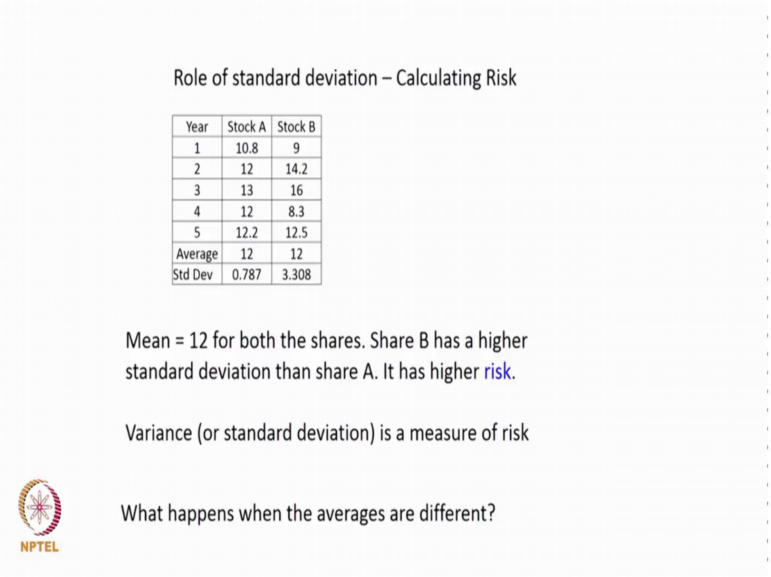
Introduction to Probability and Statistics
Prof. G. Srinivasan
Department of Management Studies
Indian Institute of Technology, Madras

Lecture – 07
Describing Numerical Data (continued)

In this lecture, we continue the discussion on Describing Numerical Data.

(Refer Slide Time: 00:24)



Role of standard deviation – Calculating Risk

Year	Stock A	Stock B
1	10.8	9
2	12	14.2
3	13	16
4	12	8.3
5	12.2	12.5
Average	12	12
Std Dev	0.787	3.308

Mean = 12 for both the shares. Share B has a higher standard deviation than share A. It has higher risk.

Variance (or standard deviation) is a measure of risk

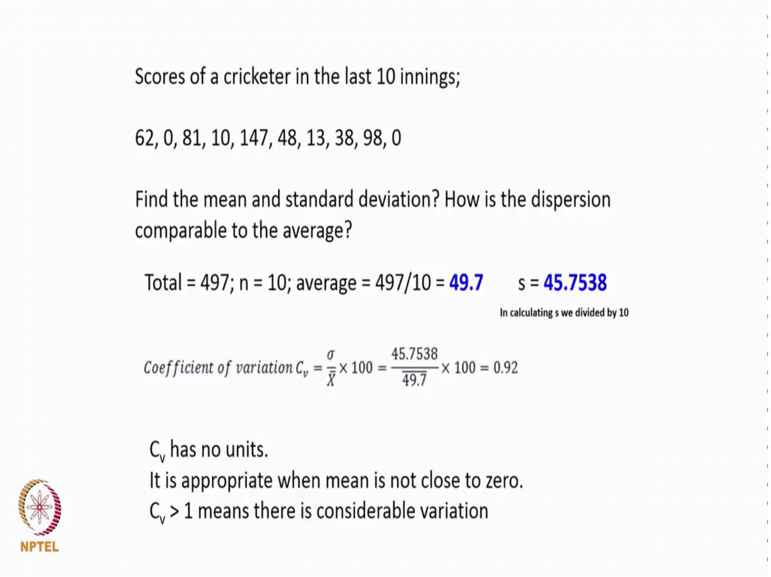
What happens when the averages are different?

In the previous lecture, we were looking at this slide and studying standard deviation and using standard deviation as a measure of evaluating risk. So, we gave the example of two stocks whose returns over 5 years are given here we call them stock A these are the 5 year returns and stock B, these are the 5 year returns. We find the average and in this case we observe that both stock A and stock B have the same average of 12 percent return.

Now, we compute the standard deviation using the formula that was discussed in the earlier lecture and we observed that stock A has standard deviation of 0.787 while stock B has a higher standard deviation of 3.308. So, share B has a higher standard deviation than share A and therefore, we can assume or conclude that share B or stock B has a higher risk than that of stock A.

So, variance or standard deviation can be used as a measure of evaluating risk. Now, in this example we had a situation where the averages were the same and therefore, when the averages were the same we said that that particular stock which has a higher standard deviation or variance has higher risk. Now, what happens when the averages are different?

(Refer Slide Time: 02:11)



Scores of a cricketer in the last 10 innings;

62, 0, 81, 10, 147, 48, 13, 38, 98, 0

Find the mean and standard deviation? How is the dispersion comparable to the average?

Total = 497; n = 10; average = 497/10 = **49.7** s = **45.7538**

In calculating s we divided by 10

$$\textit{Coefficient of variation } C_v = \frac{\sigma}{\bar{X}} \times 100 = \frac{45.7538}{49.7} \times 100 = 0.92$$

C_v has no units.
It is appropriate when mean is not close to zero.
$C_v > 1$ means there is considerable variation

Now, in order to understand that let us look at another example and try to observe this. So, let us assume that these are the 10 scores of a cricketer in the last 10 innings. So, these numbers are given 62, 0, 81, 10 and so on. First let us find out the mean or the average and the standard deviation and let us answer this question how is the dispersion comparable to the average. So, the total of these 10 scores is 497. So, n is equal to 10 the number of observations. The average is 497 divided by 10 which is 49.7. We also compute a standard deviation of 45.7538 and in this case we have divided it by 10. Now, the average is 49.7 the standard deviation is 45.7538. The question before us is how is the dispersion comparable to the average?

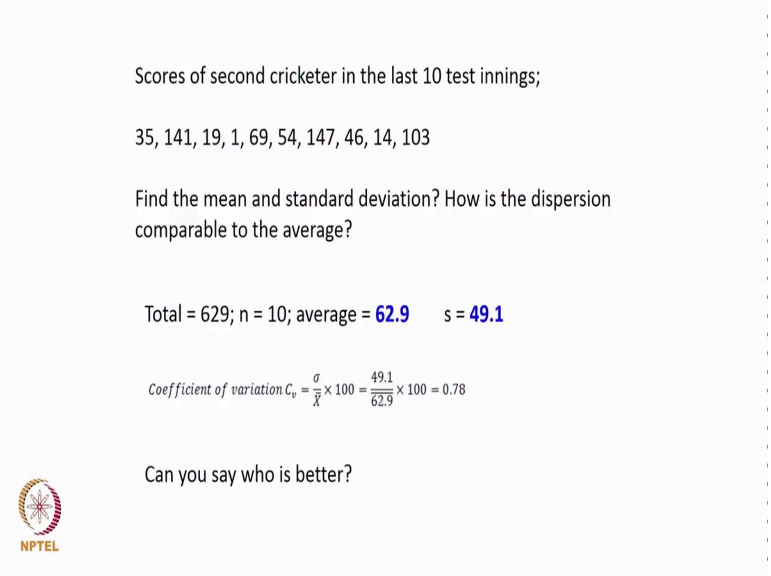
So, we now use standard deviation as a measure of dispersion and in order to compare the standard deviation with the average or we compare the measure of dispersion with a measure of central tendency and find out the ratio which we call a sigma by X bar and in this case that ratio sigma by X bar is 45.7538 divided by 49.7 and then expressed as a that gives us a value of 0.92 and when multiplied by 100 we would get a 92 percent. So, sigma by X bar is a measure that we define now which compares the dispersion with the average.

We also observe that in this case the average is runs scored a standard deviation is also runs scored. We may recall that we would first calculate the variance and then take the positive square root of the variance and the variance would have a unit of runs square and standard deviation would also have the unit of runs and therefore, sigma by X bar will not have a unit and it will just be a number which compares the dispersion with average.

Now, this new measure that we have defined or we have computed now which is sigma by X bar is called coefficient of variation. So, coefficient of variation is a measure which compares or which tries to find out how much the dispersion is compared to the average in this case the coefficient of variation is 0.92. CV or coefficient of variation has no units because the standard deviation and the average have the same unit. It is appropriate when the mean is not close to 0. We should also understand that when X bar is close to 0, CV becomes very high because denominator becomes 0 and it becomes quite close to dividing by 0 which is infinity.

So, coefficient of variation is meaningful in situations where the average is not close to 0 and CV greater than 1 means there is considerable variation. In this example CV is close to 1, but is on the lower side. So, we now realize that the coefficient of variation for this particular cricket player is 0.92.

(Refer Slide Time: 06:09)



Scores of second cricketer in the last 10 test innings;

35, 141, 19, 1, 69, 54, 147, 46, 14, 103

Find the mean and standard deviation? How is the dispersion comparable to the average?

Total = 629; n = 10; average = **62.9** s = **49.1**

$Coefficient\ of\ variation\ C_v = \frac{\sigma}{\bar{X}} \times 100 = \frac{49.1}{62.9} \times 100 = 0.78$

Can you say who is better?

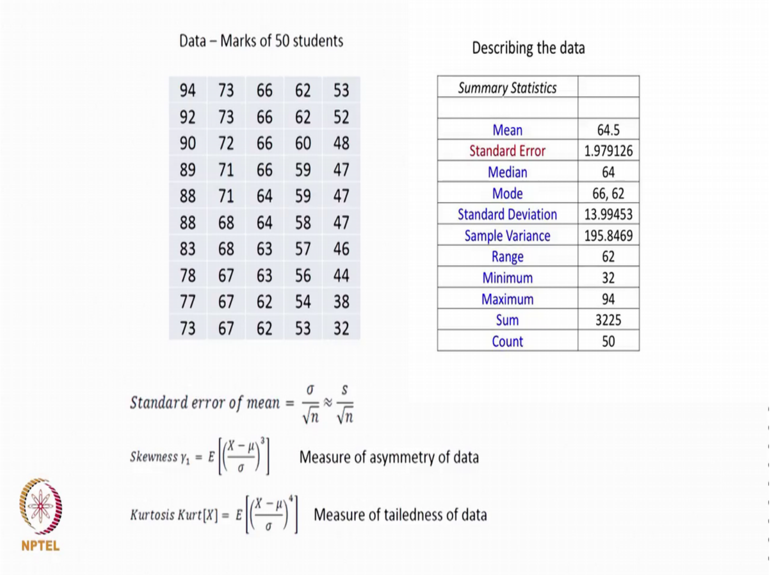
Now, let us look at another cricketer, a second cricketer who has played last 10 innings of score has been taken. So, these scores are 35, 141, 19, 1 etcetera. Now, we want to find the mean and standard deviation and we want to answer the question how is the dispersion comparable to the average.

Now, in a similar manner we calculate the sigma as well as the X bar and then we say that coefficient of variation is sigma by X bar into 100 which is 49.1 which is the average or that is a standard deviation divided by the average which is 62.9 and that figure gives us 0.78. So, now, if we compare these two cricketers, so, a total for the second cricketer is 629, n is equal to 10; so, average is 62.9, standard deviation is 49.1 and coefficient of variation is 0.78. Now, based on the coefficient of variation can we say between the two players who is a better player. So, the answer to that is that player who has a lower coefficient of variation can be taken as a better player.

So, coefficient of variation can also be used as a measure to compare when sigma and X bar are different for different players or different samples as the case may be as long as though they represent the same thing under consideration in which case in this case they are batsmen. So, the cricketer with the lower coefficient of variation of 0.78 can be taken as better compared to the other cricketer whose coefficient of variation was 0.92. Again we have to note that X bar in both the cases are not close to 0 and therefore, CV is a reasonable measure to compare the performance of both these cricketers.

Now, after defining variance and standard deviation we have now defined another method called coefficient of variation, which can also be used to compare or can be also used as a measure that uses a dispersion measure and a measure of central tendency.

(Refer Slide Time: 08:40)



Data – Marks of 50 students

94	73	66	62	53
92	73	66	62	52
90	72	66	60	48
89	71	66	59	47
88	71	64	59	47
88	68	64	58	47
83	68	63	57	46
78	67	63	56	44
77	67	62	54	38
73	67	62	53	32

Describing the data

Summary Statistics	
Mean	64.5
Standard Error	1.979126
Median	64
Mode	66, 62
Standard Deviation	13.99453
Sample Variance	195.8469
Range	62
Minimum	32
Maximum	94
Sum	3225
Count	50

$\textit{Standard error of mean} = \frac{\sigma}{\sqrt{n}} \approx \frac{s}{\sqrt{n}}$

$\textit{Skewness } \gamma_1 = E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right]$ Measure of asymmetry of data

$\textit{Kurtosis Kurt}[X] = E\left[\left(\frac{X-\mu}{\sigma}\right)^4\right]$ Measure of tailedness of data

NPTEL

Now, let us look at one more set of data which we have already seen this data. This data is the marks obtained by 50 students of a class in an examination and we have already calculated the mean median and so on. So, in this with this data now we can also calculate several statistics. Now, these summary statistics that are shown in the right are usually computed for a given set of data. So, we quickly go through this the mean is 64.5, standard error of the mean is sigma by root n and in this case approximated to s by root n.

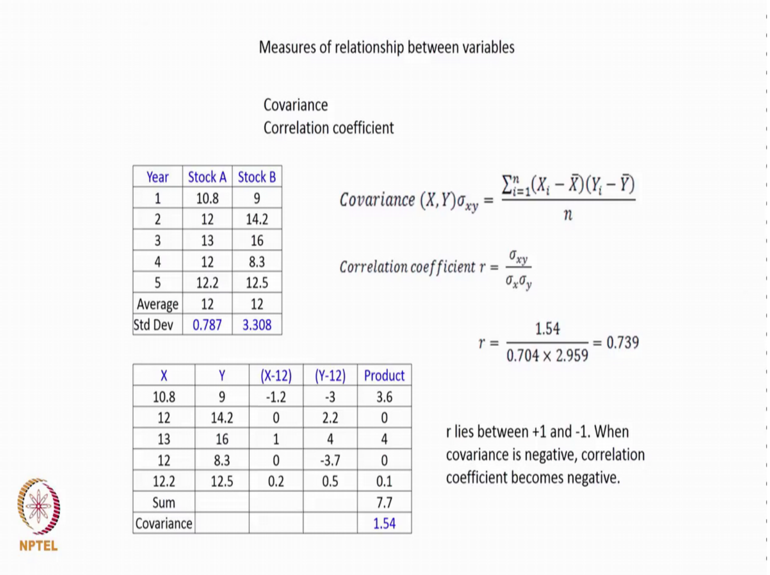
Now, we know that sigma represents standard deviation, s also represents standard deviation, n is the number of observations, the difference between sigma and s is sigma represents the standard deviation of the population while s represents the standard deviation of the sample we use smaller s or up lower case s and in this example these 50 students are seen as a sample and therefore, s by root n gives the standard error 1.979.

The median is 64, the mode is 66 and 62 standard deviation is 13.99, variance is 195.84 and then we define two more measures called kurtosis and skewness which we give the formula here. So, skewness and kurtosis are expected value of X minus mu by sigma the whole cubed and x minus mu by sigma to the power 4. So, normally what we do is the average, the variance, the skewness and the kurtosis are four measures where variance is X minus mu whole square and so on.

A skewness and kurtosis use a sigma as well and they are measures of symmetry or asymmetry of data and tailedness of data respectively. We look at a little bit about skewness particularly with respect to some distributions and data, but just for the sake of completion we are looking at these values of kurtosis and skewness. Range is the difference between the maximum and the minimum which is 62, minimum is 32, maximum is 92 sum is 3225 and count this 50.

So, given a set of numerical data we can calculate all these summary statistics which can also be taken through a Microsoft excel or any other software that can help you generate these summary statistics.

(Refer Slide Time: 11:31)



Measures of relationship between variables

Covariance
Correlation coefficient

Year	Stock A	Stock B
1	10.8	9
2	12	14.2
3	13	16
4	12	8.3
5	12.2	12.5
Average	12	12
Std Dev	0.787	3.308

$$Covariance\ (X,Y)\sigma_{xy} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n}$$

$$Correlation\ coefficient\ r = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

$$r = \frac{1.54}{0.704 \times 2.959} = 0.739$$

X	Y	(X-12)	(Y-12)	Product
10.8	9	-1.2	-3	3.6
12	14.2	0	2.2	0
13	16	1	4	4
12	8.3	0	-3.7	0
12.2	12.5	0.2	0.5	0.1
Sum				7.7
Covariance				1.54

r lies between +1 and -1. When covariance is negative, correlation coefficient becomes negative.

Now, let us also look at measure of relationship between variables now we go back to the same example of the stock. So, again two stocks A and B are given, the returns for the 5 years are given for both A and B. We have already seen that the average is 12 in both the cases the standard deviation is 0.787 and 3.308.

Now, we define another measure called covariance and covariance is defined as sigma X i minus X bar into Y i minus Y bar divided by n. So, we are showing the calculations here. Now, X bar which is the average is 12 in all the cases. So, Y bar is also 12 so, for the case when X is 10.8 X minus X bar is 10.8 minus 12 which is minus 1.2.

Similarly, Y minus Y bar is minus 3, but the product of X minus X bar and Y minus Y bar which is given in this formula as X i minus X bar and Y i minus Y bar is positive because it is a multiplication of two negative numbers. So, what we observe from this table is, there can be instances where the X value is lower than the mean in which case X minus X bar will be negative, there will be instances where it could be higher than the mean where X minus X bar could be positive and wherever it is equal to the mean it is 0.

Similarly, Y minus Y bar also behaves in a similar manner. So, we could have situations where one of them is negative and the other is positive which could give us a negative value of the product. In our example we do not come across a case where one of them is negative and the other is positive that can happen in which case the product will be negative. Now, in our example in all the five cases we observe that either both are negative or both are positive or one of them is 0.

Certainly there will be cases where for example, if the first one had been 14 and 9 then we realized that X minus X bar would be positive for stock A while Y minus Y bar will be negative and the product will be negative. So, when we take the sum of the products the negatives if there are any in this column will actually reduce the sum. So, covariance is sigma X i minus X bar into Y minus Y bar by n and in our case 7.7 is the sum of the products and 7.7 divided by 5 is 1.54 which is shown here as sigma X Y which is the covariance.

We already have the curl the standard deviations of X and Y calculated and therefore, we define another measure called correlation coefficient given by r which is sigma XY divided by sigma X sigma Y. Now, in our computation r is 1.54 divided by 0.704 into 2.959 which is 0.739. So, at this point you might wonder why the values have changed. Here I have shown 0.787 and 3.308 while I have used 0.704 and 2.959.

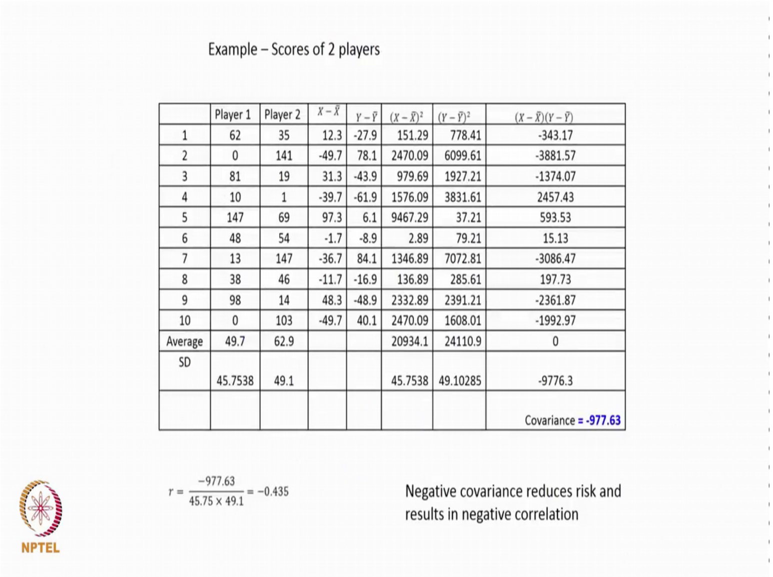
The reason that was done is here when I divided when I found the covariance I divided by n which is the number of observations which is 5; whereas, when I did the standard deviations here or the variance here I had used the n minus 1 sample formula. So, to be consistent I have divided it by n in both the cases. Therefore, instead of dividing by 4 which was done here I have divided by 5 and then you realize that this 787 becomes smaller because I have divided it by 5. So, to be consistent since I have divided it by 5 here I have to divide by 5 to get these values and the correlation coefficient is 0.739.

Now, r which is called the correlation coefficient lies between plus 1 and minus 1. Now, let us take a look at that now r is equal to covariance divided by sigma X, sigma Y. Sigma X and sigma Y are non negative quantities they cannot be negative because they represent the positive square root of variance which cannot be negative therefore, sigma X and sigma Y are either 0 or positive.

Now, sigma XY can become negative because that will depend on some of the products. In this case it is 1.54 we could have a situation where there are lot of negatives in this column and the sum has become negative. Therefore, correlation coefficient can become negative that is the first thing that we need to understand. That is because covariance can become negative and therefore, correlation coefficient can also become negative.

Now, it is also possible to show that the value of sigma X Y can only be within the range of sigma X sigma Y on the negative side and the positive side and therefore, correlation coefficient will be between plus 1 and minus 1. When covariance is negative correlation coefficient becomes negative, in this case correlation coefficient is 0.739.

(Refer Slide Time: 17:32)



Example – Scores of 2 players

	Player 1	Player 2	$X-\bar{X}$	$Y-\bar{Y}$	$(X-\bar{X})^2$	$(Y-\bar{Y})^2$	$(X-\bar{X})(Y-\bar{Y})$
1	62	35	12.3	-27.9	151.29	778.41	-343.17
2	0	141	-49.7	78.1	2470.09	6099.61	-3881.57
3	81	19	31.3	-43.9	979.69	1927.21	-1374.07
4	10	1	-39.7	-61.9	1576.09	3831.61	2457.43
5	147	69	97.3	6.1	9467.29	37.21	593.53
6	48	54	-1.7	-8.9	2.89	79.21	15.13
7	13	147	-36.7	84.1	1346.89	7072.81	-3086.47
8	38	46	-11.7	-16.9	136.89	285.61	197.73
9	98	14	48.3	-48.9	2332.89	2391.21	-2361.87
10	0	103	-49.7	40.1	2470.09	1608.01	-1992.97
Average	49.7	62.9			20934.1	24110.9	0
SD	45.7538	49.1			45.7538	49.10285	-9776.3
							Covariance = -977.63

$$r = \frac{-977.63}{45.75 \times 49.1} = -0.435$$

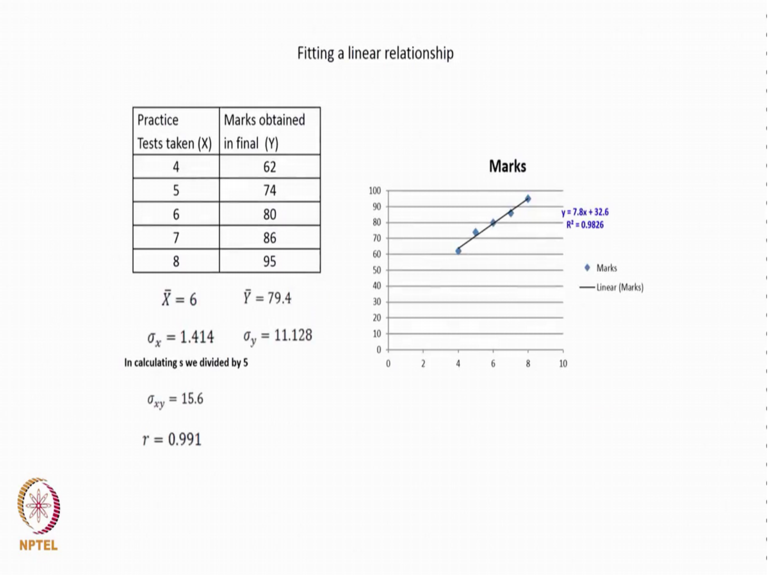
Negative covariance reduces risk and results in negative correlation

Now, we have also calculated this for the same two players we have calculated covariance becomes negative in this case. You can observe now that in this situation covariance has become negative because X minus X bar is positive Y minus Y bar is negative therefore, the product is negative and then we have found out the standard deviations and correlation coefficient is minus 0.435 in this.

So, one may make a general observer that a negative covariance reduces risk and results in negative correlation. So, one can make a general kind of a conclusion that since these two players have a negative covariance one can expect a lot of balance when both of them are playing. So, in situations where one is playing and getting a high score, the other actually does not seem to get a very high score, but they seem to balance out each other because more importantly the days when one of the players is getting a lower score you can observe that the other player has actually got a reasonable high score.

For example; you can see a 98 and 14 here and you also see a 0 and 103 here you also see a 0 and 141 here you see an 81 and a 19 here. So, you realize that together they balanced it and they seem to average about that the sum seems to average about 50 in or more in these cases. So, negative covariance reduces risk.

(Refer Slide Time: 19:09)



Fitting a linear relationship

Practice Tests taken (X)	Marks obtained in final (Y)
4	62
5	74
6	80
7	86
8	95

$\bar{X} = 6$ $\bar{Y} = 79.4$

$\sigma_x = 1.414$ $\sigma_y = 11.128$

In calculating s we divided by 5

$\sigma_{xy} = 15.6$

$r = 0.991$

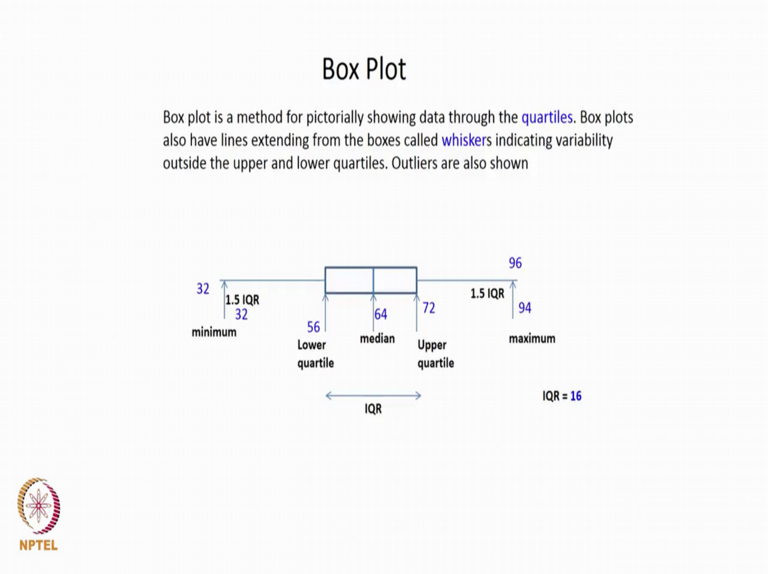
The next thing we can do is we can also try and fit a linear relationship if there is association between these quantitative variables. So, let us look at some data and try to do this. Let us assume that a set of students one particular student or 5 students we have collected data from them and let us assume that the data is the number of practice tests they have taken before a final exam and then the marks obtained in the final exam.

So, let us assume that these have taken these kind of number of practice tests and the marks that they have got. So, now, we can do many things we can first find out X bar is equal to 6, Y bar is 79.4, we can calculate sigma X and sigma Y and then we compute sigma X Y which is the covariance in this case the covariance is positive and the correlation coefficient is 0.991.

So, there is a good correlation and one can assume that if you take more practice tests it is possible to get good marks in the final examination. There is another interesting thing that we see which is the picture on the right which has been drawn using an excel software. So, we have just plotted a line there and along with the line we get some statistics. So, this statistics gives Y is equal to 7.8 X plus 32.6 and more importantly r square is equal to 0.9826.

Now, this r square represents the goodness of the fit and so on and it is possible to show that this r square which is 0.9826 is actually the square of the correlation coefficient 0.991. So, the goodness correlation coefficient also represents the goodness of the fit.

(Refer Slide Time: 21:04)



Box Plot

Box plot is a method for pictorially showing data through the quartiles. Box plots also have lines extending from the boxes called whiskers indicating variability outside the upper and lower quartiles. Outliers are also shown

We now go on to explain the box plot. The box plot is a method for pictorially showing the data using the quartiles. Box plots also have lines extending from the boxes which are called whiskers indicating the variability outside of the upper and lower quartiles, outliers are also shown. Because this also shows the whiskers this plot is sometimes called box and whisker plot. Now, if we go back to the data which we have seen earlier where we looked at marks obtained by 50 students and we calculated the interquartile range the lower quartile the upper quartile and so on, the box plot is drawn and that is shown here in this picture.

Now, you realize that the median which in this case was 64 the lower quartile is shown here as 56, the upper quartile is shown here through this arrow as 72, the IQR is shown here we can even write the value IQR is equal to 16 can be written here. So, IQR Interquartile Range is shown here. Then we draw these two lines it is customer again there are several ways of describing the box plot and we are going to use one of them. So, what we do is we draw a line that is equal to 1.5 times the interquartile range, now we can draw this to scale. So, we can actually do this to scale so, 64 would come here the differences will all be here and so on and in this case we are not drawn it to scale because 56 to 64 the differences it is the 64 is here 72 is here. So, this seems to be drawn to scale.

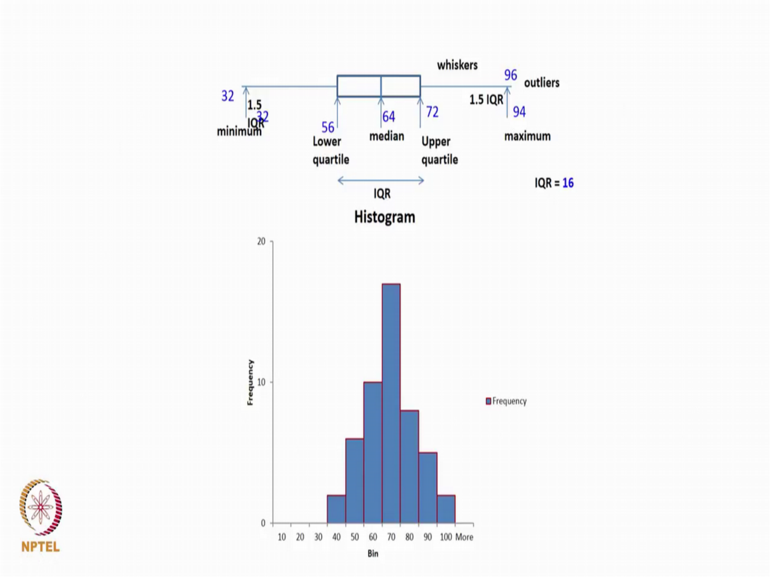
Now, 1.5 times IQR; IQR is 16 so, 1.5 times IQR is 24 and therefore, we draw this thing up to 96 and you can see carefully that is kind of slightly extending beyond the maximum which is 94. Now, on this side once again 24 is the interquartile range so, 56 minus 24 is 32. So, it kind of coincides with 32 which is shown here which is the minimum as well as the interquartile range.

Now, all the data points which are between 72 and 94, note that we do not have the maximum being 94, we do not have values more than 94. Therefore, in this box plot this can actually end with 94 itself because we do not have anything more than 94 though the 1.5 times IQR is 96 here we can end this with 94 because the maximum is 94. Here it coincides with 32 and therefore, it ends with 32.

Now, we can have situations where 1.5 times IQR is below the maximum or 1.5 times IQR is more than the minimum that can also happen. In the example of marks from the lower side it coincided with the minimum and on the upper side the maximum exceeded the 1.5 value, but we can have situations where the one point in this case the 1.5 IQR exceeded the maximum. So, we could have cases where the maximum is more than 1.5 IQR. So, when maximum is more than 1.5 IQR, some points can lie between the 1.5 inter quartile range point and the maximum and these are called outliers. Similarly, on this side we can have situations where the minimum is still lower than the point 1.5 IQR are on the other side and there could be points which lie between the minimum and the 1.5 IQR and they are all called outliers.

All the points that lie between this upper quartile and the end in this case it is the maximum or in some other case it would be 1.5 times IQR whichever is smaller and that points are called whiskers. Similarly, in this case the 1.5 IQR coincides with 32, but if we look at a situation where the minimum is still lower than 1.5 IQR, all these points between 1.5 IQR and the lower quartile are called whiskers and those to the left of the 1.5 IQR are called outliers.

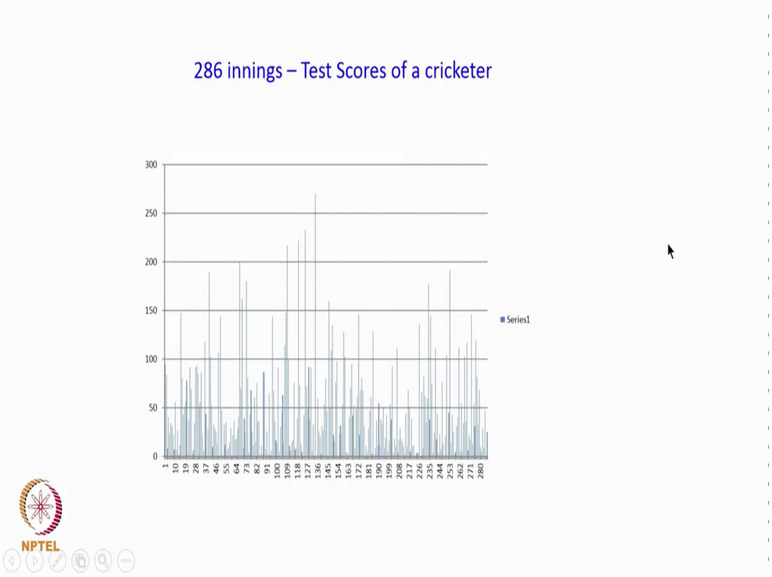
(Refer Slide Time: 26:18)



So, it is also customary the same picture is shown here it is also customary to show that show this above the histogram so, we can understand that and the only thing requirement is that this has to be drawn to scale. Right now they are not drawn exactly to scale, but we can try and appreciate a few things. The median is actually somewhere here which is 64, you can see the 56 is here in this picture. So, 56 is somewhere here in this picture 72 is here in this picture, 72 is here in this picture and so on. The maximum is somewhere here, but since it is not drawn to scale the maximum is outside.

So, also customary to show the box above the histogram, but the box plot by it itself tells us a good description of the 5 point summary of the data, because it contains a minimum, it contains the lower quartile, contains the median, contains the upper quartile and the maximum and shows the inter quartile range.

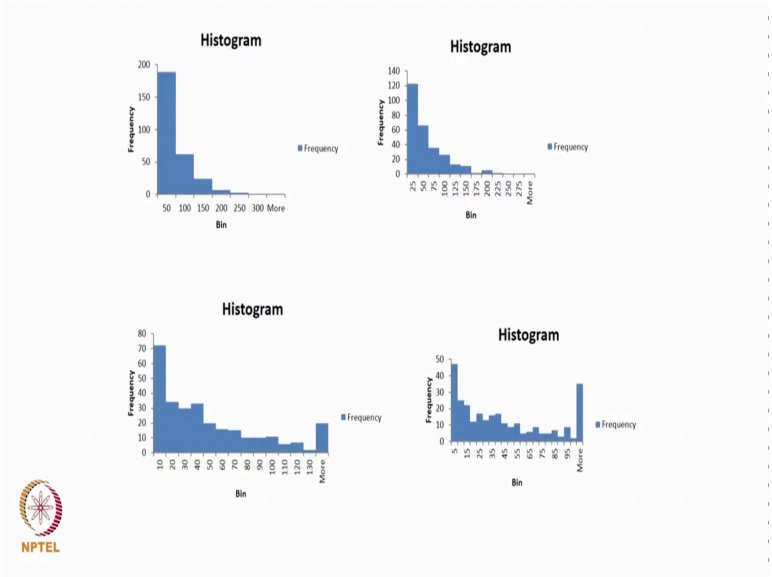
(Refer Slide Time: 27:20)



286 innings – Test Scores of a cricketer

We just try to show how the histogram looks for 286 test scores of a cricketer. So, this is plotted out to the 286 test innings scores.

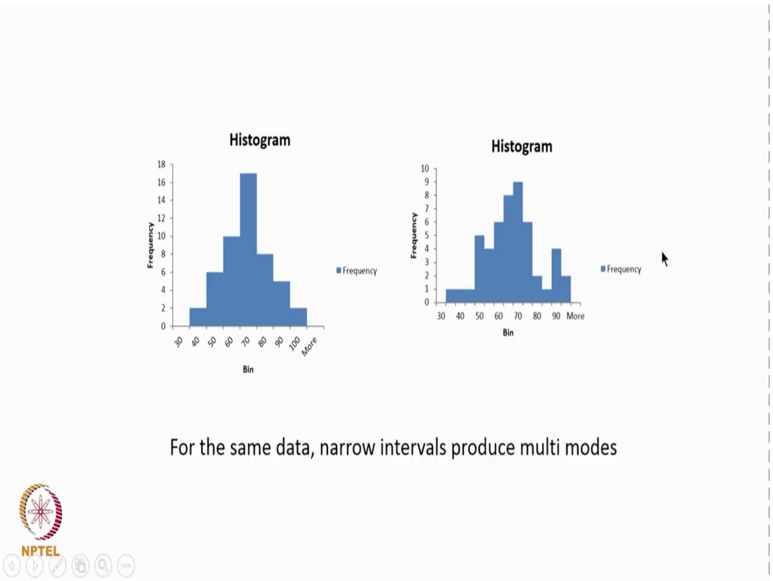
(Refer Slide Time: 27:34)



And, we could also we are going to show here how the histogram looks different depending on the way we draw the histogram. So, what we do is in this case we look at frequencies of 0 to 50, 50 to 100 like that 250 and 300 and more you realize this is how the histogram behaves. On the other hand if we say that it is 0 to 25, 25 to 50, 50 to 75 and this is how the histogram behaves you can see it is slightly different from the earlier one.

Now, here is a case where we do 0 to 10, 10 to 20 and we do this still about 130 and say greater than 130 you can see a small peak where data greater than 130 is aggregated and in this case we show 0 to 5, 5 to 10 and go on till 100 and then say greater than 100 you see a higher aggregation. So, all that we want to tell here is and we have a large amount of data depends on how we present the data and we are just looking at the histogram one has to also before making any decision look at it very carefully to understand the frequencies that are there and is there something like a more or an outlier and so on.

(Refer Slide Time: 28:49)

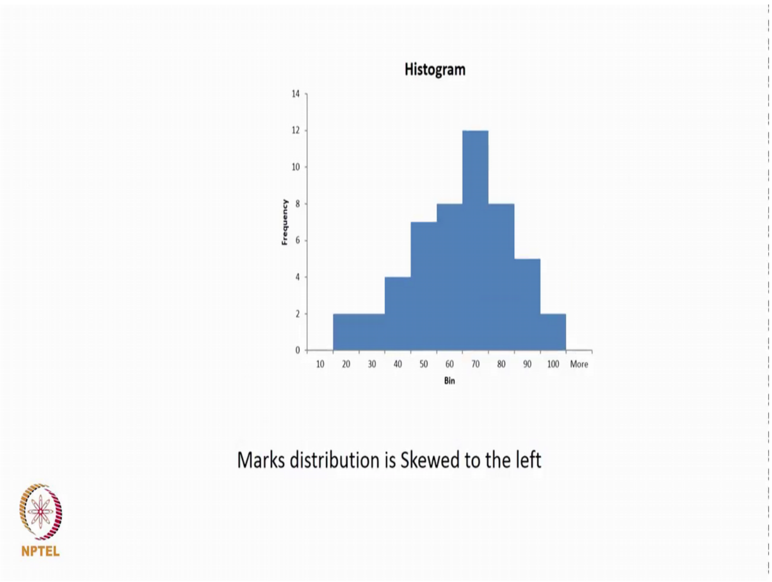


For the same data, narrow intervals produce multi modes

Now, we also want to show this for the same data when we do this 0 to from 30 to 40, 40 to 50 and so on this is how the histogram looks like, but then if we divide it from 30 to 40, 40 in a width of 5 now you realize it behaves slightly differently and it gets if you get a feeling that there are multi modes you know there is a mode here which is the largest.

But, there is also a small mode here and so on. So, as we try to reduce the interval on the x-axis we realize that it could show us more notes.

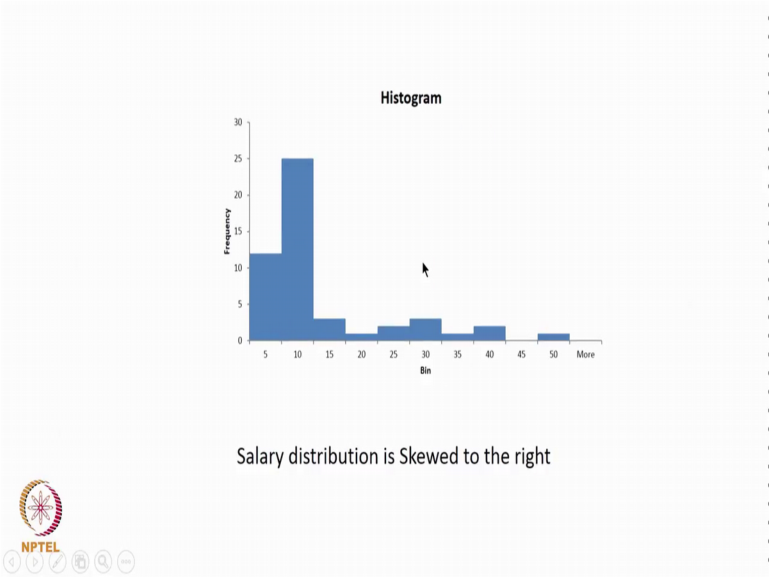
(Refer Slide Time: 29:27)



Marks distribution is Skewed to the left

And, in general observation is that Marks distribution is skewed to the left, you can see a small tail here which is skewed to the left.

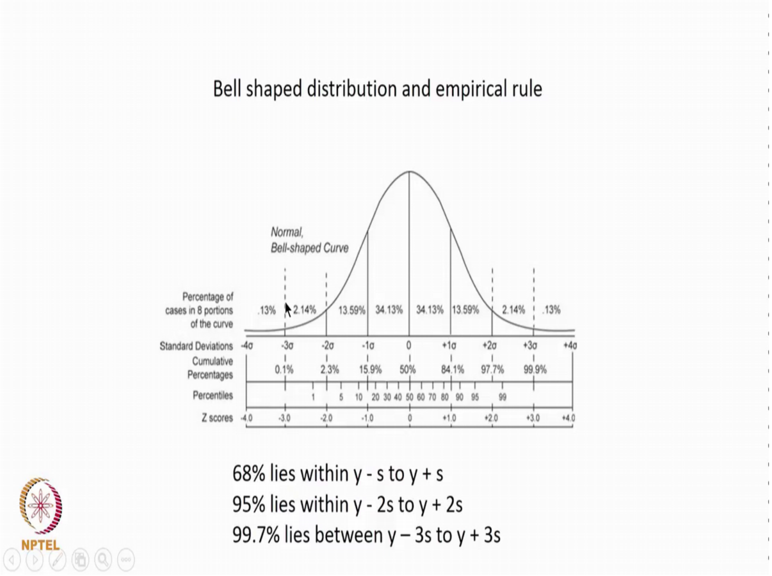
(Refer Slide Time: 29:36)



Salary distribution is Skewed to the right

And, salary distribution is skewed to the right and you can see a long tail here fewer and fewer people will get very high salaries and so on. Remember we are marking frequency on this smaller large number of people would be getting a smaller kind of a salary.

(Refer Slide Time: 29:52)



Bell shaped distribution and empirical rule

68% lies within y - s to y + s

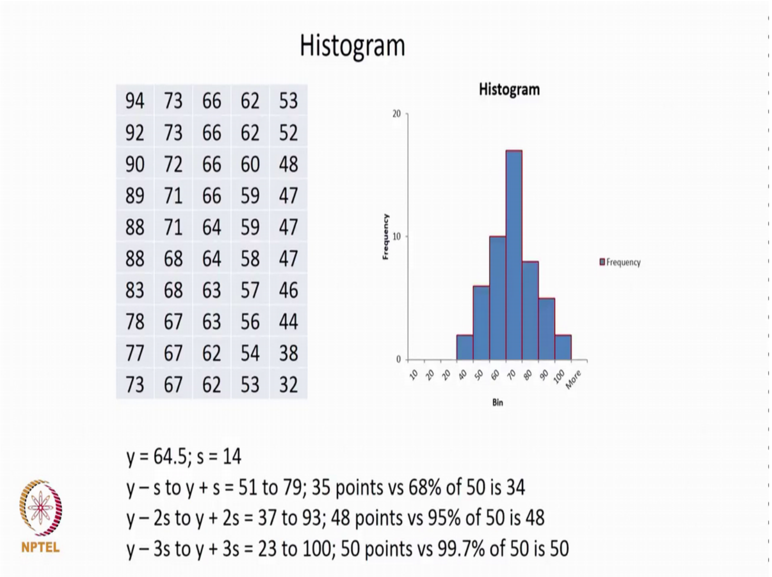
95% lies within y - 2s to y + 2s

99.7% lies between y – 3s to y + 3s

Now, this leads us to the bell shaped distribution called the normal distribution about which we will look at it in more detail towards the end of the this course. So, the normal distribution is a bell shaped curve also called a Gaussian distribution.

So, here this is the mean. So, 68 percent of the data will live within y just to y plus s, 95 percent will lie within y minus 2 s to y plus 2 s and so on; where s is the standard deviation and so, on.

(Refer Slide Time: 30:24)



Histogram

94	73	66	62	53
92	73	66	62	52
90	72	66	60	48
89	71	66	59	47
88	71	64	59	47
88	68	64	58	47
83	68	63	57	46
78	67	63	56	44
77	67	62	54	38
73	67	62	53	32

y = 64.5; s = 14

y – s to y + s = 51 to 79; 35 points vs 68% of 50 is 34

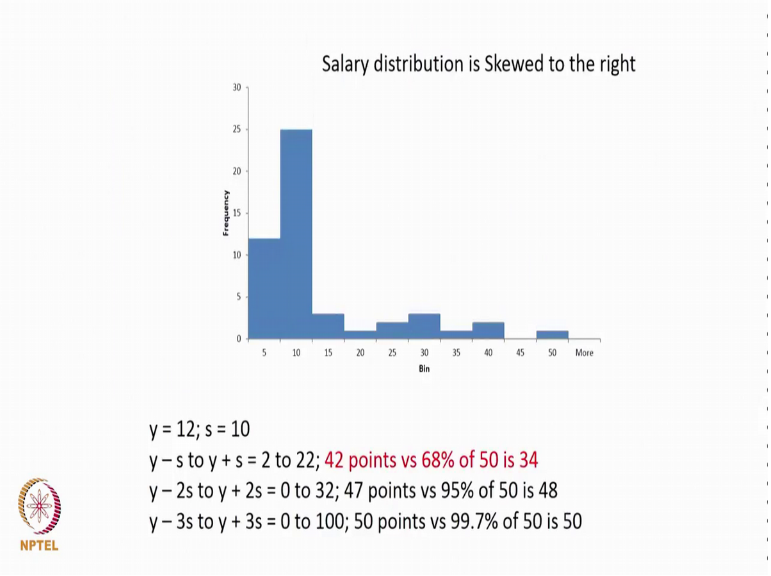
y – 2s to y + 2s = 37 to 93; 48 points vs 95% of 50 is 48

y – 3s to y + 3s = 23 to 100; 50 points vs 99.7% of 50 is 50

So, for this data we have plotted the histogram and we have tried to check based on our consideration whether it looks normal, it looks reasonably normal y is 64.5 the standard deviation is roughly taken as 14. So, y minus s to y plus y s is 51 to 79, 35 points are there and if we compare it with the previous slide which said 68 percent will be there. So, 34 should be there, 35 points are there. So, when we take y minus 2 s to y plus 2 s which is 37 to 93, we have 48 points versus an estimate of 48 points; y minus 3 s to y plus 3 s is between 23 and 100 and we have 99.97, 50 points we also have 50 points.

So, this can be taken as reasonably as a normal distribution.

(Refer Slide Time: 31:15)



Salary distribution is Skewed to the right

y = 12; s = 10

y – s to y + s = 2 to 22; 42 points vs 68% of 50 is 34

y – 2s to y + 2s = 0 to 32; 47 points vs 95% of 50 is 48

y – 3s to y + 3s = 0 to 100; 50 points vs 99.7% of 50 is 50

And, we also told that salary distribution is skewed to the right. So, we just show some example of this salary and if we take y is equal to 12 and s is equal to 10. So, in case y minus s to y plus s 42 points are there in that range versus a 34. Similarly, between y minus 2 s and y plus 2 s, 47 points are there against 48 and 50 points are in 50. Now, this is an indication that it is skewed and it is skewed to the right.

So, with this we come to the end of this lecture which talks about describing numerical variables. In the next lecture we will look at some revision problems and then move to the next topic which is association among categorical variables.