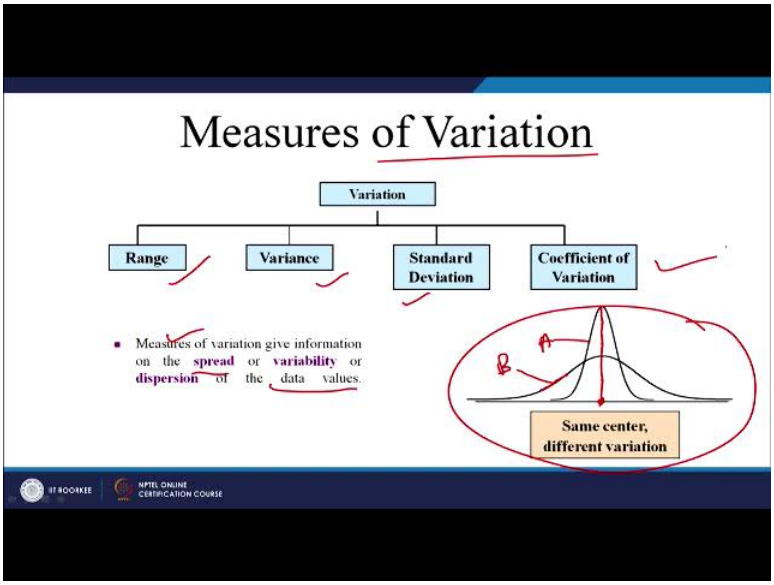
Business Statistics
Prof. M. K. Barua
Department of Management Studies
Indian Institute of Technology – Roorkee

Lecture – 08
Measures of Variation

Good morning, friends I welcome you all in this session. As you are aware in previous session we discussed about measures of Central tendency and the measures, which we discuss, were the Arithmetic mean, Geometric mean, weighted mean, mode and median. So in this session, we will talk about measures of variation.

(Refer Slide Time: 00:52)



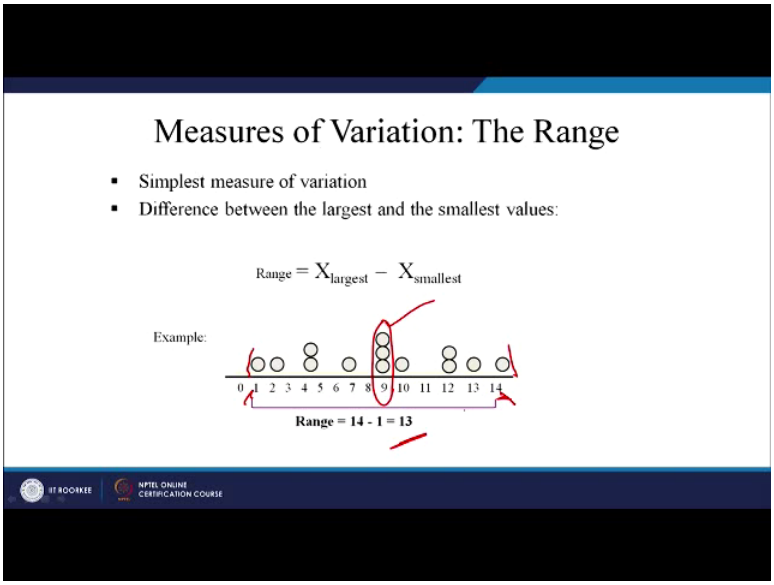
Now what is the need of measuring variation? Why you want to measure variation? Because in one of the classes, I have said that data generally speak. You need to hear them. So, if you hear them means what? You need to look at various properties of the data. So when you know more and more properties of the data, you can make a better decision. For that only you need to measure variation in the data.

So, it will help you or it the measures give you information on spread or variability or Dispersion of the data. Let us look at this distribution. So this is what? This is nothing but mean, mode, median are all these three are measures of Central tendency. What is central tendency? A point

where the distribution rests, this is the point where distribution is resting. Now if you look at these two. There are two distributions: Distribution A, Distribution B.

So which is which is got more dispersion where there is more variability in A or in B. So there is more variability in this distribution. Here data are concentrated. Here there more dispersed. So we will look at couple of methods of variation. First is very simple but the most inaccurate is the Range. Then, we will see variance, standard deviation and coefficient of variation. So there are four measures of variation.

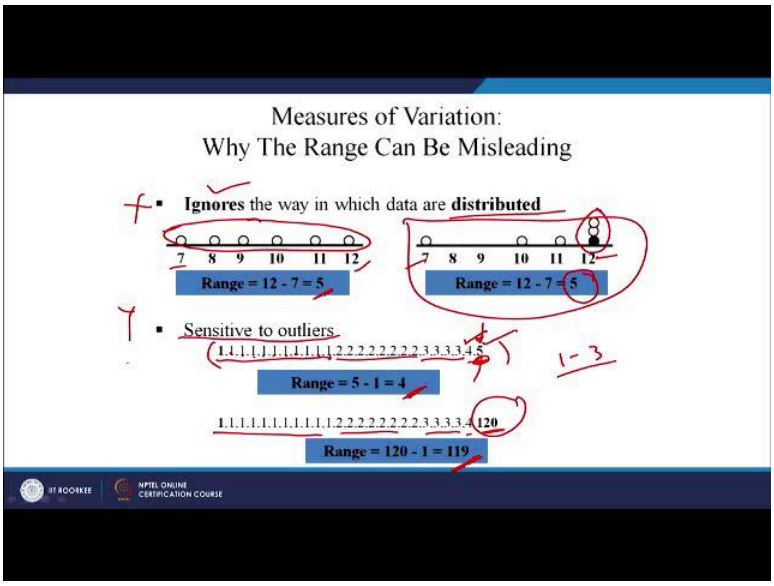
(Refer Slide Time: 02:50)



So, Range is as I said impressed the simplest. The difference between maximum value and minimum value of the data set, right. So this is your range. So let us look at this example. So you have got different data points. Let us say 1, 2, 3 and so on right. So, let us say the frequency of 9 number is 3. Suppose, if I ask you to calculate mode of this particular data set. How would you do it?

This is the mode, right because the number which is occurring most number of times. But we are not talking about mode. So we will calculate range to the minimum values from here to here, right. The minimum is 1, maximum is 14, 13 Range is 13.

(Refer Slide Time: 03:42)



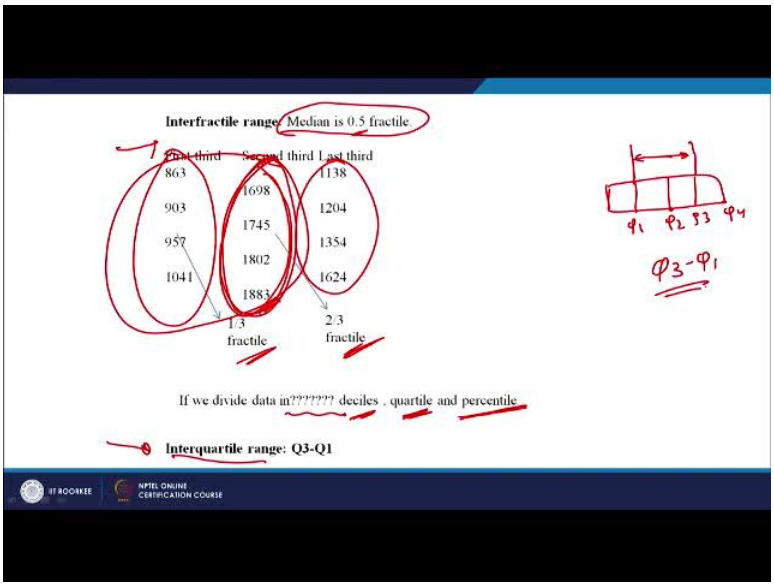
Let us look at the problems of range. Why range is not a good measure of variation? It ignores the way in which data's distribute. Let us look at this example. So there are 6 data points. 7, 8, 9, 10, 11, 12. Range is minimum to maximum, right. So, 12 – 7 = 5 is the range and if you look at this example. So, data points are 7, 10, 11 and 12. What is the range? 12 highest value, least value is 7. So what is the difference?

So, here the data are evenly distributed. Here there is a concentration of this particular data point. So but the answer in both these cases is the same. So what we will say? Range ignores the way in which data are distributed. Let us look at second problem. It is sensitive to outliers. So, let us look at this data set 1 1 1 2 2 2 2 3 3 4 and 5, right. If you look at the frequencies, just see. How many times 1 2 3 4 5 6 7 8 9 10 11 frequency is 9. Frequency is again approximately 8 to 9.

This is the frequency 4, frequency is 1 and 1. Then, Range is 5 - 1. So if you look at the 90% of data points in this data set are between 1 and 3. So there is only one point which is 5 and this has affected the range and their range is 4 here. So this is sensitive to outliers. Another example 1,2,3,4, and 120 right now, the range would be 119; so just one outlet has changed the range. So that is why the because of these two problems we do not use range as a measure of variation.

Now there are different types of range. We have seen in case of measures of Central tendency median. What is Median? It is nothing but a kind of range. It is it is 50% of the data.

(Refer Slide Time: 06:29)

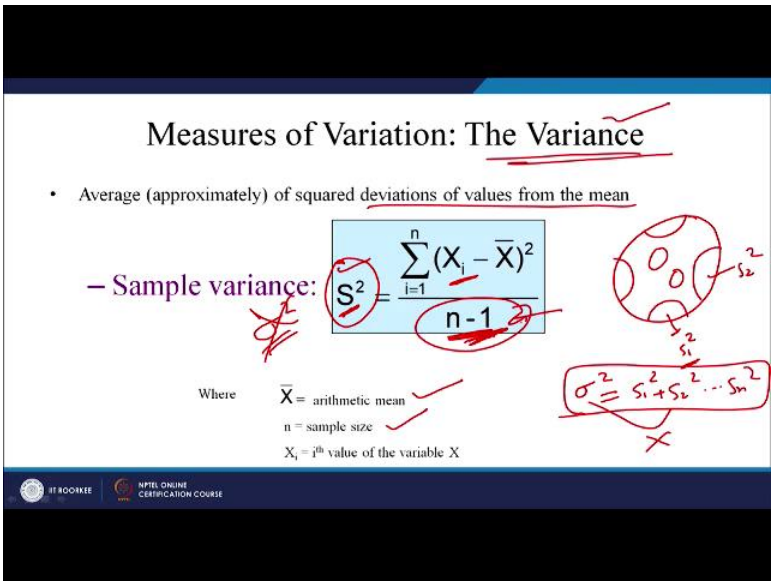


This 50% off the data we will call it median is in .5 fractile. Fractile is nothing but how you are dividing data, in how many parts you are dividing data. So each part is known as a fractile. So let us say you have got twelve data points. So these are 4 data points again, 4 data points, again 4 data points. So the first three data points we will call it one third of the fractile or one-third fractile. 2nd, 3rd are the four data points, right.

It is two third of the fractile. So one third when I say one third it means one third of the total this is two third upto this point so all these are two third. So this is two third fractile. If we divide data set into let us say, 10 equal parts, each one would be called a decile. If we divide data set into four equal parts it would be called quartile. If we divide dataset into 100 equal parts what it would be called? A percentage.

So you can have different types of ranges. So, one more important type of range is interquartile range is the difference between Q3. Let us say, in a data set there are first quarter Q1 and Q2, this also median, right. Q3 and Q4, right. This, this difference or this distance is known as IQR Interquartile range. It is Q 3 minus Q1.

(Refer Slide Time: 08:24)



Let us look at a better method of measuring variation is called variance so average of squared deviation of values from the mean. The good point about variance is that it takes into account each and every data points in calculation of variance. Why in case of range what we were doing? You are taking only minimum value and maximum value. The all other values we ignored. Why in case of variance we will take into each and every data point of the data set, ok.

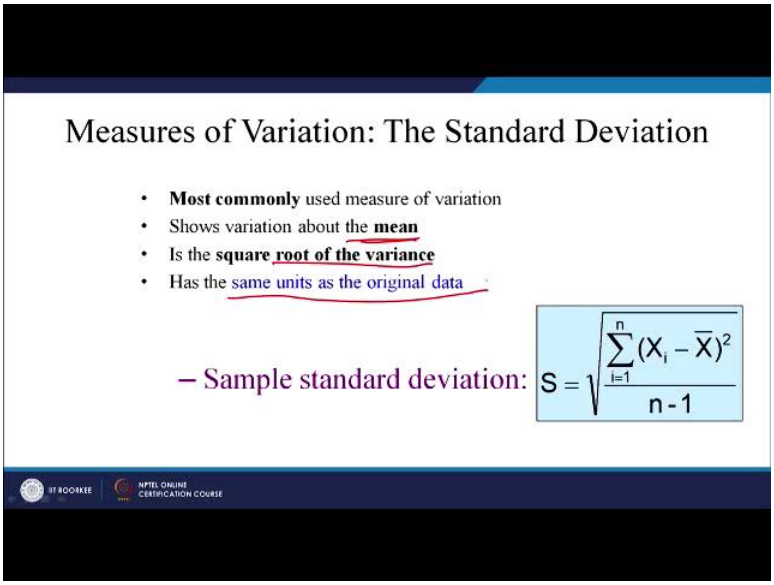
So sample variance is this is the symbol for sample variance is S square. So what would be the symbol for population variance? In one of the classes, I talked about these things. I said I talked about statistics and parameter right. So statistics are about samples and parameters are about population. So the sample variance is this population variance would be this sigma square. So let us look at first sample variance. Forget about this ok.

So, $S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$

So here X bar is arithmetic mean any sample size and X is the highest value of the data set. Now why this is n minus 1, why we have divided this n -1. This we have to be you know very firm. Why, why this is and n-1 why not 10. So the answer to this is what happens let us say there is a population and this is a sample, sample and if you calculate its variance, so it would be as S1 square right. So this is variance of 4 samples.

So take another sample this would be S2. So you can have different samples right. Differences samples, so at the end of the day there is something called Population variance. So this population variance should be equal to what? Sum of all these sample variances: S2 variance, S1 Variance so these two should be equal. But it is not this, it is not equal. Statisticians and researchers have proved that if you put n hour here instead of n -1 then these two will not be equal. If you put n - 1 over here in denomination then only this two would be equal so this, one of the reasons of having n -1 value here, sample variance right.

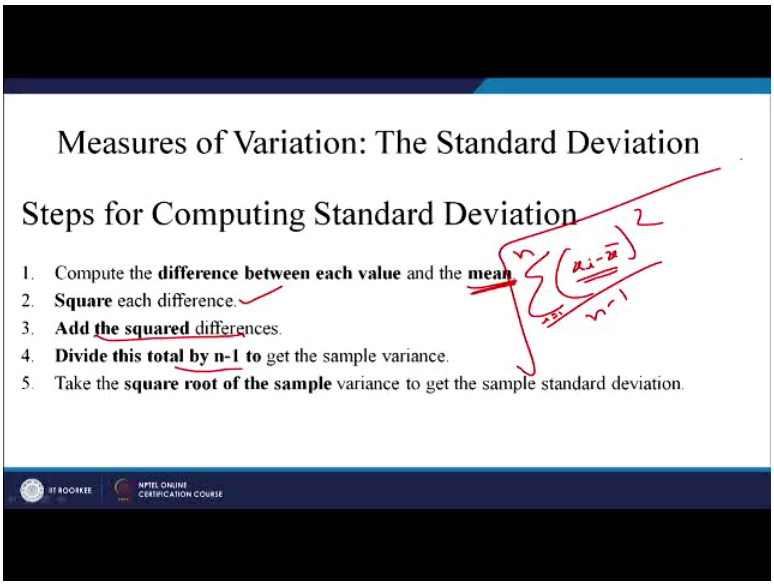
(Refer Slide Time: 11:45)



The next measure is standard deviation. So if you take square root of variance it becomes standard deviation. Just see this. Square root of variance is standard deviation, most commonly used measure, shows variation about the mean. It is the square root of the variance that is what we have already talked about. It is the same units as original data. So this is your sample standard deviation using three measures so far.

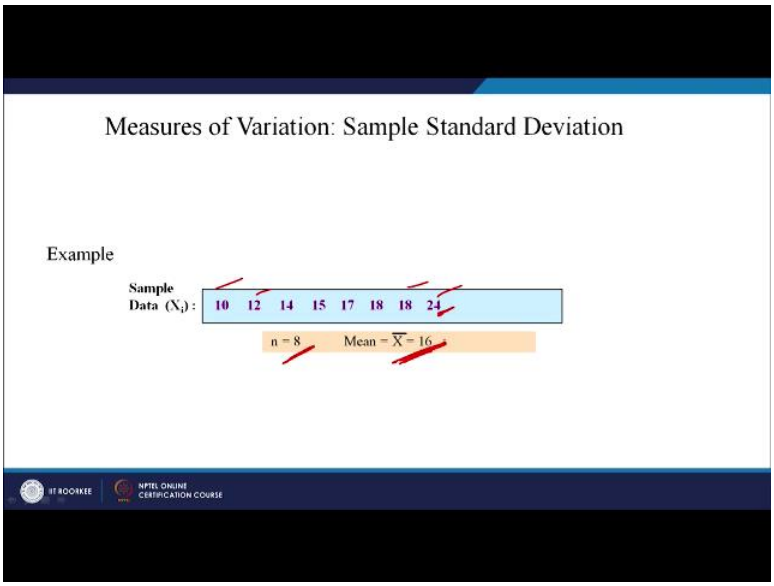
What are those measures range then variance then standard deviation right? So will work out couple of examples on standard deviation.

(Refer Slide Time: 12:35)



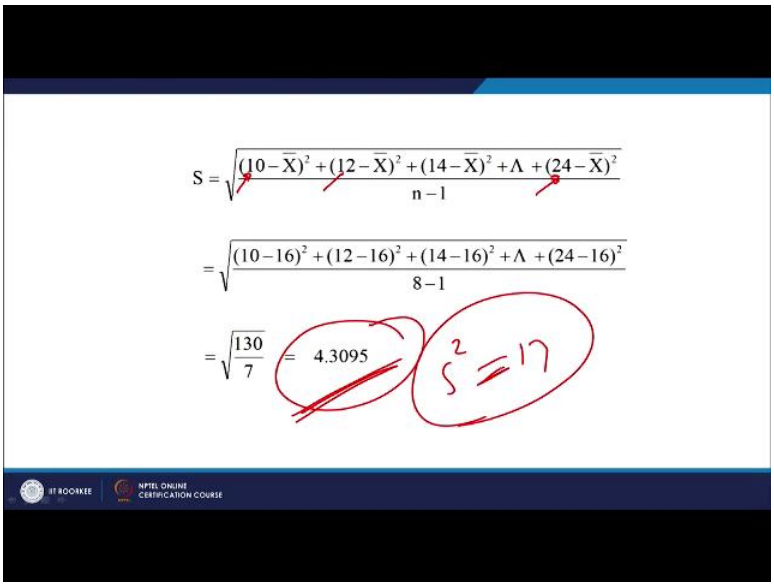
So first of all how to calculate standard deviation? It is very simple. So you will have mean first then difference of every observation with mean right. Then this difference is to be square right then you have to add all these differences. So this is the squared difference is divided by n -1. So here is ranging from 1 to 9, right, ok. Take the square root of sample variance to get the sample standard deviation. This is what you should do and this is the same formula right. So let us calculate standard deviation.

(Refer Slide Time: 13:27)



We have got eight data points 10, 12, 18 and final one is 24 right, So n is 8 mean is 16 so you take all these values and divided by 8 right. So you will get mean as 16. Now how would you proceed for calculating standard deviation, right?

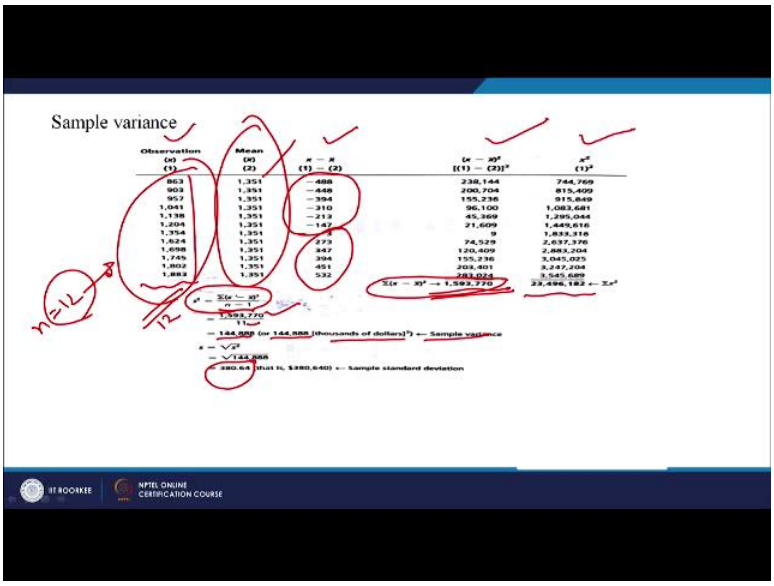
(Refer Slide Time: 13:54)



So this is your first data point, right this is your n first data point minus 16 whole square then next data point is 12 and so on then the last one is 24 divided by n -1 so you will get standard deviation is 4.30, right. Since you know standard deviation if you take square of this it would become variance. Or if you know variance take under root of it would become standard deviation. Let us take one more example on sample variance.

In fact in this case in this example we have we have calculated sample standard deviation first and then sample variance. So it would be around 17, right. If you take square of this, so sample variance approximately is 17.

(Refer Slide Time: 15:00)

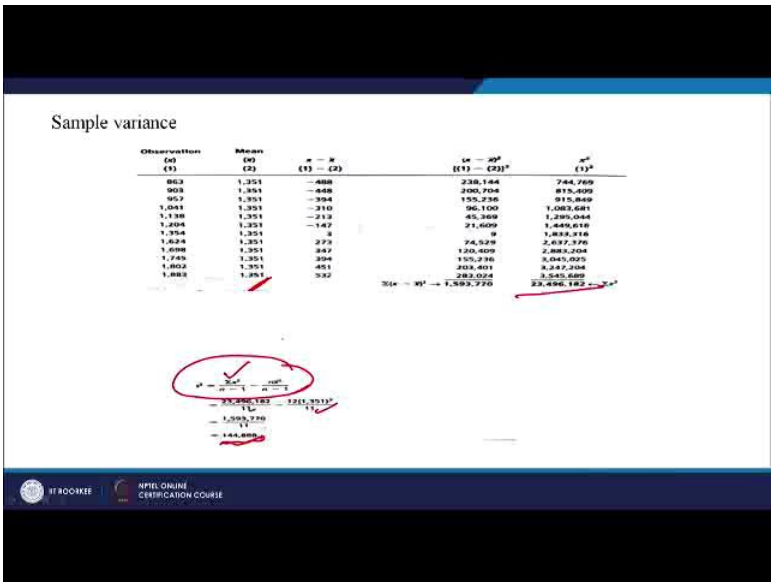


Let us look at another example. We will calculate sample variance. So these are different observations. Now first of all to calculate variance or standard deviation is what, you need to calculate mean. So this X bar is 1351. How did you get this? There are how many observations? n is what? n is equal to 12, right. These are 12 observations. Mean is this so how did we get this means we just summed up all these observations and divided that sum by 12, right.

So that sum if you want you just multiply 12 with this you will get the sum. Now the third one is X minus X bar. So X is this, X minus this, so these are first few negative values and then positive values. Take the square, summation of that, take square of X value, square of all these observations and the summation is this. Now put all these values in this formula X minus X bar whole square divided by n -1 so this is X minus X bar whole square summation is this divided by 12 - 1 is 11.

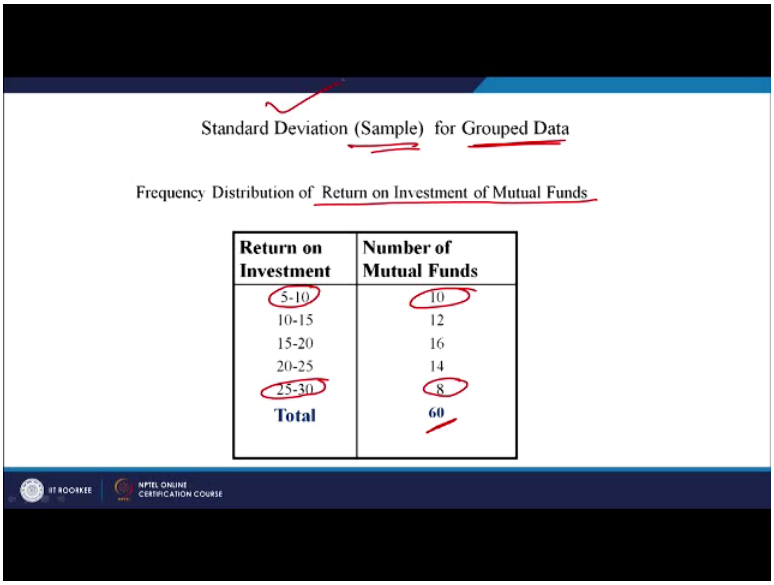
So we will get this value as this. So the sample variance is 144888 thousands of dollars because these are some observations in terms of dollars, since you know variance standard deviation is this. 38.6 is the standard deviation. So very simple, only the point you have to remember is that in case of sample variance in the formula this is in minus one. While in case of population, it is just simple n. So this is another method of calculating variance using this formula.

(Refer Slide Time: 17:29)



This is simple, just put summation of X square, which is this, n into X bar. X bar is what 1351 n is 12, right and take a square of this. So X bar square, 11 n-1, 11 n-1. This is the sample variance are you getting same answer 144888. We will check it 144888 is the answer.

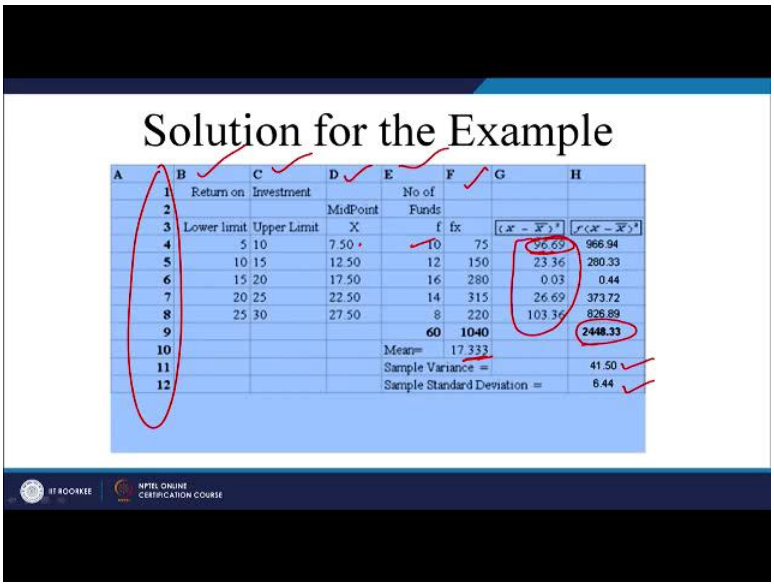
(Refer Slide Time: 18:01)



So this sample, you can have one more example on standard deviation of grouped data. So this very simple this is basically a grouped data right. So for examples we have seen for having ungrouped data. Now let us look at this question. It is it is a question on Frequency distribution of return on investment of mutual funds. Let us say the return on investment is between 5 to 10 rupees 5 to 10 dollars and there were such 20 mutual funds.

And the return was between 25 to 30 and there was at 8 funds. So total 60 Mutual Funds. You want to find out standard deviation. You can solve this question using Excel rather than solving it through a particular formula, right.

(Refer Slide Time: 19:15)

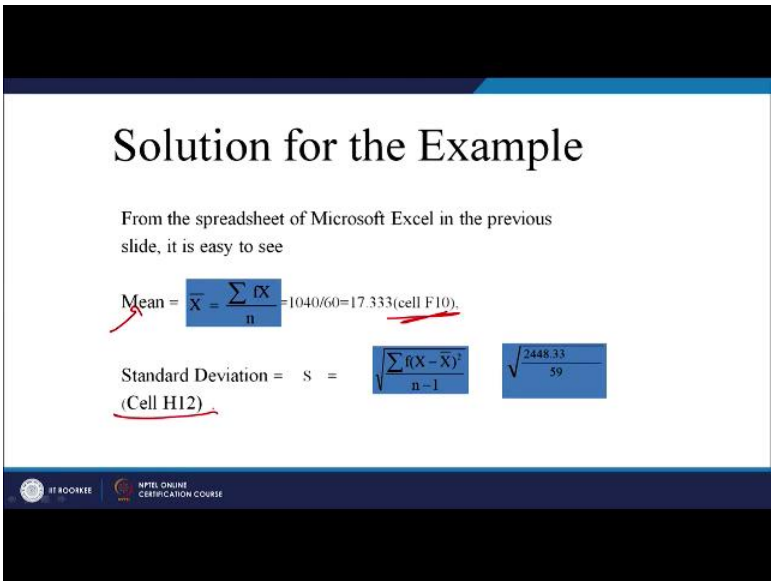


	A	B	C	D	E	F	G	H
1		Return on	Investment		No of			
2				MidPoint	Funds			
3		Lower limit	Upper Limit	X	f	fx	$(x - \bar{x})^2$	$f(x - \bar{x})^2$
4		5	10	7.50	10	75	96.69	966.94
5		10	15	12.50	12	150	23.36	280.33
6		15	20	17.50	16	280	0.03	0.44
7		20	25	22.50	14	315	26.69	373.72
8		25	30	27.50	8	220	103.36	826.89
9					60	1040		2448.33
10					Mean=	17.333		
11					Sample Variance =			41.50
12					Sample Standard Deviation =			6.44

So you just write down serial number here. So there are total 12, Rose first is this column lower limit, upper limit, midpoint, number of funds, is frequency is there, multiplied f identified into X. x minus X by whole square so this would be 96.69. So this X minus X bar. So 7.5 minus mean, which is 17.3 3 whole square it would be 96.69. It cannot be negative value all these values cannot be negative because these are squared. Similarly multiply this value by if f frequency.

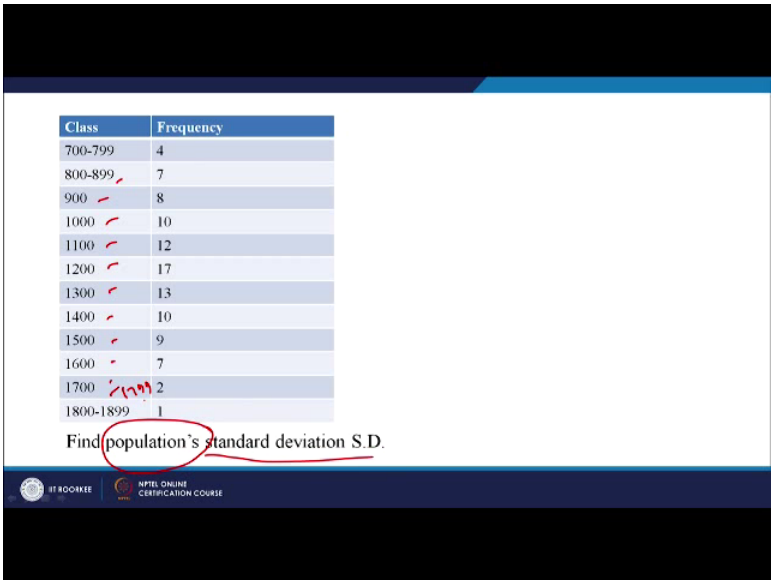
So you will get this one and this is the total. So mean is this, sample variance is this and standard deviation is this just take root of this.

(Refer Slide Time: 20:25)



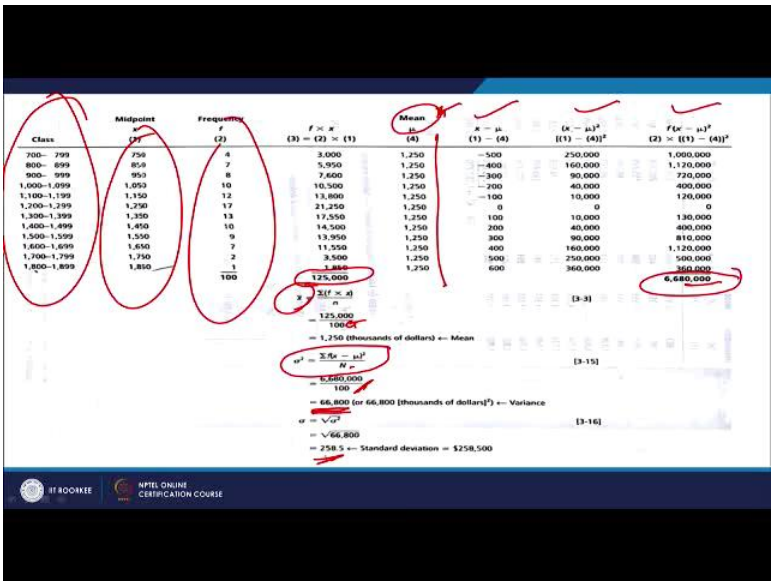
So if you look at if you look at this slide what we have done? We had mean in cell number F10 is this. This is f10 so f10 is mean. Similarly, you can have standard deviation in cell H12.

(Refer Slide Time: 20:49)



This is another example now here to calculate standard deviation of the population. So find out population standard deviation ok. So this again group data so class interval is 700 to 799, 899 remaining values you can write over, not a problem. So this would be 1799 right and so on Frequencies are given over here. We have to find out standard deviation of the population. So for standard deviation of the population you need to use this particular formula.

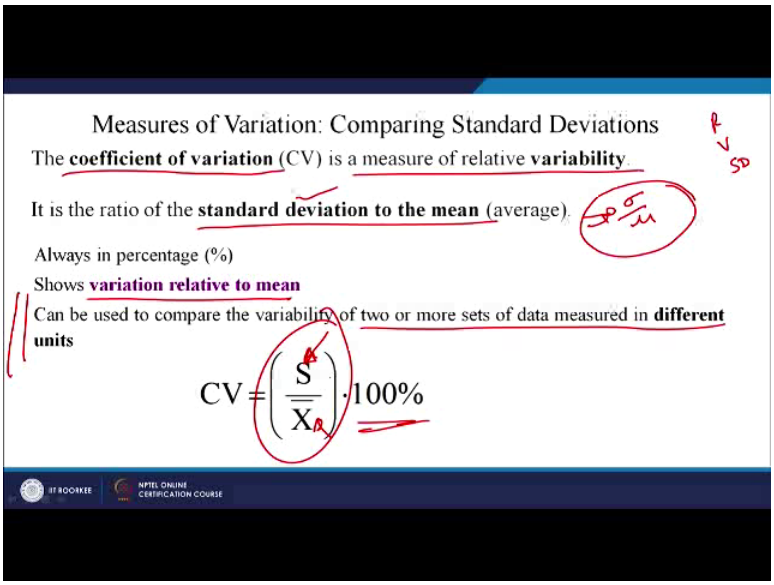
(Refer Slide Time: 21:35)



This is summation of f into X minus Mu whole square divided by n. These are different class intervals, these are midpoints, frequencies, f into X and summation. So this is a F into X which becomes 125000 divided by n. So this small n is its frequency right. So this is how you will get X bar 1250. This X bar is nothing but mean of the population that is why you get symbol as Mu. You do not use X bar. So this is mean of the population just write over here.

And then calculate X - Mu whole square multiplied this column by f. So this is f into X minus μ whole square this is 668 four times zero divided by 100. So this is your standard deviation, sorry variance of the population. Once you know the variance just take root of this it becomes standard deviation. So this is how you can solve equation for finding standard deviation of the population. We will just move on to next slide.

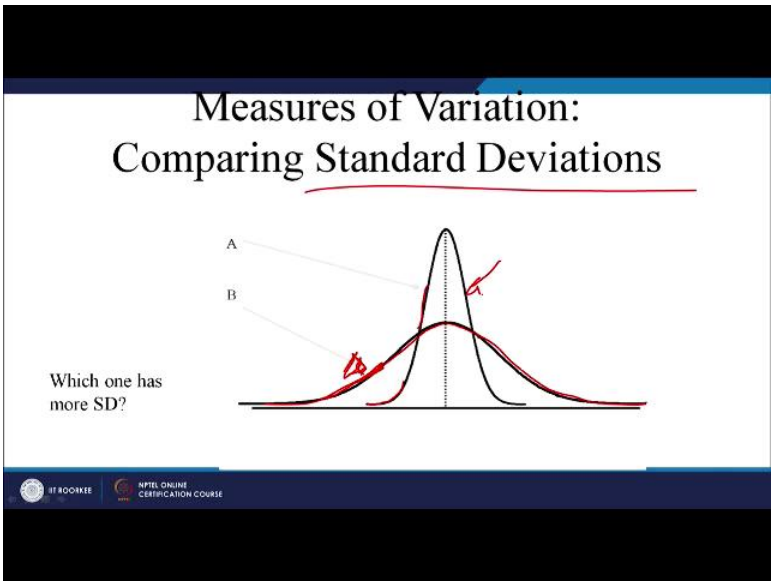
(Refer Slide Time: 23:45)



Now we will look at the fourth measure of variation, why fourth, because we have seen range, we have seen variance, we have seen standard deviation and the fourth one is the coefficient of variation, right. So if you got two distributions having different means and standard deviation you can compare them so coefficient of variance is a measure of relative variability. It is the ratio of standard deviation to the mean. This is the point you should remember. It is ratio of standard deviation to the mean.

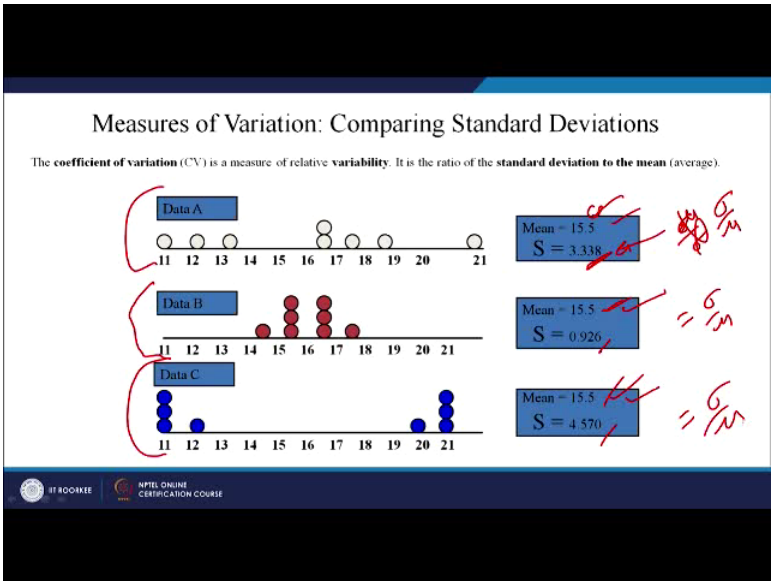
It shows variation relative to mean. How a data is away from mean; can be used to compare relative variability of two or more sets of data measuring different units. This is the most important application of coefficient of variation. As I said this is population coefficient of variation and this is sample. S and X bar of sample. S is what Standard deviation of the sample and this is mean of the sample. We just take percentage of this ratio ok. Now again I will ask one more question.

(Refer Slide Time: 25:14)



If you want to compare standard deviations and there are two standard deviations, so let us say there are two distributions. And this is the distribution A and distribution B. Now in which distribution do you think you will have more standard deviation, more standard deviation or more variability or more dispersion? You will have more standard deviation in case of B because it is dispersed more compared to A. Now we will work out couple of examples.

(Refer Slide Time: 26:08)



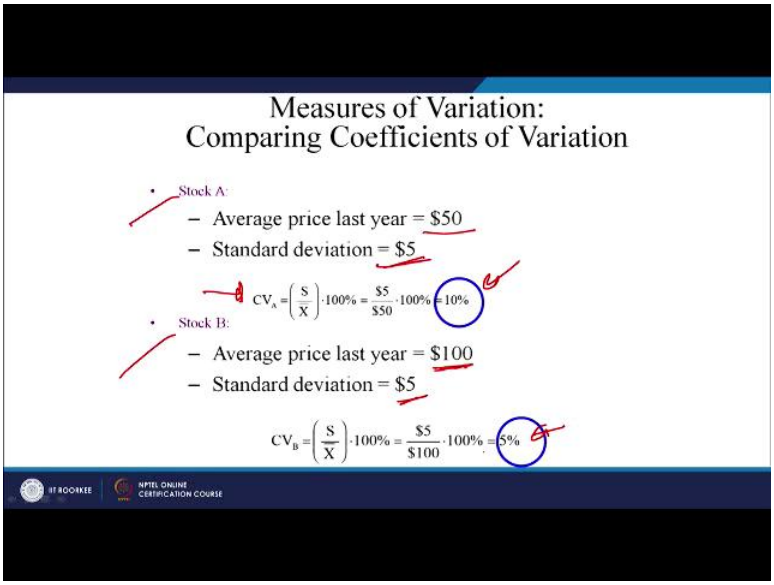
So have you got three distributions. You have data set A. So the mean of this particular data set is 15.5 ok and standard deviation is 3.38. I have already taught you how to calculate standard deviation. Data Set B and data set C. So mean 15.5, standard deviation this, mean this, if you

look at all these three data sets mean is same. 15.5, 15.5, and 15.5. Now you want to find out coefficient of variation.

Let me put it in different way. Let us say there are three Cricket players, mean of all the cricket player is 15.5 and their standard deviation is this, this is second and third. So which player will you choose which would be the best player out of 3. While selecting players, you should have you should see a player just got the highest mean and least standard deviation. So you just divide this is, if Mu by sigma right what we have seen.

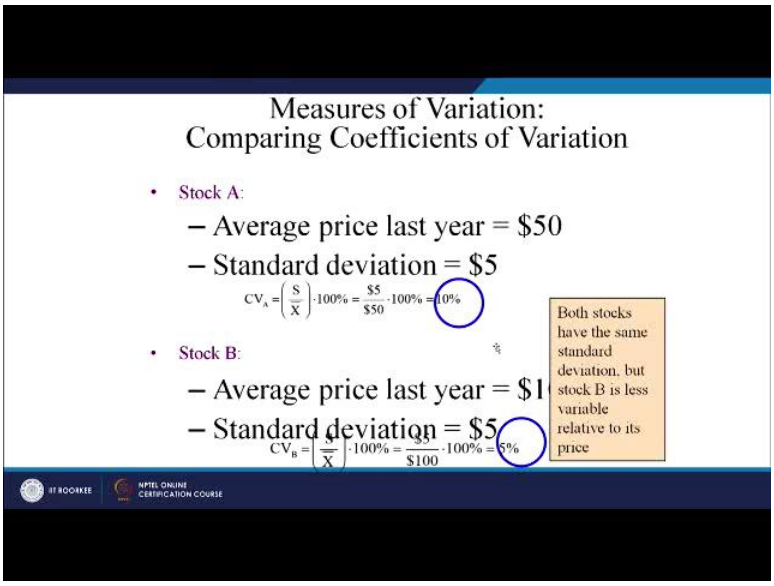
This is S by X bar. Let us call it Standard deviation by mean, standard deviation by mean. Just calculate all these three values. So select the players wherein you get me the least value so that would be the best player.

(Refer Slide Time: 28:26)



This another example let us say there are two stocks. Stock A and stock B. So average price last year was Dollar 50 standard deviation Dollar 5 coefficient of variation of is 10%. Average price is last year was this much, standard deviation 5 coefficient of variation 5%. This is how you can compare them. Which is the better one, which is the better stock? This one; the least out of these two right.

(Refer Slide Time: 29:12)



This is how you can compare different distributions and select the best one. So let me summarise what we did in today's session. We have seen different methods of variation namely Range, which is not a good measure, you should outrightly reject it. You should use standard deviation right because it takes into account each and every data points. We use standard deviation or variance and the fourth one was coefficient of variation, which we generally used for comparing different distributions.

So with this let me finish today's session. We will have next section and in which will talk about standardized normal distribution. Thank you very much.