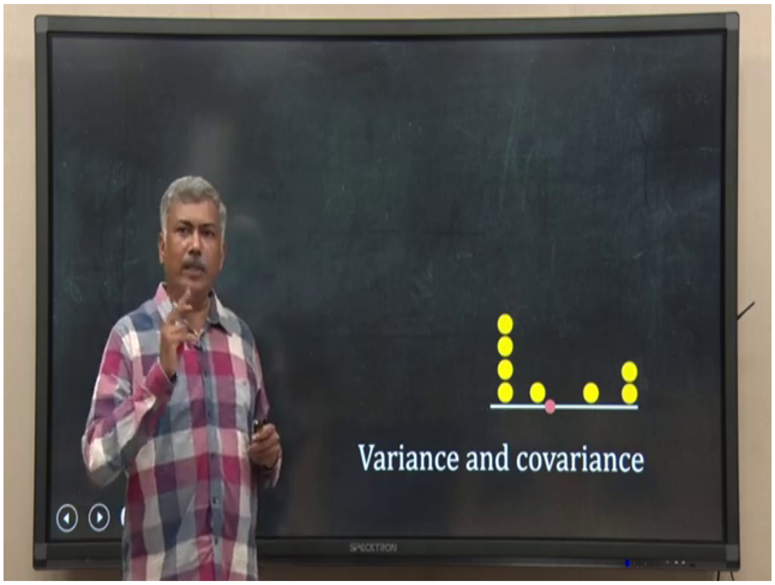
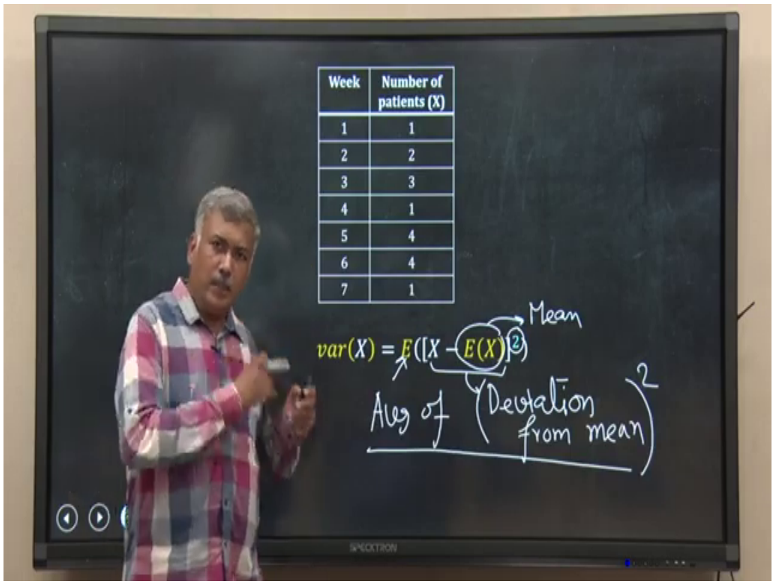
Data Analysis for Biologists
Professor Biplab Bose
Department of Biosciences & Bioengineering
Mehta Family School of Data Science & Artificial Intelligence
Indian Institute of Technology, Guwahati
Lecture: 5
Moments Variance and Covariance

(Refer Slide Time: 00: 31)



Welcome back. In the last lecture, we learned the concept of moments. And from moments we will learn about mean and variance. In this lecture, we will learn how to calculate variance for a particular set of data. And we will also learn about covariance. So, let us start. Let us start with the data set that we have used in the last lecture.

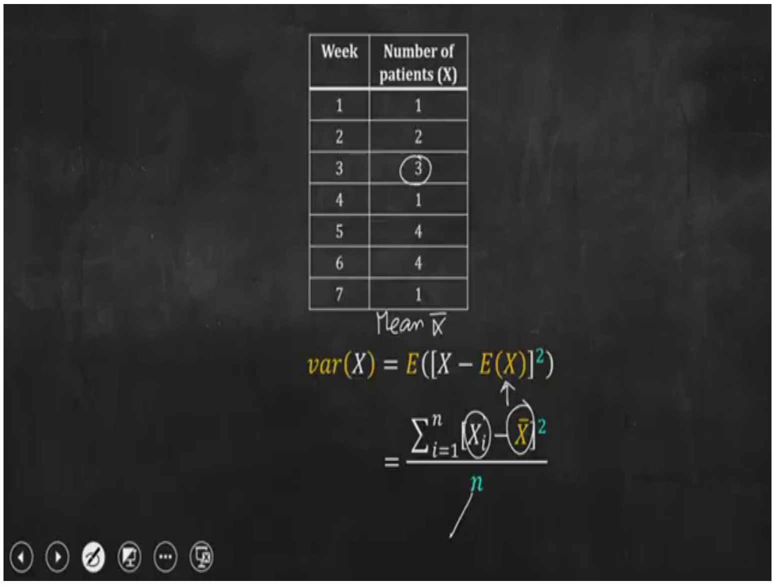
(Refer Slide Time: 00:54)



Week	Number of patients (X)
1	1
2	2
3	3
4	1
5	4
6	4
7	1

$$var(X) = E([X - \overset{\text{Mean}}{E(X)}]^2)$$

$$\underline{\text{Avg of } \left(\text{Deviation from mean}\right)^2}$$

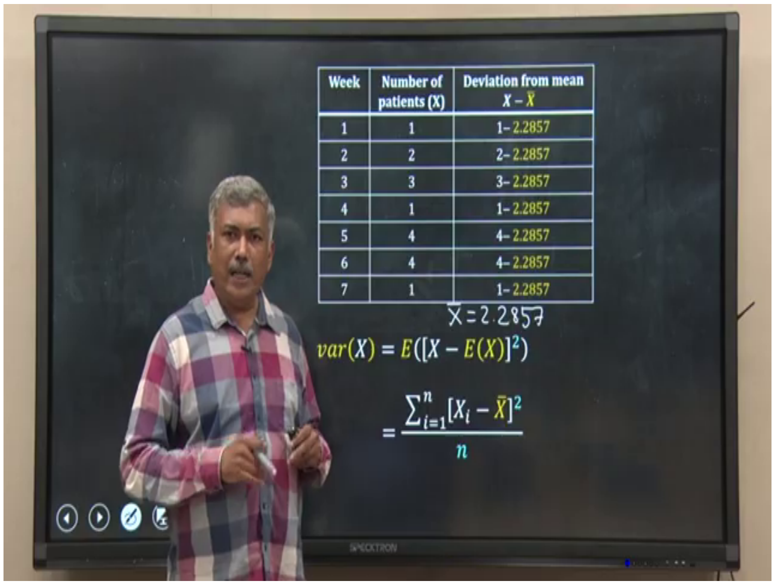


Week	Number of patients (X)
1	1
2	2
3	3
4	1
5	4
6	4
7	1

Mean $\bar{X}$

$$var(X) = E([X - E(X)]^2)$$

$$= \frac{\sum_{i=1}^{n} [X_i - \bar{X}]^2}{n}$$



Week	Number of patients (X)	Deviation from mean $X - \bar{X}$
1	1	1- 2.2857
2	2	2- 2.2857
3	3	3- 2.2857
4	1	1- 2.2857
5	4	4- 2.2857
6	4	4- 2.2857
7	1	1- 2.2857

$\bar{X} = 2.2857$

$$var(X) = E([X - E(X)]^2)$$

$$= \frac{\sum_{i=1}^{n} [X_i - \bar{X}]^2}{n}$$

So, I have data for number of patients', throat cancer patients coming, and at a time getting admitted to a hospital week wise. And for 7 weeks, I have the data, and I want to calculate the variance of this data. So, let us start with the basic definition of variance that we have learned in the last lecture. Variance is the second central moment.

So, what we have,

$$\text{var(X)} = \text{E([X – E(X)]}^2)$$

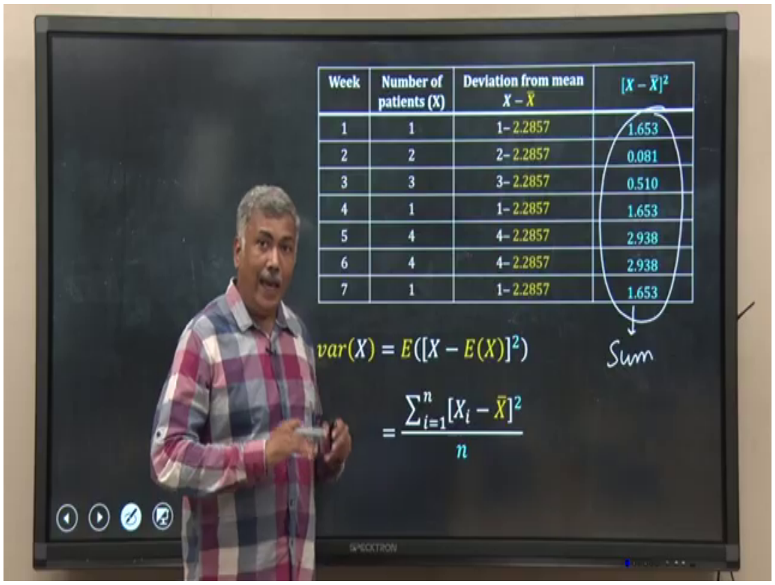
Let us break down this thing, what we have, this inner thing, this is nothing but deviation. Deviation from mean because this E(X), is nothing but the mean. And we are squaring the whole deviation.

So, I have square of that, and then I have an E outside expectation outside, that means I am taking the average of the squared deviation. So, that is my definition of variance. So, we will use this definition to calculate the variance for this particular data set. So, what I have to do, I have to obviously calculate the mean of that.

So, I need to calculate the mean of that, that I will represent by X bar, this one. So, I will replace E(X) by X bar. And X_i is one specific data, for example, this one 3 for third week. So, I calculate the deviation of that data point from the mean X bar and square the whole thing. And then I do average by dividing n here, in this case obviously n is 7, because I have 7 samples.

So, let me do the calculation, I have created one other column here, on this column, I have calculated deviation of each of these data from the mean, because if you remember, in the last class also, we said the mean is here, in this case 2 point roughly 2.2857. So, I have subtracted the individual value of each of the data from the mean of those data points.

(Refer Slide Time: 03:15)

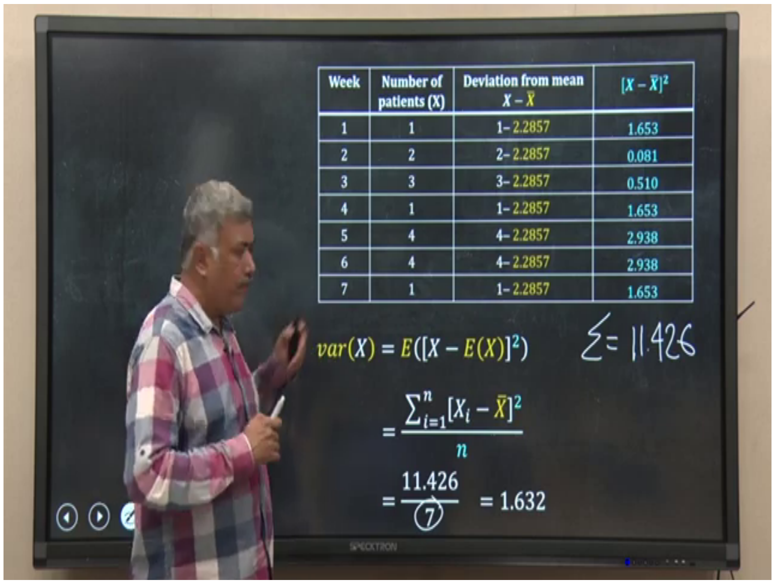


Week	Number of patients (X)	Deviation from mean $X - \bar{X}$	$[X - \bar{X}]^2$
1	1	1− 2.2857	1.653
2	2	2− 2.2857	0.081
3	3	3− 2.2857	0.510
4	1	1− 2.2857	1.653
5	4	4− 2.2857	2.938
6	4	4− 2.2857	2.938
7	1	1− 2.2857	1.653

Sum

$$var(X) = E([X - E(X)]^2)$$

$$= \frac{\sum_{i=1}^{n} [X_i - \bar{X}]^2}{n}$$



Week	Number of patients (X)	Deviation from mean $X - \bar{X}$	$[X - \bar{X}]^2$
1	1	1− 2.2857	1.653
2	2	2− 2.2857	0.081
3	3	3− 2.2857	0.510
4	1	1− 2.2857	1.653
5	4	4− 2.2857	2.938
6	4	4− 2.2857	2.938
7	1	1− 2.2857	1.653

$\Sigma = 11.426$

$$var(X) = E([X - E(X)]^2)$$

$$= \frac{\sum_{i=1}^{n} [X_i - \bar{X}]^2}{n}$$

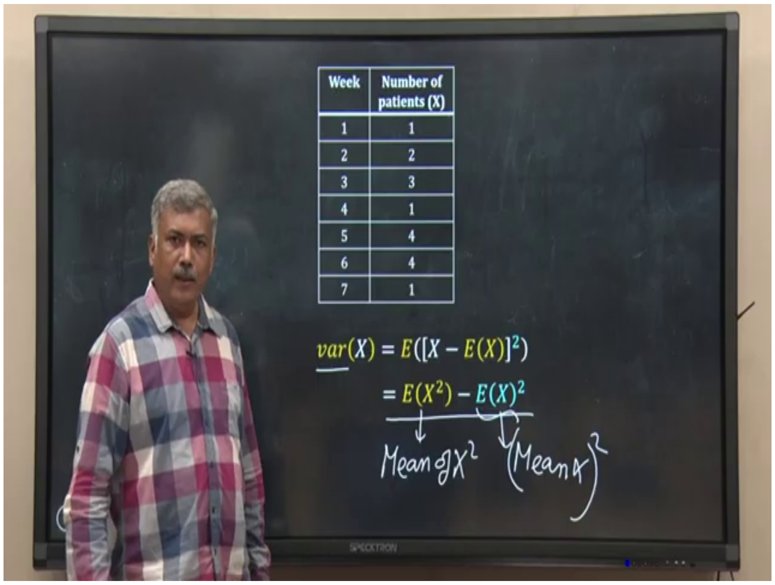
$$= \frac{11.426}{7} = 1.632$$

And then I squared, I add another column and square them, then I sum all these values, the square values, and I sum them. So, by summing them, I get a particular value. And then I will divide that by 7 because I have 7 samples. So, that is what I do

$$= \frac{\Sigma\,[Xi - Xbar]^2}{n}$$

Roughly 11.426 and I divide that by 7, because I have 7 samples. So, this is the average of square of deviation from mean for all the data in my table and that comes to 1.632, you can calculate the same variance in a different way also.

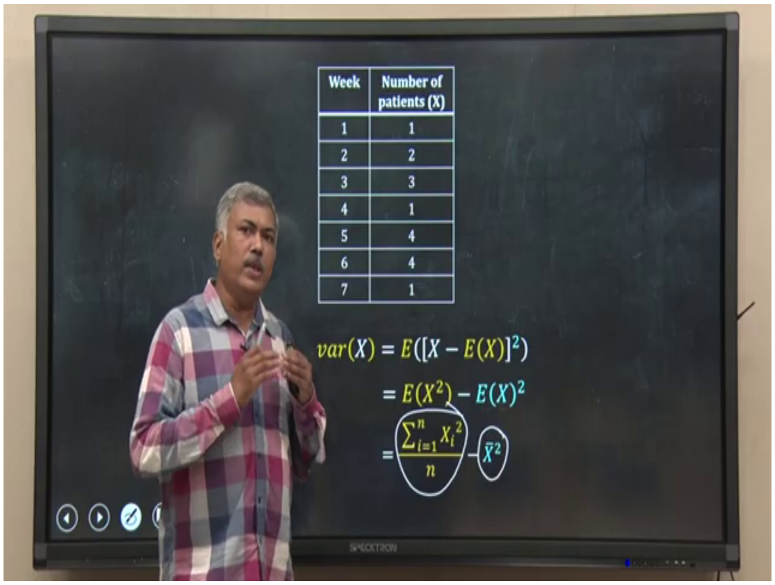
(Refer Slide Time: 04:02)



Week	Number of patients (X)
1	1
2	2
3	3
4	1
5	4
6	4
7	1

$$\underline{var}(X) = E([X - E(X)]^2)$$

$$= \underbrace{E(X^2)}_{\text{Mean of } X^2} - \underbrace{E(X)^2}_{(\text{Mean } X)^2}$$

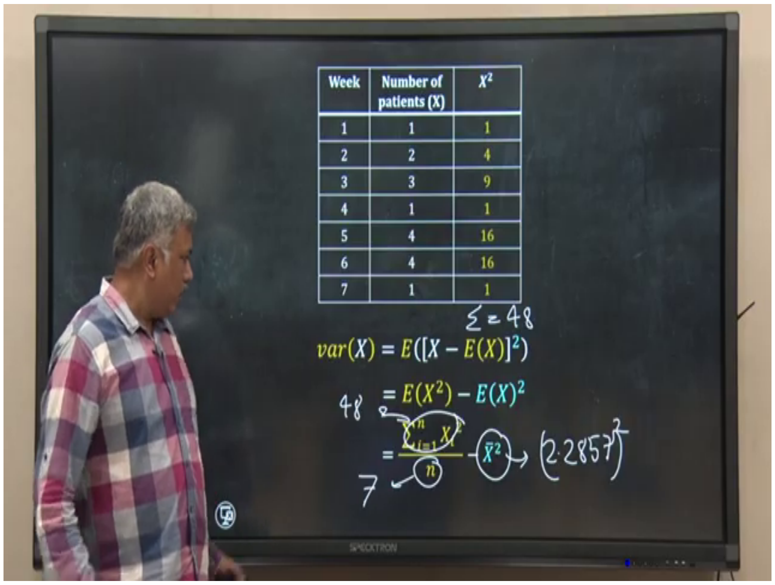


Week	Number of patients (X)
1	1
2	2
3	3
4	1
5	4
6	4
7	1

$$var(X) = E([X - E(X)]^2)$$

$$= E(X^2) - E(X)^2$$

$$= \frac{\sum_{i=1}^{n} X_i^2}{n} - \bar{X}^2$$



Week	Number of patients (X)	X^2
1	1	1
2	2	4
3	3	9
4	1	1
5	4	16
6	4	16
7	1	1

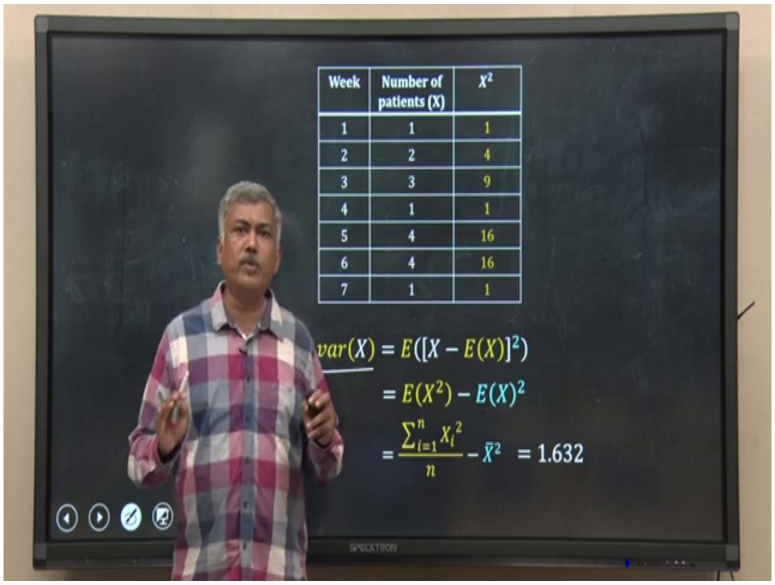
$\Sigma = 48$

$$var(X) = E([X - E(X)]^2)$$

$$48 \quad = E(X^2) - E(X)^2$$

$$= \frac{\sum_{i=1}^{n} X_i^2}{n} - \bar{X}^2 \rightarrow (2.2857)^2$$

7



Week	Number of patients (X)	X^2
1	1	1
2	2	4
3	3	9
4	1	1
5	4	16
6	4	16
7	1	1

$$var(X) = E([X - E(X)]^2)$$

$$= E(X^2) - E(X)^2$$

$$= \frac{\sum_{i=1}^{n} X_i^2}{n} - \bar{X}^2 \quad = 1.632$$

If you remember, when we were discussing about the moments and the rules derived from this moment concept, we have said the variance can be written in this fashion also,

$$\mathrm{var}(X) = E(X^2) - E(X)^2$$

What is this expectation of X square, is nothing but the average mean of X square.

So, this is mean of X square and what is this, this is nothing but mean of my data and we are squaring that, because we have a squared term there. So, I can use this formula also to calculate the variance of the same data set. So, what I have to do, mean I already knows that is Xbar

2.2857 and square that, I calculate the mean or average of X^2, here. So, I take the square of each of these data points and sum them and then divided by number of data points that is 7.

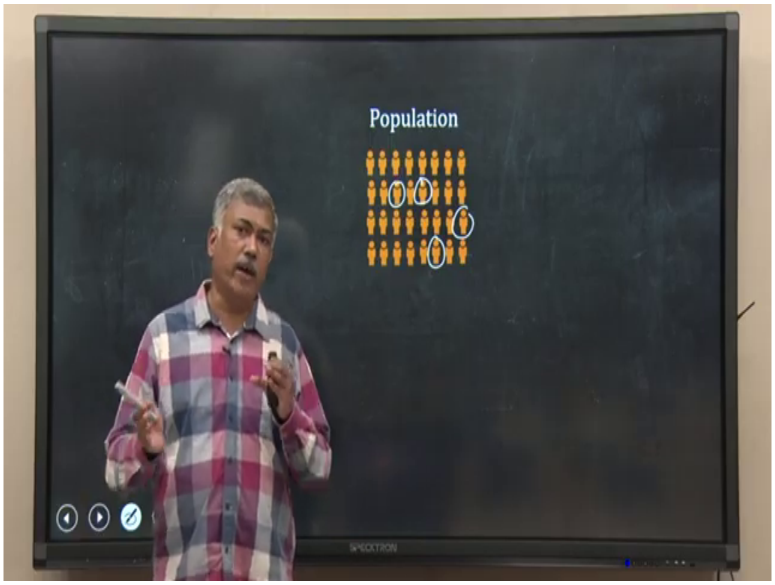
So, what I have to do, I have to calculate the X^2. So, I have calculated the square of each of these data points. So, when it is 2, X^2 is 4, when it is 3, X^2 is 9, something like that, and I sum them and most probably the sum will be 16 plus 16, 32 and it will be 48. So, what I get, I get 48 here, what is n, n is 7.

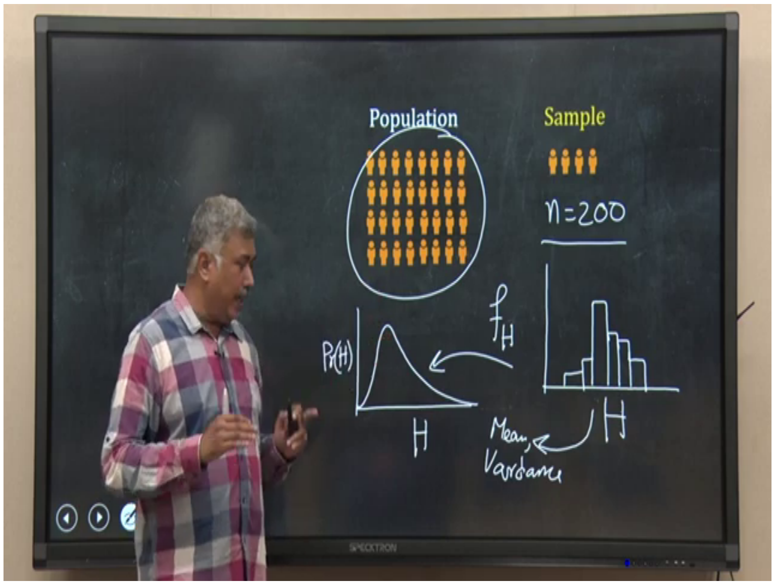
What is this one, this one is 2.2857 whole square if you remember, we have calculated the mean, the Xbar earlier, the square of that. So, if you now add all this things, subtract up mean square also, add up all this thing, what you will get, you will get the same variance 1.632. Now, you can use either of this technique.

Either of these two, formulas whatever you choose to do and that will give you the variance of a data set. But there is a catch, it can be said, it can be shown mathematically that this variance, this variance, that way I have calculated here, from this data set of 7 samples, n = 7 here, sample number of sample is 7 here, the way I have calculated this variance, this variance is not a correct estimate of the variance of the population.

That is a mouthful word, isnt it? We have not discussed what is variance of population or variance of samples, something like that earlier. So, let me discuss what do I mean by population variance and why this variance is actually not a correct estimator of that population variance. To understand that we have to understand the difference between what is called population and what is called sample in probability theory and statistics.

(Refer Slide Time: 07:02)





Let us take an example. Suppose I want to know the average height of children going to primary school in a particular region, suppose in a particular state of the country and suppose in that state, there are 2 lakh primary school going children. Now, obviously, you cannot go to house of these 2 lakh children or school of each of these students and measure their height and then give me average value and the variance of that obviously, you cannot do that, that does not make sense.

So, what you will do, you will take some of those students, judiciously, we will plan so essentially, you may take around the height, measure the height of around suppose 200 student judicially picked up, for example, you may maintain, I will maintain the ratio of girls and boys equal, equal. So, 100 boys, 100 girls and you sample from different region of the state because, regional variation can be there and then you measure the height of each of these children. And that is what you will call a sample. So, the sample in this example may be, suppose I have 200 students. Now, from this sample data, what you want to know?

Are you really interested to know the height of these students whom you have sampled or you actually, in generalized way you want to know, what is the average height of primary school going children in the particular state? So, you want the generalized answer. So, you want to know the population behavior.

But you do not have the population data, you have the sample data. So, that means, everything that you calculate from the sample data is actually an some sort of estimate what will happen in the population, what is the true picture, what is the real values? And you really do not know what the real value is; you really do not know what is the real average height of students in the whole population.

We do not know that, we want to predict it; we want to estimate it from our sample, which is just 200 students. So, remember the mean, variance that we discussed in the last class, we discussed, we derived them from the idea of moments, and moments are related to probability distribution. So, this population, in this population of 1 lakh or 2 Lakh School going children in a particular state, if I plot like this height in the horizontal axis.

And probability in the, probability of height in the vertical axis, maybe there is a probability distribution for height among these children. From this probability distribution, if you tell me the probability distribution, then I can use the definition of mean and variance from moments and can calculate very easily the mean and variance of the population.

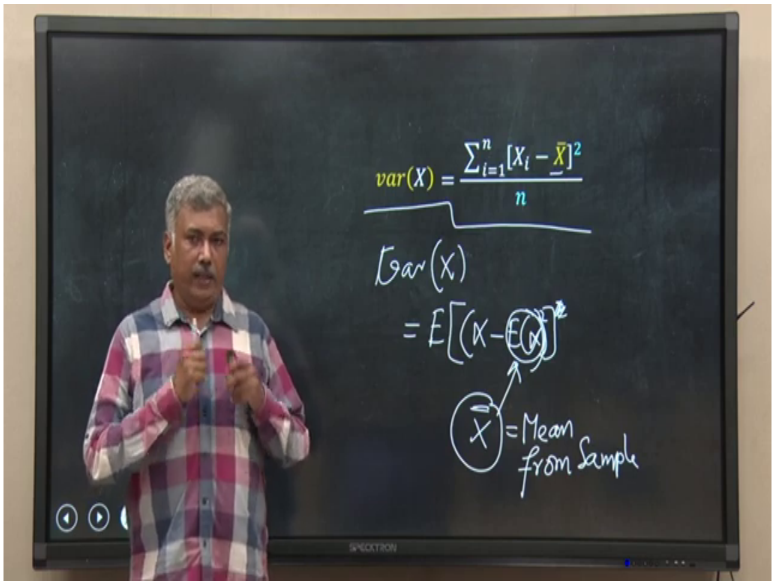
But I do not know this population distribution. I do not know what is the population distribution of height in this particular population? I may assume something, I may assume that the population has a height distribution which follows normal distribution something like that I can assume from my education, from I can make an educated guess.

But we do not know really what is this population distribution, what I know, I have the data for 200 students from that I can make a plot something like this, H in the horizontal axis and the frequency of observing a particular height in the vertical axis and then I can represent the data maybe in suppose some bar plot, something like this.

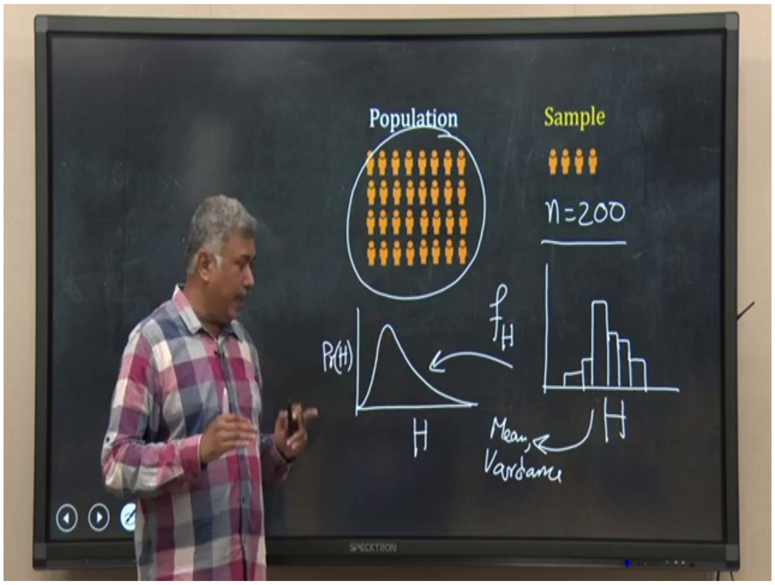
So, from this data, from this data I want to calculate the mean and variance, which are defined by probability distribution. So, I have a frequency data so, you have data from n sample, 200 Students sample, from that you want to understand the behavior of the population. So, obviously there will be your common sense obviously, there will be some amount of error.

So, there will be some amount of deviation from the real value of mean and variance of the population because we do not know the real probability distribution of height in the population, let us take the special case of variance and see rather than going into details of mathematics that how much will be the variation and why it will not be a real estimate of the population variance, the variance of the height of this particular population, let us take a approximate way of understanding where our variance estimate is going wrong.

(Refer Slide Time: 11:55)



$$var(X) = \frac{\sum_{i=1}^{n}[X_i - \bar{X}]^2}{n}$$



So, if you remember we defined variance of X as,

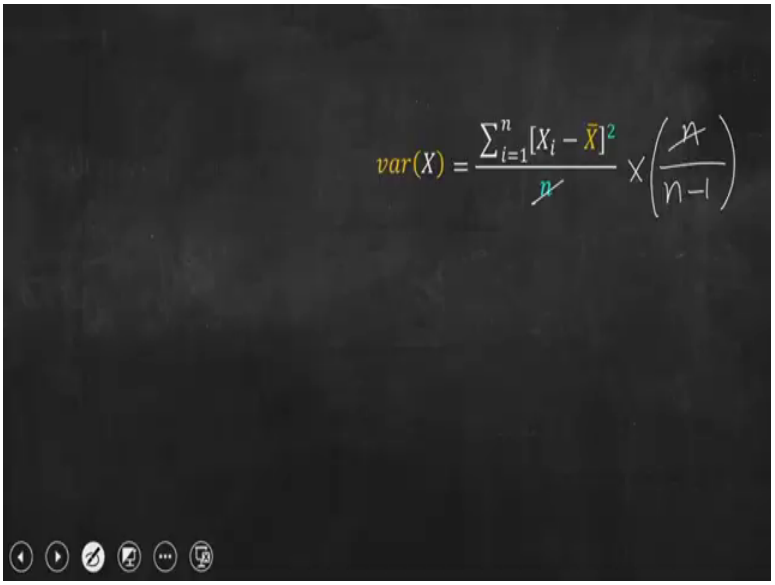
$$\mathrm{var}(X) = E([X - E(X)]^2)$$

Sorry, let me put the square inside not here. So, what we are doing, we really do not know this E(X), because we do not know the population distribution. So, we do not know the population distribution.

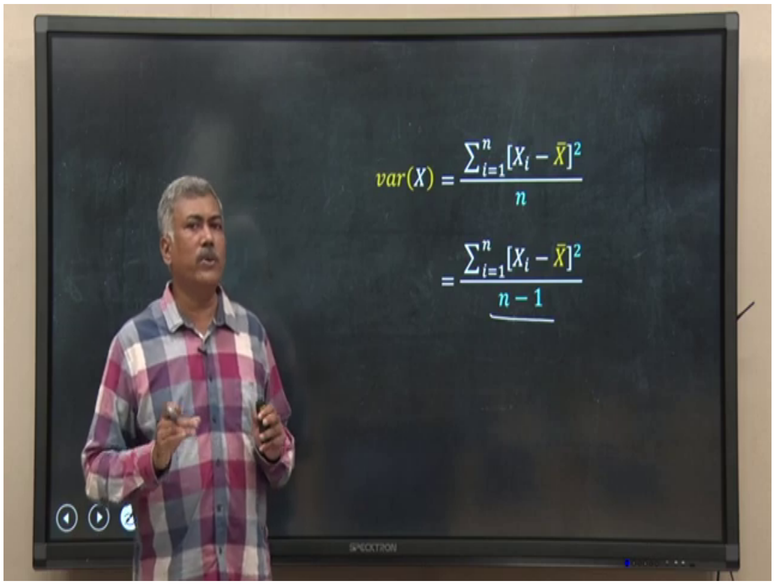
So, we do not know the exact expectation of X, X is the height here, what I will do in my calculation, I replace this by Xbar, the mean from the sample, that is what we have done here. So, we are replacing E the expectation of X, E(X) by Xbar, which is the mean from the sample and as we know that mean will not be exactly same as the expectation of X, because we do not know the population distribution of height.

So, in that case, this mean will be slightly different, obviously, it is slightly different from the expectation of X. So, every time I am putting this Xbar here, I am actually incorporating an error in my calculation. So, it can be shown mathematically that this definition, this definition of variance is actually under estimating the variance of the population. So, we have to correct it. So, how should I correct it, let me clean the board a bit to show the correction.

(Refer Slide Time: 14:00)



$$var(X) = \frac{\sum_{i=1}^{n}[X_i - \bar{X}]^2}{\cancel{n}} \times \left(\frac{\cancel{n}}{n-1}\right)$$



$$var(X) = \frac{\sum_{i=1}^{n}[X_i - \bar{X}]^2}{n}$$

$$= \frac{\sum_{i=1}^{n}[X_i - \bar{X}]^2}{\underline{n-1}}$$

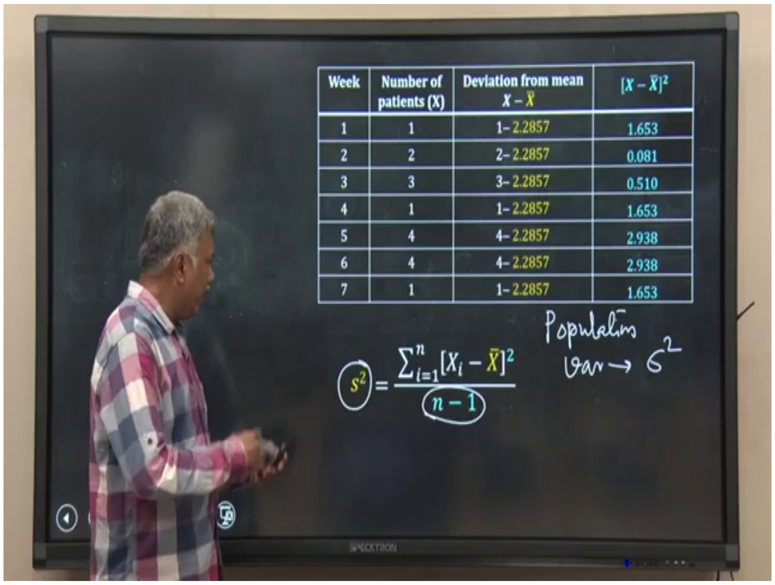
So, the way to correct it is known as using the bessel correction, I will not go into details of how it has been derived, let us just use the correction, what we will do remember we are, by this particular formula, we are under estimating the population variance, the variance in the population.

So, what you have to do, you have to put a bessel correction factor here, which is nothing but n by (n-1). So, I will multiply our defined variance with (n/n – 1) and as I have n both in the numerator and denominator, I will cancel them. So, what I will get? I will get,

$$= \frac{\Sigma\,[Xi - Xbar]^2}{n-1}$$

So, you are taking the deviation for each data point from the sample mean, that Xbar is the sample mean, and squaring that and then summing them and then you are dividing by (n – 1), not n. So, this is called the bessel correction for variance. And we are doing this, just to repeat it, we are doing this because our, the earlier definition of variance from the sample data does not represent the population variance adequately, that is biased estimate. So, we are un-biasing it. So, now, let us try to use this formula, modified formula, the corrected formula to calculate the variance of our original data, the data for patients coming to the, getting throat cancer patient getting admitted to the hospital.

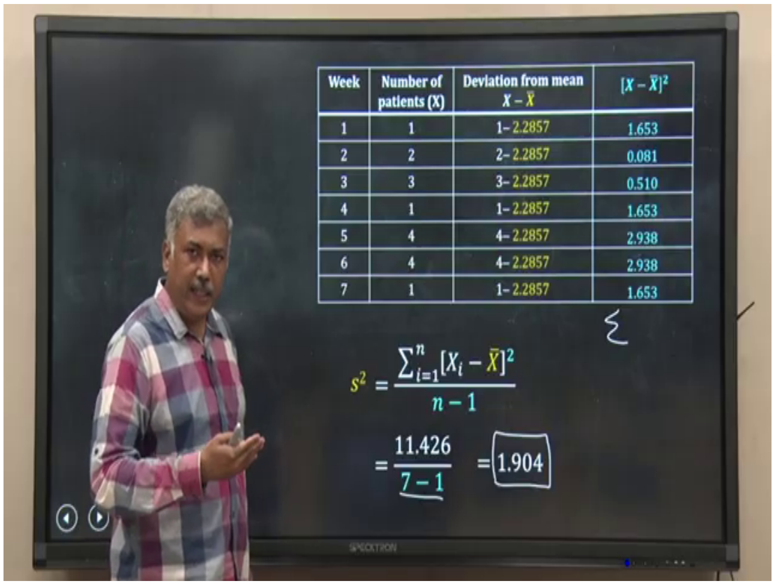
(Refer Slide Time: 15:36)



Week	Number of patients (X)	Deviation from mean $X - \bar{X}$	$[X - \bar{X}]^2$
1	1	1– 2.2857	1.653
2	2	2– 2.2857	0.081
3	3	3– 2.2857	0.510
4	1	1– 2.2857	1.653
5	4	4– 2.2857	2.938
6	4	4– 2.2857	2.938
7	1	1– 2.2857	1.653

$$s^2 = \frac{\sum_{i=1}^{n}[X_i - \bar{X}]^2}{n-1}$$

Population
var → σ^2



Week	Number of patients (X)	Deviation from mean $X - \bar{X}$	$[X - \bar{X}]^2$
1	1	1– 2.2857	1.653
2	2	2– 2.2857	0.081
3	3	3– 2.2857	0.510
4	1	1– 2.2857	1.653
5	4	4– 2.2857	2.938
6	4	4– 2.2857	2.938
7	1	1– 2.2857	1.653

Σ

$$s^2 = \frac{\sum_{i=1}^{n}[X_i - \bar{X}]^2}{n-1}$$

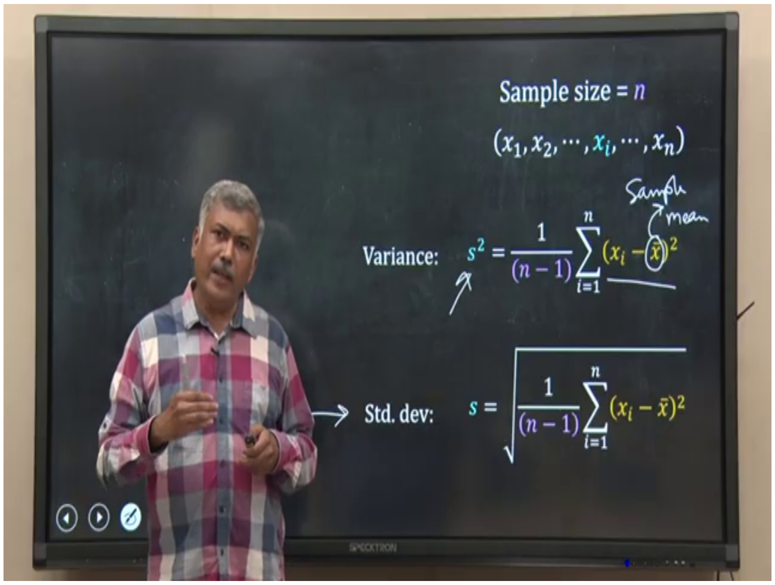
$$= \frac{11.426}{\underline{7-1}} = \boxed{1.904}$$

So, this is my data, the first column is week, second column is the number of patient then the deviation from the mean, mean is 2.2857 and the square of those deviation. What I will do I will use the same formula except this one will be, the denominator will be (n – 1) and I will not write explicitly variance will write it s^2.

So, in general in statistics textbooks usually s represents the standard deviation from sample. So, s^2 is the variance of the sample, the variance you are calculating from the sample not for the population, for population variance people will use the symbol σ^2. So, now, let us do the calculation.

So, I will sum these values, I will sum these values all those values, and that is 11.426 and I will divide by (7 - 1), 6 because (n – 1) = 6 and I will get 1.904. So, you can see here the variance has increased. So, as I was saying earlier that our earlier definition of variance was actually under estimating the population variance, this one has a higher value. So, this is called the corrected or unbiased sample variance.

(Refer Slide Time: 17:08)



Sample size = n

$(x_1, x_2, \cdots, x_i, \cdots, x_n)$

Sample mean

Variance: $s^2 = \frac{1}{(n-1)} \sum_{i=1}^{n} (x_i - \bar{x})^2$

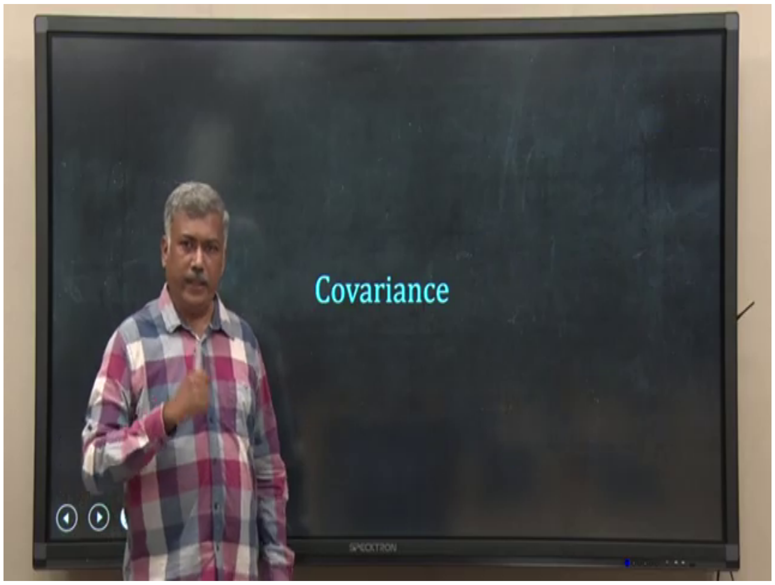
Std. dev: $s = \sqrt{\frac{1}{(n-1)} \sum_{i=1}^{n} (x_i - \bar{x})^2}$

So, let us generalize, I have a sample size of n, in the earlier example, I took the sample of 200 or suppose in the throat cancer patient sample I have 7 samples. For each of the sample I have a measurement X1 X2 Xi up to, Xn. So, the variance should be, the corrected variance will be, s^2 will be,

$$s^2 = \frac{\Sigma\,(Xi - Xbar)^2}{n-1}$$

Xbar is sample mean then you sum them up and then you divide by (n – 1), standard deviation is nothing but the square root of variance. So, I simply take the square root of this s^2 and I get the standard deviation from my sample. Now, let us try to understand something else which is related to variance, it is called covariance.

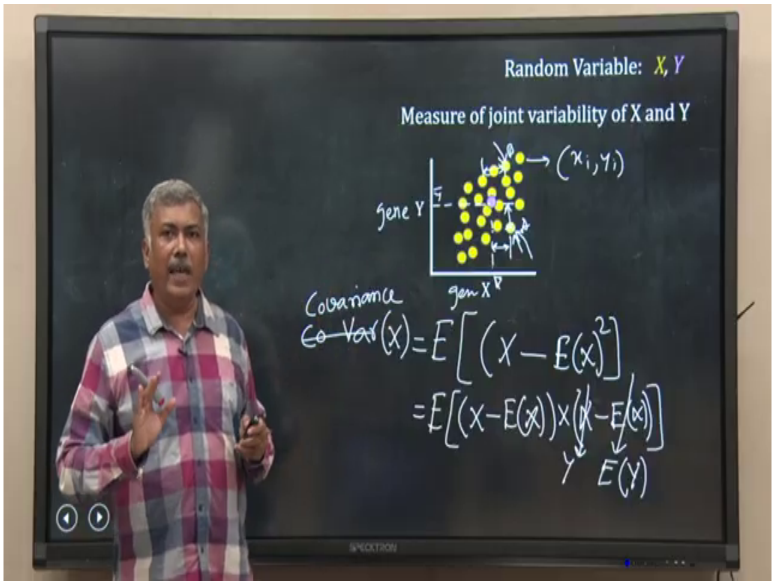
(Refer Slide Time: 18:04)



Covariance

Now, what is covariance?

(Refer Slide Time: 18:07)



Random Variable: *X*, *Y*

Measure of joint variability of X and Y

Covariance

$$\text{~~Co-Var~~}(X) = E\left[\left(X - E(X)\right)^2\right]$$

$$= E\left[(X - E(X)) \times (\underset{Y}{\cancel{X}} - \underset{E(Y)}{\cancel{E(X)}})\right]$$

Let us try to define it first, till now I was discussing about one single random variable like for example, the height of the students or the number of patient, throat cancer patient getting admitted to the hospital. Only one random variable, one variable. Now, suppose I have two variables X and Y. For example, suppose you are measuring the expression of two genes, in a tumor sample collected from 100 patients. So, you have measurement for two genes, obviously, both the genes will not get a expressed an equal amount in all the tumor samples, there will be variation.

So, expression of both the genes, are two different random variable for example, X and Y. So, now, I want some sort of measurement, which will give me the joint variability between X and Y. Now, what do I mean by this joint variability? Let us check that first with the example. So, the same thing, suppose X is a gene Y is another gene and you are measuring their expression in some tumor samples collected from a patient using quantitative PCR.

So, each of this data points in my plot is one sample and it has a expression of Xi and Yi values, pairwise values. So, that is a particular patient, that circle is a particular patient and that patient's tumor has Xi amount of relative amount of expression of X and Yi relative amount of expression of Y.

So, I have suppose 20 samples and I have shown them my 20 circles, and this pink one is the mean value, simple arithmetic mean. So, I will draw, so this is Xbar. And this one is Ybar, mean of Y. So, you can calculate the variance of X and Y, separately, just now you have learned, for example, suppose take this particular sample, and then the deviation for this sample is the one from the mean of X. You can square that deviation.

And for each data point, you can do that and then sum them and divide by (n – 1) to get the corrected sample variance. For X, particular gene, gene X. You can do the same thing for Y. For example, I can take for the same sample from same sample, I can take the deviation from Y, so, this is the deviation from Y, and I can square that, then I can do the same thing for all other data points and sum them together divided by (n – 1), that will give me the variance of my Y data.

But I want something which will give me a hint about how both X and Y are vary. For example, take this data point, here the data point is on the right-hand side of the X so, I have positive deviation. I have another data point here, which also have a positive deviation in X. But check those two different data points. Suppose let us make it A and this one is B. So, A and B, both X are positive deviation, X as positive deviation.

But for Y, A has negative deviation from its mean. And for B, Y has positive deviation from its mean. So, how can I combine all this information, deviation in both the direction to create a measure, which will give me joint variability? Let us take the help of variance, what is the definition of variance?

$$\mathrm{var}(X) = E([X - E(X)]^2)$$

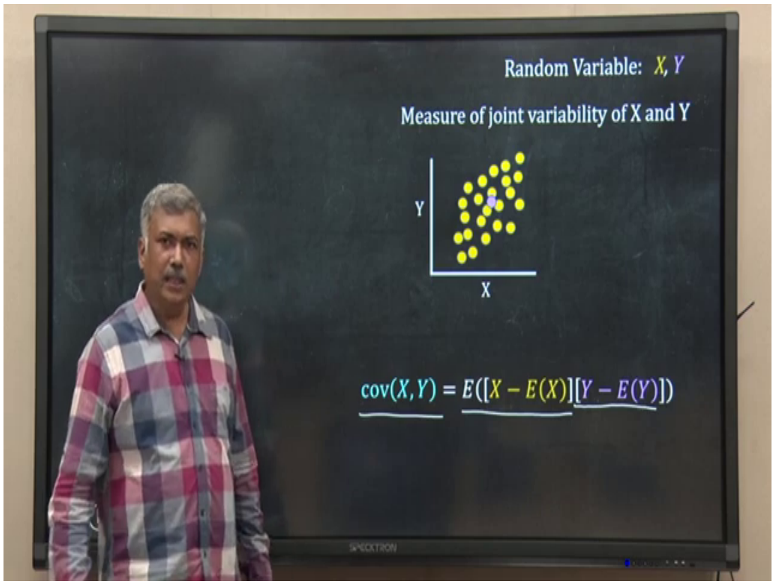
Let me write down these explicitly, I can write down,

$$[X - E(X)]^2 = [X - E(X)] * [X - E(X)]$$

I am multiplying twice, so, that is equivalent to square. So, now this is a same thing as the definition of variance we have, now what if I replace this X by Y and I replace this one, E(X) by E(Y) and then I will call that covariance.

So, I have replaced one set of X thing with Y things, I have replaced one set of X – E(X) with the Y – E(Y), E(Y) is the mean of Y. So, now, I have combined deviation from of X from its mean and deviation of Y from its mean. So, both direction, deviation in both the horizontal and the vertical direction of this graph has been incorporated in my definition, and now, I do not call it variance I call it covariance.

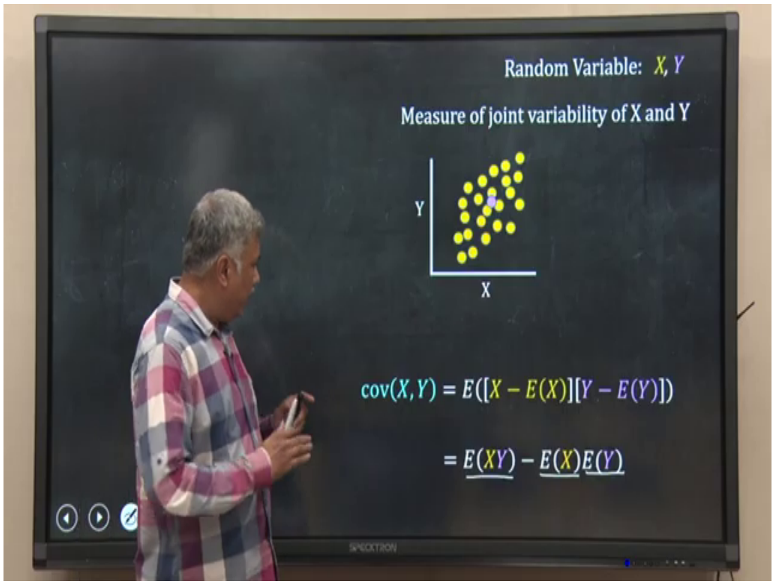
(Refer Slide Time: 23:50)



Random Variable: X, Y

Measure of joint variability of X and Y

$$\underline{\mathrm{cov}(X,Y)} = E(\underline{[X - E(X)]}\underline{[Y - E(Y)]})$$



Random Variable: X, Y

Measure of joint variability of X and Y

$$\mathrm{cov}(X,Y) = E([X - E(X)][Y - E(Y)])$$

$$= \underline{E(XY)} - \underline{E(X)}\underline{E(Y)}$$

So, covariance by definition,

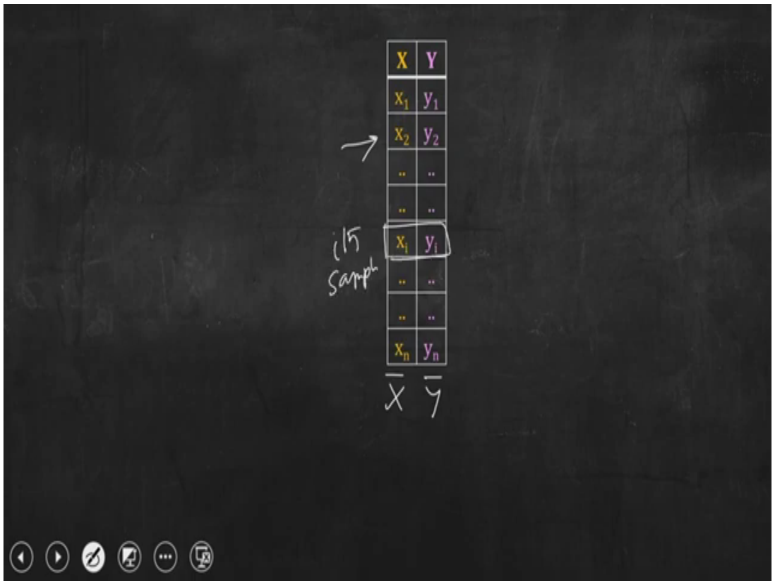
$$\mathrm{cov}(X,Y) = E([X – E(X)][Y – E(Y)])$$

where expectation is the mean of X and Y. So, now, you can actually show, you can do some algebra and use the probability theory and you can show that it can be written as,

$$= E(XY) – E(X)E(Y)$$

So, now, we can use this definition of covariance to calculate covariance in a data set.

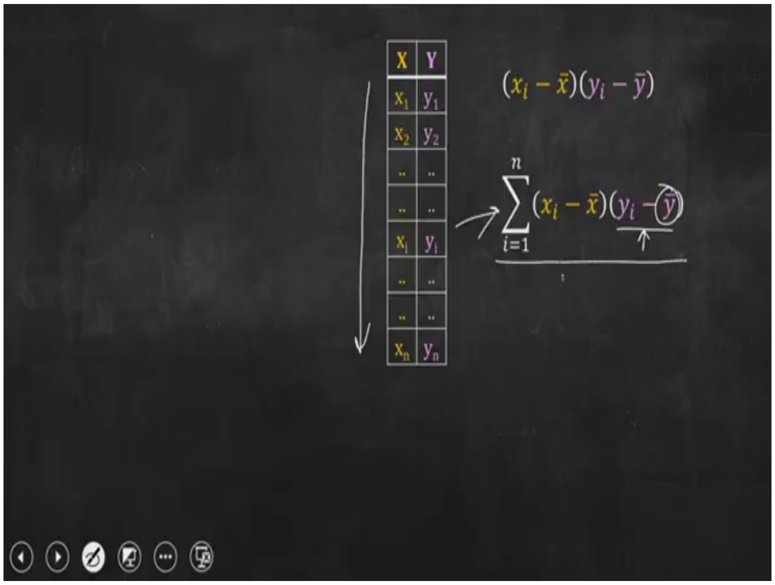
(Refer Slide Time: 24:31)



X	Y
x_1	y_1
x_2	y_2
..	..
..	..
x_i	y_i
..	..
..	..
x_n	y_n

ith sample

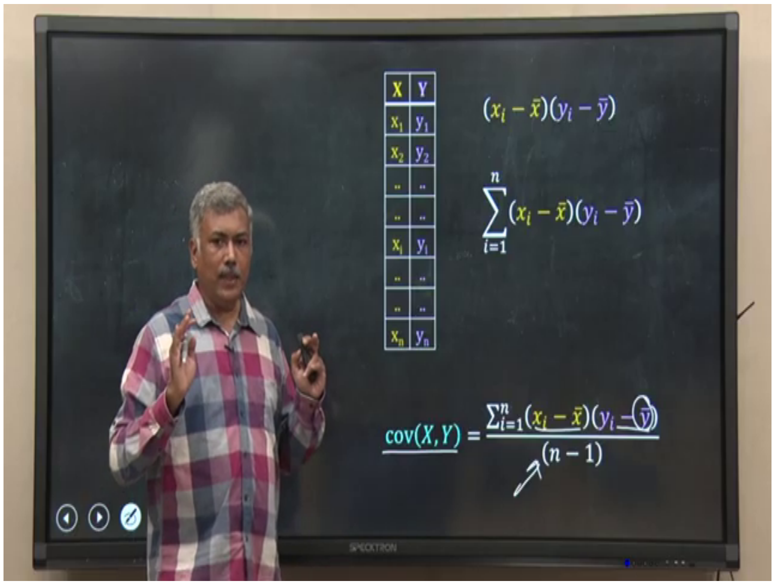
$\bar{X}$ $\bar{Y}$



X	Y
x_1	y_1
x_2	y_2
..	..
..	..
x_i	y_i
..	..
..	..
x_n	y_n

$$(x_i - \bar{x})(y_i - \bar{y})$$

$$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$



X	Y
x_1	y_1
x_2	y_2
..	..
..	..
x_i	y_i
..	..
..	..
x_n	y_n

$$(x_i - \bar{x})(y_i - \bar{y})$$

$$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

$$\underline{\mathrm{cov}(X,Y)} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)}$$

For example, suppose I have this data set again the data for two genes X and Y and I have n samples. So, for each n sample I have pair wise data, for example, for the second sample I have expression of X is x2, expression value of Y is y2 something like that. So, for this data set I can calculate the sample mean it will be Xbar remember it is sample mean not population mean it is sample mean.

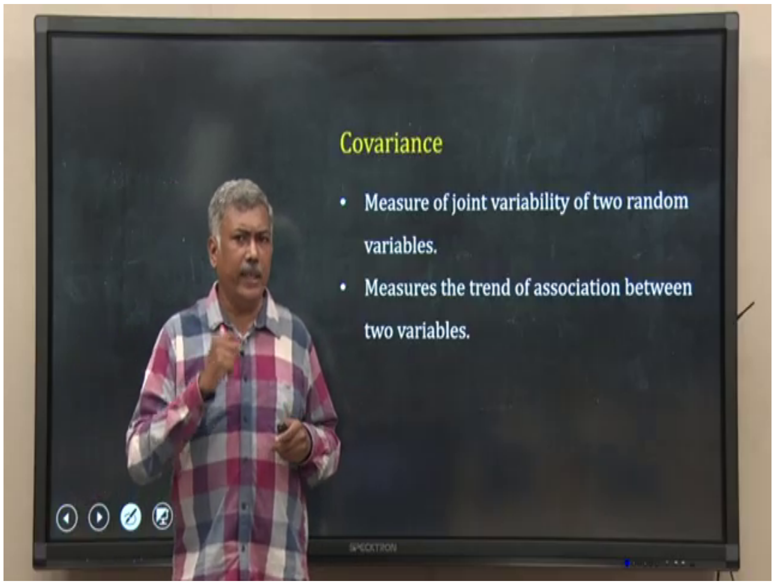
Similarly, I can calculate the sample mean for Y from this data set, then for each of this sample, suppose, if I take this ith sample, this is the ith sample, I can calculate the deviation of X from the mean and deviation of Y from the mean, its mean and then I can multiply them. Then I can do the same thing for all the data points, all the samples and then sum them together.

So, this is what I have done here, I have calculated the deviation of X for a sample from the sample mean of X and multiplied that with the deviation of Y in that sample from the sample mean of Y and then I have summed all these multiplication product for all these in data point. Now, I will do a averaging of that, I will do a averaging of that means, I have to divide these by sample number n.

Now, again we have to remember this is nothing but variance, we have come from the definition of variance. So, here also I have to do bessels correction to make it unbiased otherwise, if I have used a division by n, then it will underestimate the covariance between X and Y, it will not give me the correct representation of the population covariance.

So, what I have to do, I have to divide by (n – 1) not n. So, what we have got, covariance of two random variable X and Y is nothing, but what you do you take Xi minus X bar, Xbar is the sample mean for X into Yi minus Y bar, Ybar is the sample mean for Y and then you take this product and sum them together and divide by (n – 1).

(Refer Slide Time: 27:06)



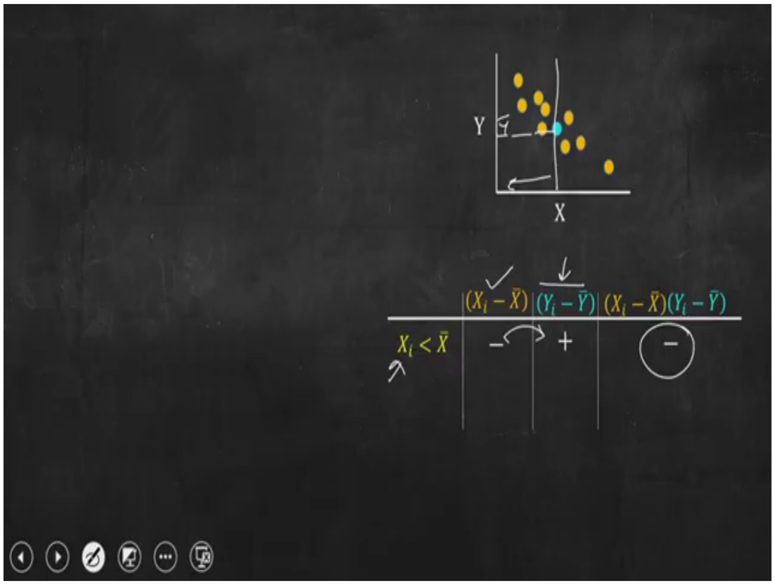
Covariance

- Measure of joint variability of two random variables.
- Measures the trend of association between two variables.

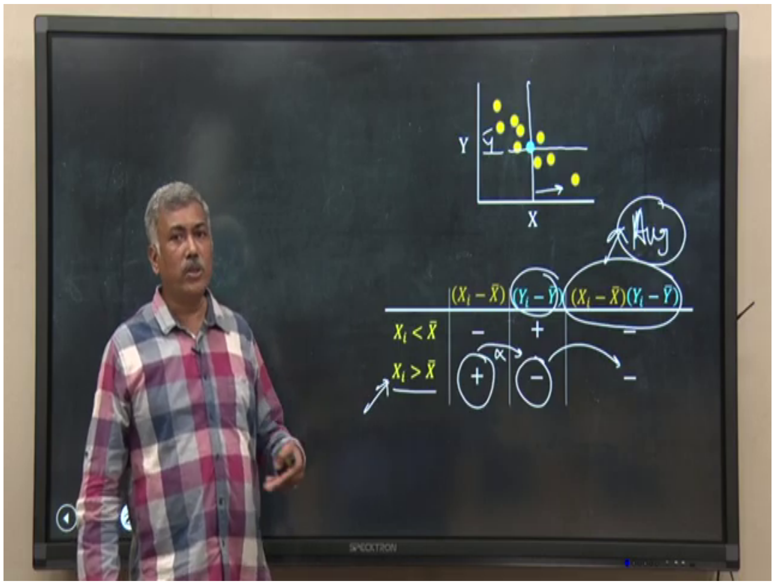
So, now, look back, what we are doing it covariance. The first thing I started discussing is that covariance is some sort of measure of joint variability of two random variables, but there is something more to this, in fact, covariance is a sort of measure of trend of association between two variables. What do I mean by association, for example, suppose you are doing gene expression study, these two gene X and Y may be controlled by the same transcription factor.

So, when that transcription factor is induced then both X and Y should increase. So, that means, there is a association between X and Y or suppose X is itself is a transcription factor when it increases it inhibits expression of Y. So, then there should be opposite relation between X and Y. So, you want this type of sort of you want to know the trend in association from your data and covariance can actually help in that. So, let us look into one you know hypothetical data to understand this.

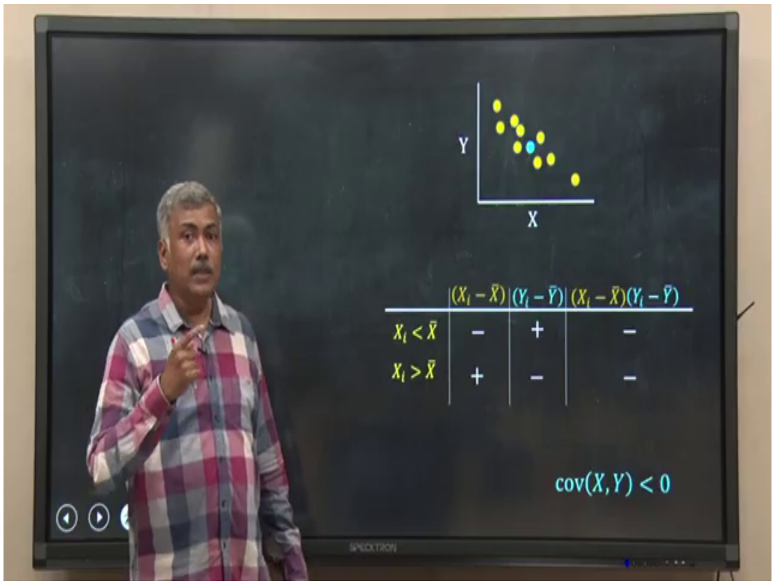
(Refer Slide Time: 28:12)



	$(X_i - \bar{X})$	$(Y_i - \bar{Y})$	$(X_i - \bar{X})(Y_i - \bar{Y})$
$X_i < \bar{X}$	−	+	−

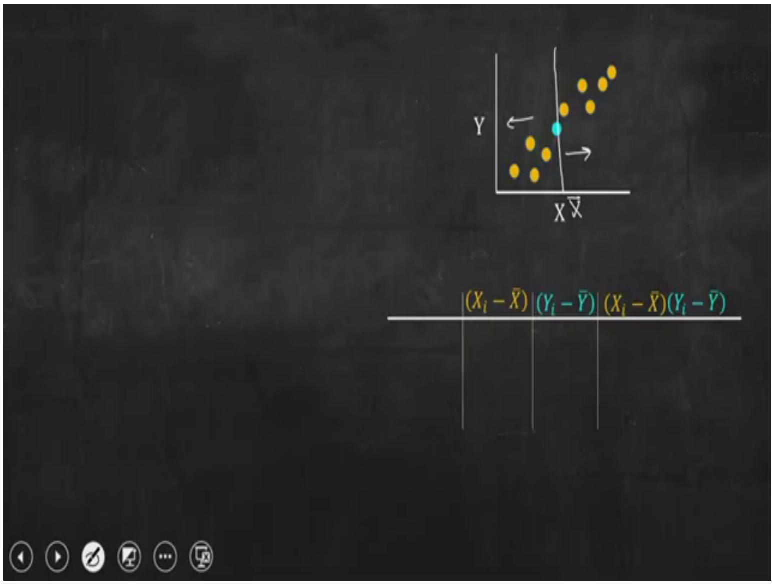


	$(X_i - \bar{X})$	$(Y_i - \bar{Y})$	$(X_i - \bar{X})(Y_i - \bar{Y})$
$X_i < \bar{X}$	−	+	−
$X_i > \bar{X}$	+	−	−

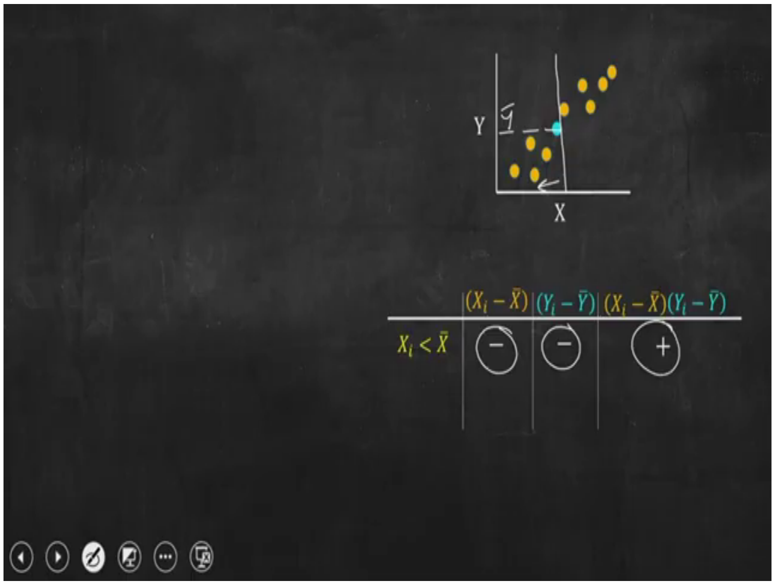


	$(X_i - \bar{X})$	$(Y_i - \bar{Y})$	$(X_i - \bar{X})(Y_i - \bar{Y})$
$X_i < \bar{X}$	−	+	−
$X_i > \bar{X}$	+	−	−

$\text{cov}(X,Y) < 0$



	$(X_i - \bar{X})$	$(Y_i - \bar{Y})$	$(X_i - \bar{X})(Y_i - \bar{Y})$



	$(X_i - \bar{X})$	$(Y_i - \bar{Y})$	$(X_i - \bar{X})(Y_i - \bar{Y})$
$X_i < \bar{X}$	−	−	+

Suppose X and Y has an inverse opposite relationship that means, when X increases Y drops, so, the graph will look like this, the yellow points are the sample data and the blue point is the mean. So, I can mark like this, this is my mean Xbar and this is my Ybar sample mean. Now, I will do some rough calculation here, what I will do let me take a straight line at Xbar.

Let me mark this again as Ybar. So, think about data, which are on the lower side of Xbar means on this side. So, on this side of the line that I have drawn through the mean, the blue dot, see in all these cases, X, the value of X for a particular dot will be less than Xbar, because they are on this side, towards the 0 side of the mean. So, here what will happen Xi any data point which is on this side will be less than X bar.

So, that mean Xi minus Xbar will be a negative value for most of the values, isnt it? Similarly, if I calculate the Yi minus Ybar, deviation from the mean of Y for those, so, here I have the mean of Y. So, for most of those data, for most of those data, Y bar will be smaller than their individual Y value. So, my Yi minus Ybar will be positive.

So, now in covariance, we multiply Xi minus Xbar in with Yi minus Ybar. So, the multiplication for most of this data points on this side of mean of X will always have something minus because you are multiplying minus with plus, so, you will get a negative value. What will happen if I take data on the other side of mean of X this is mean of X.

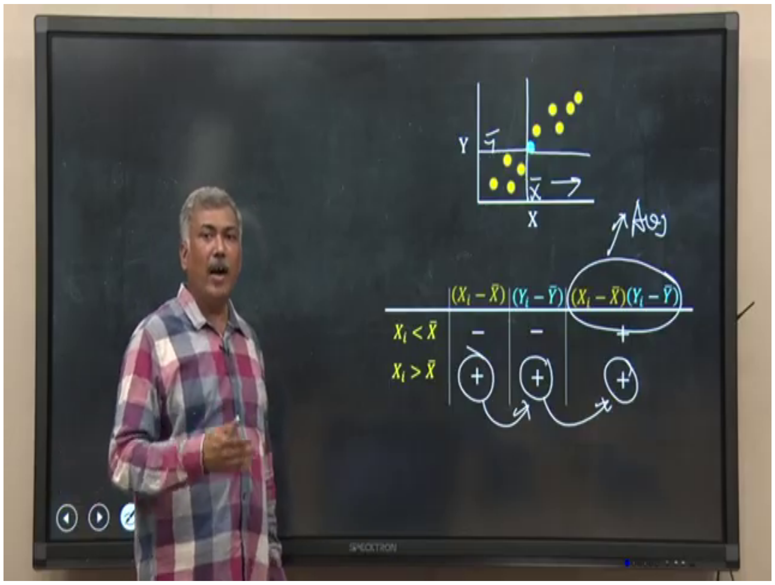
So, in this case, in this case, Xi, the individual value of X for individual data point is bigger than Xbar, obviously, Xi minus Xbar will be positive, but for most of this data, for most of this data on this side, this is my Ybar, Y value will be lesser than bar, so, this one Yi minus Ybar will be negative. So, I have a positive number, I have a negative number, if I multiply them, if I multiply these two, I should get a negative number.

So, now, I have covered both sides of the mean of X. So, I have covered all the data points to calculate the covariance what I have to do I have to make a average of these multiplication calculate the mean of that, and for most of the data point that multiplication product is a negative value. So, obviously, the average will be also negative and that average is nothing but covariance.

So, that means, for this data covariance will be less than 0, covariance will be negative. Now, let us look into the other behavior, where when X increases, Y also increases. So, in this data almost there's a linear trend, so, when X increases, Y also increases. So, in this case, again I can do the same thing I can draw a line, this is my X bar, and I will see first on this direction, and then I will check into this direction. First to this direction.

So, here Xi, individual value of X is less than Xbar. So, that means Xi minus Xbar will be negative for most of the data points, Yi minus Ybar will again be negative, you can see this is my Ybar. So, multiplication of these two negative things will give me a positive value. Now, check the other side.

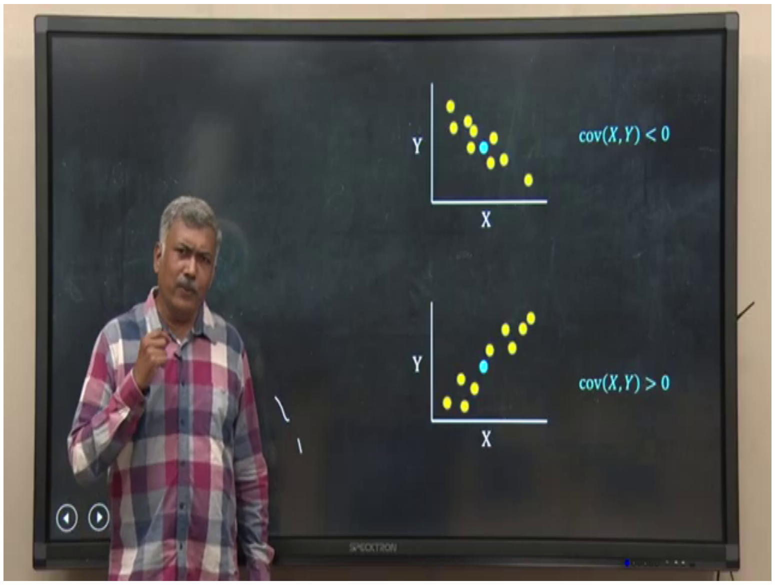
(Refer Slide Time: 32:58)



	$(X_i - \bar{X})$	$(Y_i - \bar{Y})$	$(X_i - \bar{X})(Y_i - \bar{Y})$
$X_i < \bar{X}$	−	−	+
$X_i > \bar{X}$	+	+	+

So, I am checking on this side in this case, X minus Xbar is positive Y minus Ybar is also positive, you see, the values are above the Ybar. So, again positive, multiplication of these two positive thing will give me a positive product. Now, again by definition of covariance, what I will do, I will take the average of this. So, average of lots of positive values will give me a positive value. So, that means, in this case, when X and Y both are increasing, at the same time, my covariance will be bigger than 0, mean it will be positive. So, let me jot down.

(Refer Slide Time: 33:53)



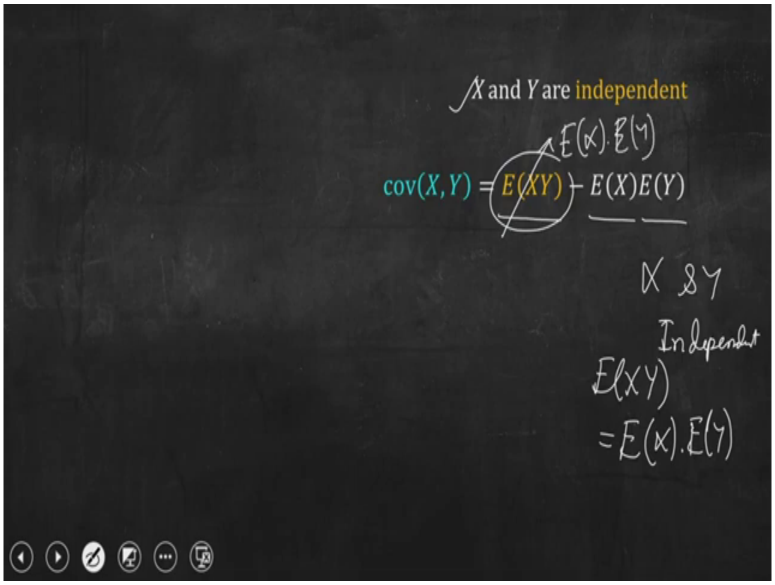
$\mathrm{cov}(X,Y) < 0$

$\mathrm{cov}(X,Y) > 0$

What we discussed in these two example, when there is a opposite inverse trend, when X increases Y decreases, Y increases, X decreases something like that, then covariance between X and Y would be less than 0 it will be negative. Whereas, Y and X if they are varying in the same fashion in the same direction, same way and then covariance of X and Y will be bigger than 0. Now, you must have noticed something extra here.

In the last lecture, we had discussed variance will be always positive, isnt it, because variance is average of some square values, but covariance can be both negative and positive, because we are multiplying two different thing, (X - Xbar) into (Y – Ybar). So, depending upon their values, I can get a negative covariance, I can also get positive covariance. Now, let me check another particular behavior of covariance.

(Refer Slide Time: 35:03)



X and Y are independent

$$\mathrm{cov}(X,Y) = \overset{E(X).E(Y)}{\cancel{E(XY)}} - \underline{E(X)}\,\underline{E(Y)}$$

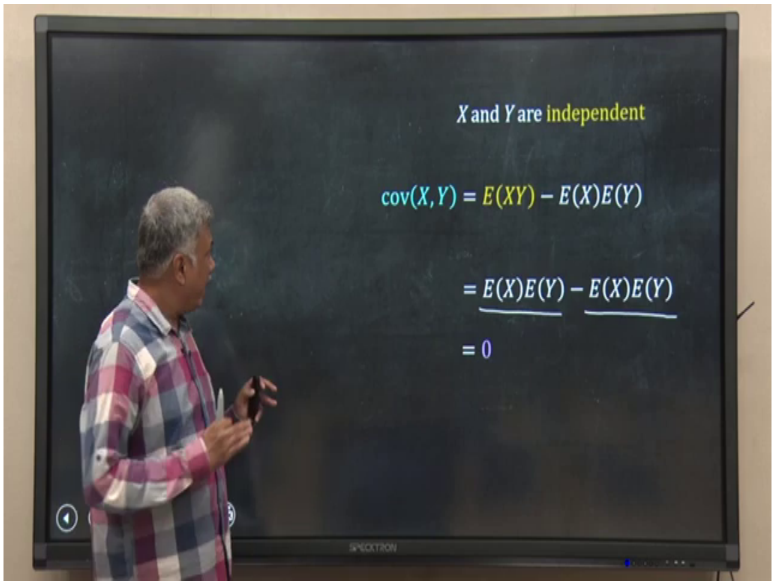
X & Y Independent

$E(XY)$
$= E(X).E(Y)$

Till now, we were assuming that we are expecting that there is some sort of association between two random variable, gene X and gene Y expression, some sort of association, but suppose these two are completely independent, there is no association between them, then how the covariance should behave, let us look into the definition and try to give a answer to this.

So, covariance if you remember, we have shown a short formula that covariance will be equal to E(XY) – E(X)E(Y). So, if you remember the last lecture, we have discussed some generalized rule derived from the definition of first moment and second central moment for mean and variance when different relation between random variables exist. So, in that lecture, we had said if X and Y are independent then E(XY) is nothing but E(X)*E(Y), you may check the last lecture. We have discussed this. So, now I can replace this when I have already assumed that they are independent, so I can replace this one by expectation of X into expectation of Y. So, let me write down it cleanly.

(Refer Slide Time: 36:41)

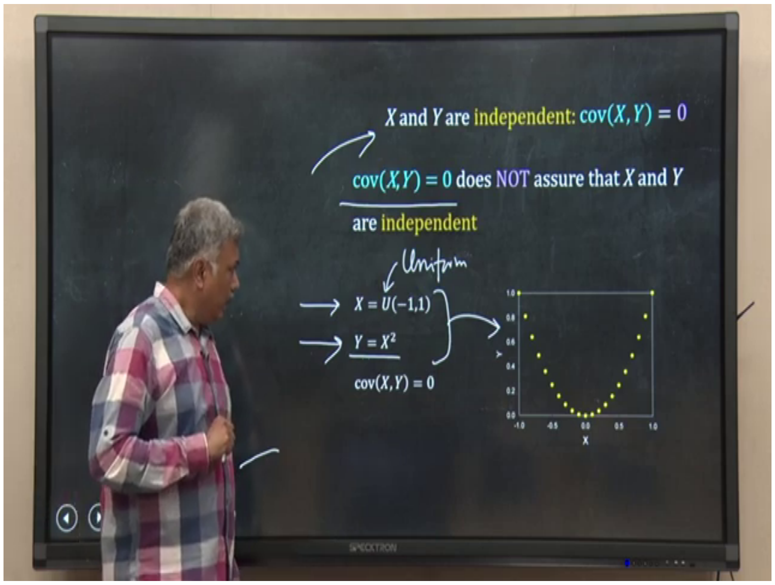


X and Y are independent

$$\mathrm{cov}(X,Y) = E(XY) - E(X)E(Y)$$

$$= \underline{E(X)E(Y)} - \underline{E(X)E(Y)}$$

$$= 0$$



X and Y are independent: $\mathrm{cov}(X,Y) = 0$

$\underline{\mathrm{cov}(X,Y) = 0}$ does NOT assure that X and Y are independent

Uniform

$\rightarrow X = U(-1,1)$

$\rightarrow \underline{Y = X^2}$

$\mathrm{cov}(X,Y) = 0$

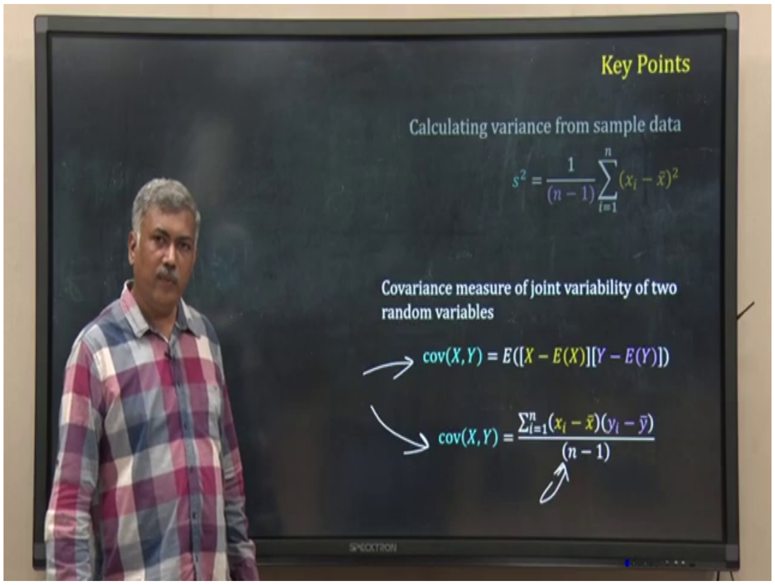
So, I have E(X)E(Y) – E(X)E(Y), so I will get 0. So, that means when X and Y are independent random variables, the covariance between them will be 0. But there is a catch in this, the reverse of this statement is not true. What do I mean by reverse of the statement, that means when covariance of X and Y equal to 0, that does not assure me that X and Y will be independent. This is a bit tricky, but we have to keep in mind. X and Y are independent, then covariance will be 0. But given that you have told me that covariance between X and Y equal to 0 that does not assure me that X and Y are independent.

I will explain this without going into detail of the mathematical derivation, I will take a textbook example and show you how it is true. Suppose X is a uniformly distributed random number, from -1 to 1, that is what you have written, (-1,1), whereas Y = X^2.

So, Y depend upon value of X, If X equal to 1, Y equal to 1, if X equal to 0.5, Y equal to 0.25, so X and Y are not independent. And if I plot this, the data will look like this, the way I have shown here, the yellow dots. Now, if you calculate the covariance for this data, you will find covariance is 0. So, we have shown earlier that X and Y, when they are independent covariance become 0.

But in this case, X and Y are not independent, Y depends on X^2, Y equal to X^2, but still covariance is equal to 0. That means, although when X and Y are independent covariance is 0, that is for sure, but just giving that telling me that covariance equal to 0 does not assure that the X and Y are independent, we have to keep this thing in mind, this tricky thing in mind.

(Refer Slide Time: 39:03)



Key Points

Calculating variance from sample data

$$s^2 = \frac{1}{(n-1)}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

Covariance measure of joint variability of two random variables

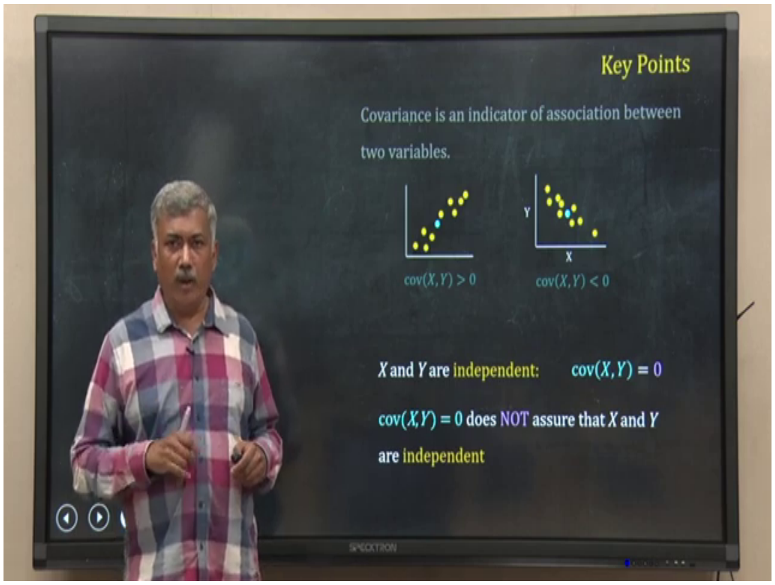
$$\mathrm{cov}(X,Y) = E([X-E(X)][Y-E(Y)])$$

$$\mathrm{cov}(X,Y) = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{(n-1)}$$

So, now, let me jot down what we have learned in this lecture. In this lecture, we have learned how to calculate the variance of a data, in that we have learned about bessel's correction, and we have to use bessel's correction where will replace the denominator n by (n – 1), so that our sample variance becomes a better predictor of the population variance.

Then we have learned about covariance, covariance is nothing but the joint variability between two random variables, we have given the definition of the covariance and then we have seen the method to calculate the covariance when data for two variables are given. And again, you have to remember here that we have to do the bessel's correction, (n – 1) in the denominator in place of n.

(Refer Slide Time: 39:56)



Key Points

Covariance is an indicator of association between two variables.

X and Y are independent: $\text{cov}(X,Y) = 0$

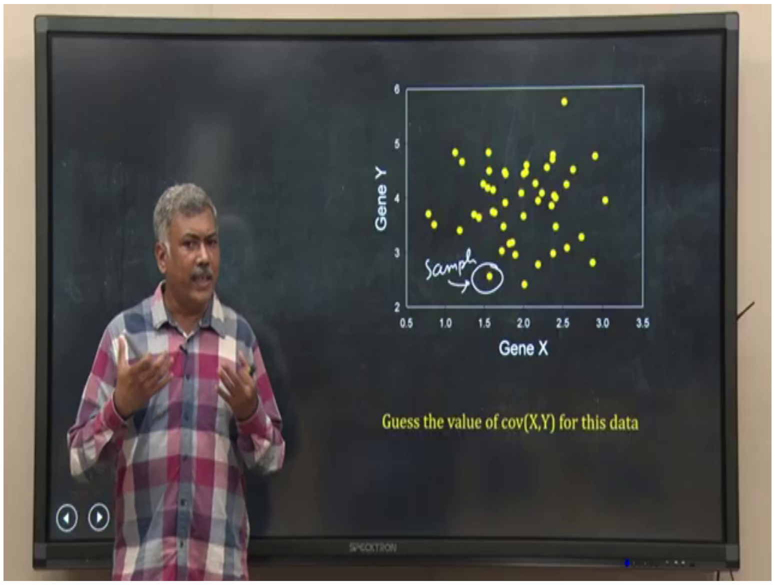
$\text{cov}(X,Y) = 0$ does NOT assure that X and Y are independent

We have learned further that apart from giving the joint variability, covariance to some extent tells us the trend of association between two variables. For example, if there is a positive trend, that means X and Y both increases simultaneously both decreases simultaneously, then a covariance will be positive.

If they have a opposing behavior, that means when X increases, Y decreases something like that, then covariance will be negative. But we have to remember here, actually, covariance is not a good measure to check the association between two variable. There are something else, for example, correlation coefficient, which we will discuss in other lectures, those are much better measure to find association between two variables.

But still, covariance gives us a broad idea. Lastly, we have discussed that if X and Y are two independent random variable, two independent measurements, then the covariance between X and Y will be equal to 0. But here we have to keep the tricky thing in mind that if covariance is, between X and Y is 0, that does not assure us that X and Y are independent. That is all for this lecture. Before we end, let me give you a problem to think about.

(Refer Slide Time: 41:20)



Guess the value of cov(X,Y) for this data

Suppose I have collected the gene expression data for two gene, gene X and gene Y from tumor samples. And suppose I have 50 tumor samples. 50 patients, each of these dots is one sample, one patient tumor. I do not want you to do lots of calculation, lots of math, just intuitively look into this data and try to guess the value of covariance between gene X and gene Y in this data.

Again, I do not want you to do exact calculation, because I am just giving you a graphical representation, the numerical data are not there. So, I do not want you to give a exact value of covariance, looking at the data, and by the idea of covariance that we have learned in this lecture. Can you guess, what will be the ballpark value of covariance of between X and Y in this case? Will they be very positive and high?

Will they be very positive and negative and very low? Will they be 0? Will it be very close to 1 or something like that? I want you to guess that. So, why do not you try it? Try to guess what will be the ballpark value of covariance, X and Y. That is all for this lecture. Happy Learning. See you in the next lecture. Thank you.