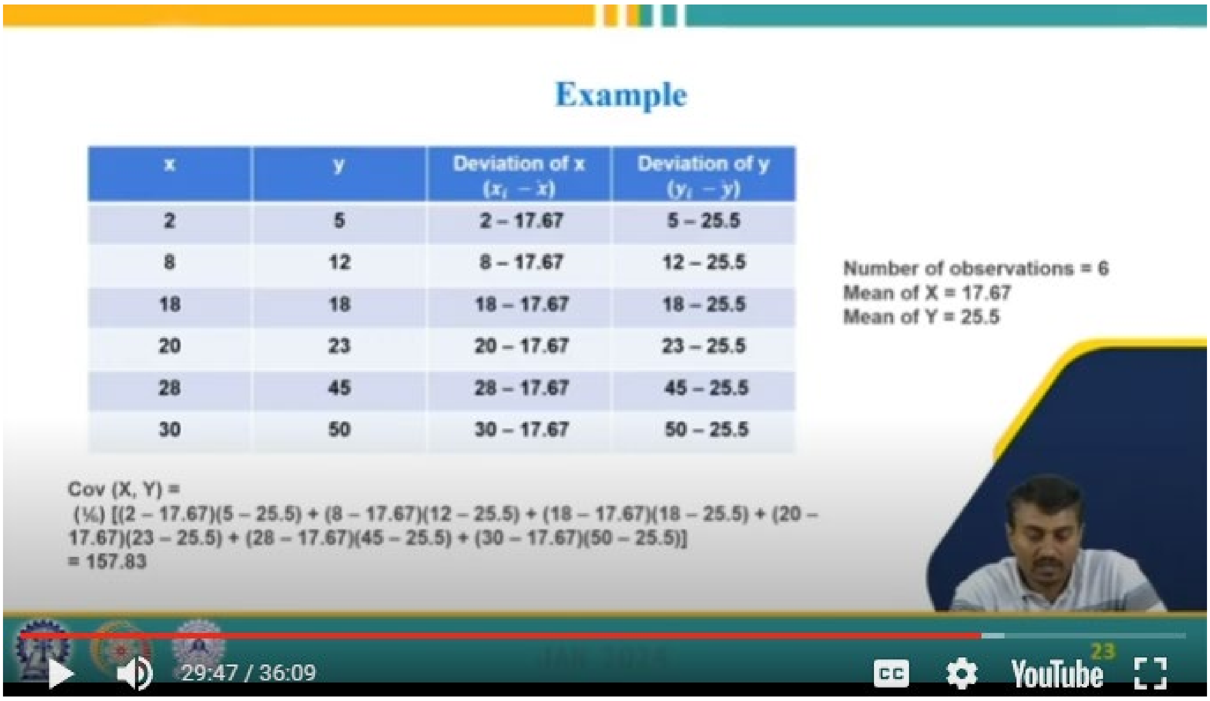
Covariance quantifies the strength of the linear association between two numerical variables. Remember, it is the strength of the linear association between the two numerical variables. So here is a scatter plot of data 1 increasing in the x-axis in the positive direction and increasing on the y-axis in the positive direction. So here, this is the covariance ellipse, so you find a significant amount of data representing a good relationship between one and another, data 1 and 2. So, from these, we have to calculate the covariance. Covariance is calculated based on this relation, so we have seen in the observation earlier that x1 x2 x1 xn, so x1 x2 xn is the observation in the sample.

Now, y1 y2 yn is another observation of another variable, so we want to find out the association between data set 1 and variable one or two. This is variable 1. So, what does variable 1 mean? The sample mean x bar variance for the second variable y1 y2 yn is y bar. So, we are going to calculate the population variance. Our population or sample variance is between these two variables, called covariance of x y. For the sample, the sum of i equals 1 to n xi minus x bar multiplied by y xi yi minus y bar and divided by n minus 1. This signifies that the sample variance for the population variance is more or the same; instead of n minus one, it is only N. So, we must remember that covariance for the sample variance is n minus 1 for the population. This is N, capital N.

$$\textbf{Population Variance : } Cov(x, y) = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{N}$$

$$\textbf{Sample Variance : } \quad Cov(x, y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

So, capital N represents the number of observations for the population, and the small n represents the number of observations for the sample—for example, the association between a variable x and y of a person. So, here, the x value is given, and here, the y value is provided, so we have to calculate the x bar, and the x bar is equal to 17.67, y bar is equal to 25.5. So, the division of the xi over the x mean for all these observations is calculated here, and for the y is also calculated. So now, based on this, the number of observations is 1, 2, 3, 4, 5, 6 okay, and the mean of x is 17.67, mean of y is 25.5. So, the covariance is one by 6 for the population calculation, and the population variance is one by 6.



## Example

| x | y | Deviation of x $(x_i - \bar{x})$ | Deviation of y $(y_i - \bar{y})$ |
|---|---|---|---|
| 2 | 5 | 2 – 17.67 | 5 – 25.5 |
| 8 | 12 | 8 – 17.67 | 12 – 25.5 |
| 18 | 18 | 18 – 17.67 | 18 – 25.5 |
| 20 | 23 | 20 – 17.67 | 23 – 25.5 |
| 28 | 45 | 28 – 17.67 | 45 – 25.5 |
| 30 | 50 | 30 – 17.67 | 50 – 25.5 |

Number of observations = 6
Mean of X = 17.67
Mean of Y = 25.5

Cov (X, Y) =
(⅙) [(2 – 17.67)(5 – 25.5) + (8 – 17.67)(12 – 25.5) + (18 – 17.67)(18 – 25.5) + (20 – 17.67)(23 – 25.5) + (28 – 17.67)(45 – 25.5) + (30 – 17.67)(50 – 25.5)]
= 157.83

So two minus this into five minus this multiplied eight minus 1.7. So this, this, this multiplied by this, this multiplied by this, multiplied by this, multiplied by this, multiplied by this, multiplied by this, multiplied by this, multiplied by this. So, this is calculated here. So, the population covariance for this data set is 157.83 unit of covariance, the size of covariance, however, is difficult to interpret because the covariance is a unit, and the units of the covariance are those of the multiplication of x variables times the y variables. Correlation, a more easily interpreted measure of linear association between two numerical variables, is a correlation. So it is derived from covariance, and to find the correlation between two numerical variables x, y divide the covariance between x and y by the product of the standard division of x and y, and then we will find out the correlation that is Pearson correlation coefficient between x and y, and that is calculated like this. As you have seen, the sample covariance, so sample covariance was, is basically the covariance divided by n minus one and divided by n minus one was there and is also n minus 1 square root; n minus 1 square root goes up.

$$r = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\ \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{Cov\,(x\,,y)}{s_x s_y}$$

So it is divided by n minus one earlier. So now it would be above, so n minus one multiplied by n minus 1. So this is finally n minus one divided by n minus 1. So this particular cancel out. So this is the sample standard division Sx, the sample standard division of Sy, or the variable y. So the Pearson correlation coefficient R is nothing but the covariance of x, y divided by the standard division of sample x and the sample standard division of y Sx and Sy.

So now you can see the covariance unit was the multiplication of this x and y, and here, the unit of the standard division Sx is the unit of x, and the Sy unit is the unit of y. So, the multiplication of units of x and y is a unit less. Measuring the association by some value rather than by unit is essential. So, this is a handy example of measuring or representing the association between two variables. So, in a unit of correlation, you have seen that there is no unit; it cancels out because this is the multiplication factor, and the value lies between minus one and 1. That is also a perfect example and a convenient representation that it depends on or varies from minus 1 to plus 1.

For example, these were the x value of x 20, 25, 30, 40, 50, 60, y value 60, 60, 70, 73, 67 and 73. Now, we have calculated the deviation of x from the mean and y from the mean, and then we calculated the square of the deviation x minus x bar. We also calculated the deviation of the square of the y minus y bar square. We multiply the deviations of x minus x mean and then the deviation of the y minus y mean, this is the value, this is the value of the square 1, this is the value for this part will be used to calculate the standard deviation. This part will be used to calculate the covariance. So the covariance is nothing but x I minus x bar multiplied by y I minus y bar, I vary from 1 to n, small n divided by n 1, n minus 1 for the sample covariance. So the covariance is several observations, which is 6, 1, 2, 3, 4, 5, 6, n minus one is 5, so the covariance is 76.5. So the variance of x is 237.5, so the standard deviation calculated for the sample is 15.41; from the variance of y, the standard deviation calculated for sy is equal to 6.41. So the correlation coefficient between these two variables, between these two variables x and y of this observation 20, 25, 30, 40, 50, 60, subsequently of the similar observations 60, 60, 70, 73, 67, and 75, so the correlation is 0.774, so it is very strongly correlated.

So these are the references, so these are the concepts we have covered in this lesson; we have calculated the variance, we have calculated the standard deviations, we have shown how to calculate the interquartile range IQR, we have interpreted the five number summary using the box plot, and we have found out the association using the scatter plot, then covariance and finally we landed on correlation. Thank you.