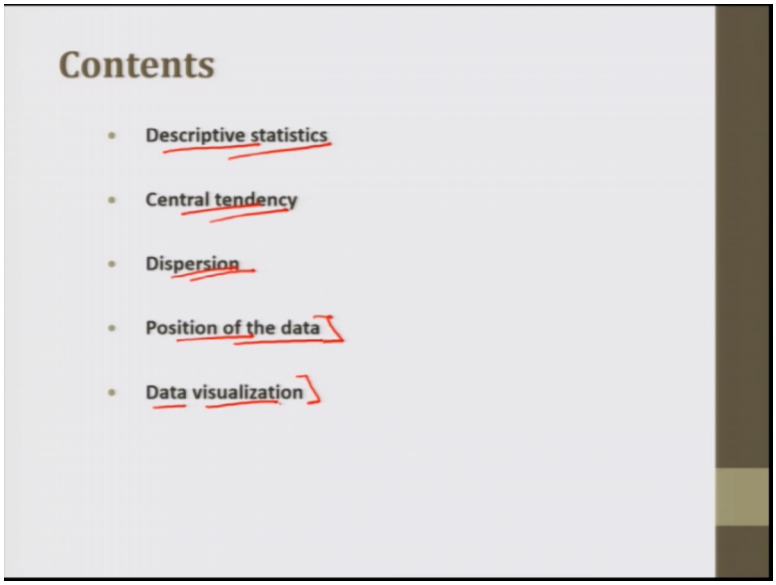
Engineering Metrology
Prof. J. Ramkumar.
Dr. Amandeep Singh Oberoi
Department of Mechanical Engineering & Design Programme
Department of Industrial & Production Engineering
Indian Institute of Technology, Kanpur
National institute of technology, Jalandhar

Lecture – 39
Statistics for Metrology, Fundamental Concepts (Part 1 of 3)

Good morning, welcome back to the course on engineering metrology and I am taking statistics in metrology. In this course, in this lecture I am going to discuss some fundamental concepts of statistics, going to the feedback that we are getting on this course, I am I have designed this structure to shed some light on the fundamental concepts of statistics, specifically in metrology. So, I will talk about descriptive analytics here.

(Refer Slide Time: 00:41)

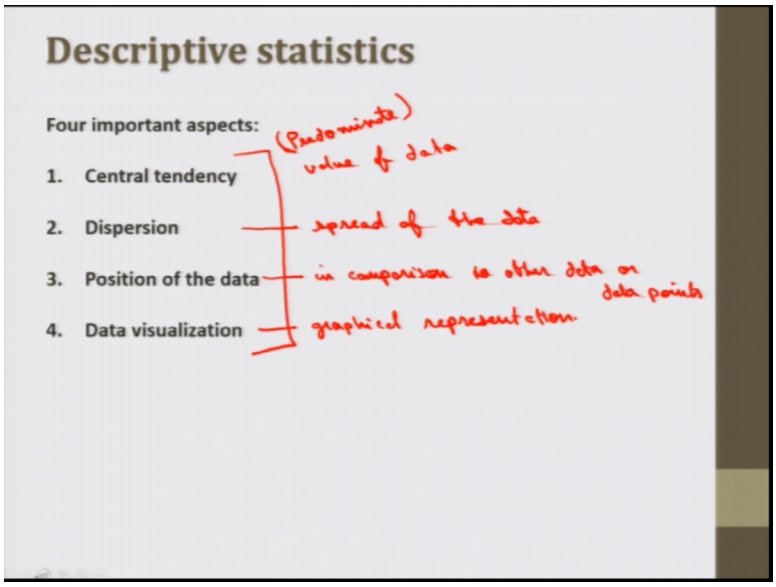


As I said before descriptive analytics tells or it directs towards defining or explaining the data or presenting the data in a proper way.

The data is present the data that, we call the observations that we have that we have done through instruments those are there, we have to describe that in some way, we have to find the average the mean, this mean fall in which class? It is the central tendency. Central tendency then we have, standard deviation that is dispersion how? What is the

spread of the data? That is dispersion then, we have position of data I also discuss about the median that, the median position of data, where does our data lie? The observation that, we are having where does it lie in the complete data set?

(Refer Slide Time: 01:44)

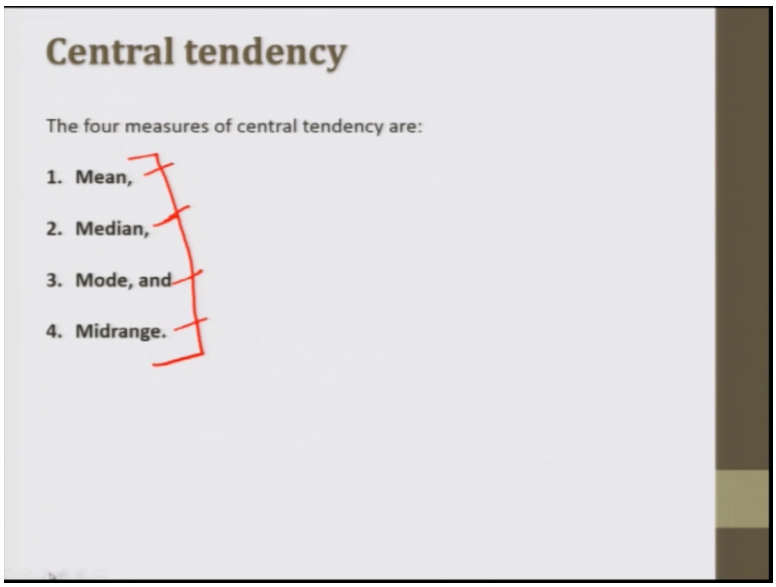


Then we will talk something about the data visualisation, how to graphically represent the data? I start from descriptive analytics or to be more focussed descriptive statistics here only, these are the 4 components that are just discussed in the contents of descriptive statistics. Number 1 is central tendency; central tendency or what we call the value of data (Refer Time: 02:01) value data, I will call the predominate value of the data, predominate value of the data that the data the complete data that values, we are trying to that is that is, will be close to the true value.

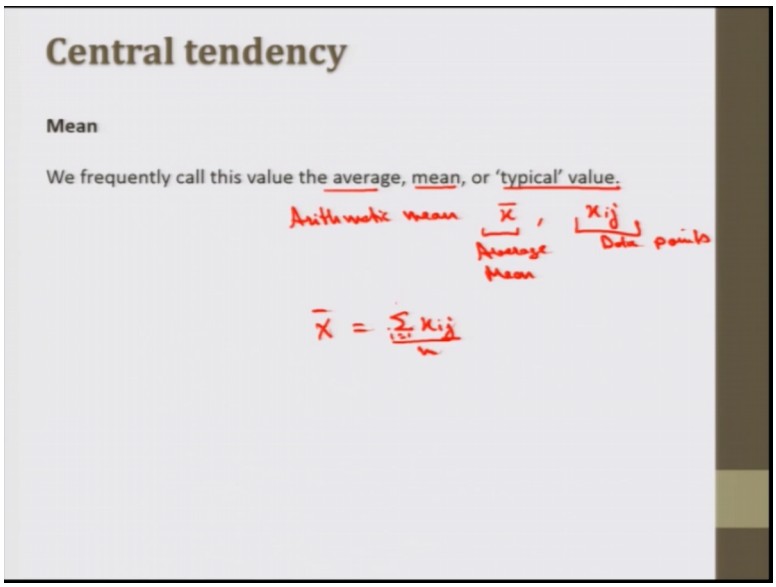
But I am telling should be whether it is or it is not. That is the different question, that we have to deal while we deal with the errors, but it has to be it should be actually close to the true value the measure of central tendency is a single value that describes the way in which a group of data cluster around a central value. To put in other words, it is a way to describe the centre of the data set. Next is the dispersion is nothing, but spread of the data the spread of the data the position of data position of data, talking about the position, position is always with respect to some other points, we have the whole data set where does our data lie? In that with respect to the available or the observed values, where does it lie?

So, it is in relation to I would say in comparison to other data or I call it data points. Next is data visualisation, data visualisation is nothing but, the graphical representation of the data, whether it is graphical or whether it is in tabular form. When we present the data, when we write the report, the conclusions or the results that we write is a different thing that is prescriptive statistics. In descriptive statistics with just need to present the data in the proper way for that, what kind of plot shall we use? So, it is important that is graphical representation of the data.

(Refer Slide Time: 04:30)



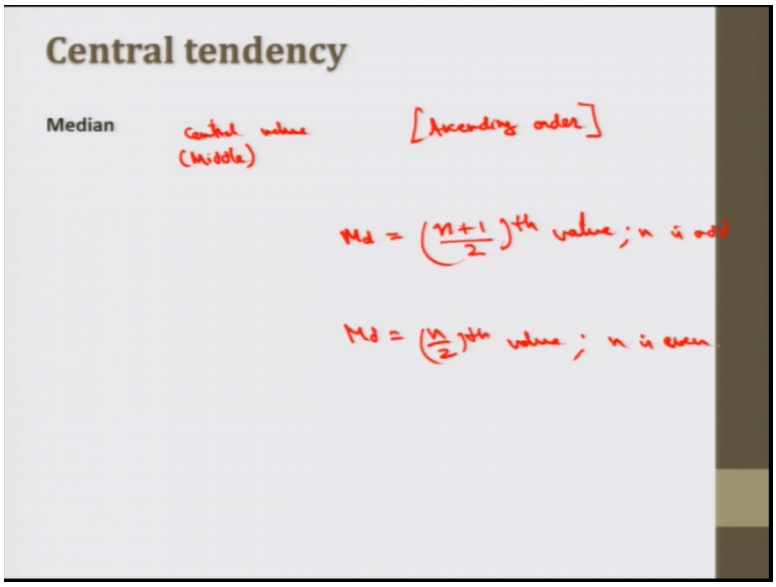
(Refer Slide Time: 04:44)



So, let me start from the central tendency, there are 4 measures of central tendency, which are mean, median, mode and mid range. So, most of you must be knowing, what is mean? We frequently call this value as the average mean or the typical value that the data should attain.

So, this value is nothing, but it is the arithmetic mean, I would better put arithmetic mean which is generally denoted by x bar, if the data values are x i j, these are my data points. So, this is my average or better we call it mean. So, this value this is actually, the value that we receive, when we ask for the average grade or the average high or low temperature for a specific environment for say, average temperature in a day or average temperature of the city in a month, something like that. So, it can be denoted x bar is equal to summation x i j by n.

(Refer Slide Time: 06:06)



Ok similar to mean, we have median; median is the central value. Mean is in continuous data, we have a set of values, we just take the average, we sum up of all those and we divided by the total number of values that, we have that is mean. Median is the central value, it can lie in the left or right of the mean, that will discuss, we will discuss the normal distribution and we will focus on the skewness in normal distribution, we can see how mean, median, mode are differently located

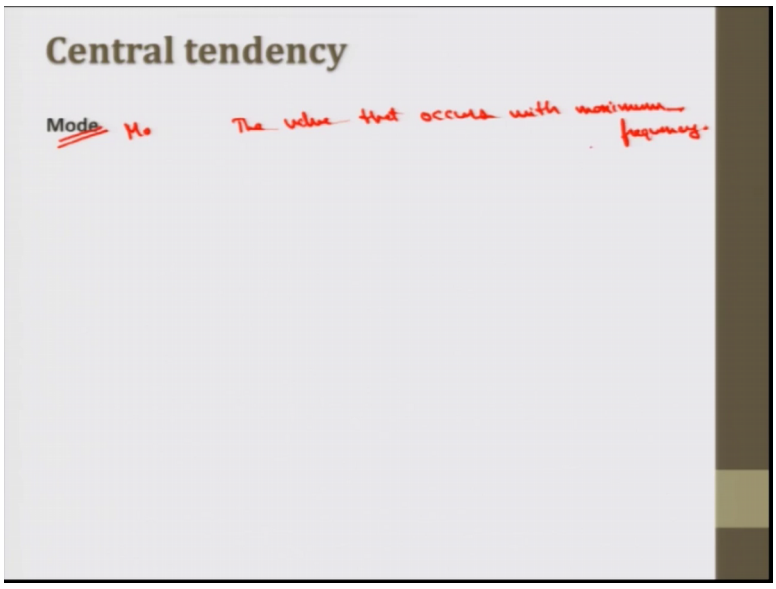
So, median is the central value, it is the central value to write it more clearly, I would put it is the middle value, middle value of the data when the data is arranged in ascending

order or in some order. It is arranged in ascending order, it may be arranged in descending order as well because, the median got the middle value should not change, but ascending order is the first criteria that we follow.

When the data set or the number of values are odd then, the middle value is the median, but when the data set contains even number of data points the 2 numbers is actually the median value. So, median value in general is denoted by subscript d is equal to this is equal to n plus 1 by 2th value. Please note, that this is not the end plus 1 by 2 value; it is the end plus 1 by 2th value. If I have 9 data points, it is 9 plus 1 10 by 2, that is the 5th value; 5thvalue can be anything in case of if you say, temperature it can be some temperature value ok. So, it is n plus, when n plus 1 by 2th position or value. So, where 1 is the position of the smallest value and n is the number of the pieces of data.

We can say it in this way as well, I will just do the calculations this is actually just introduction and the purpose of this lecture is not to just tell about the mean, median because you people might be knowing this thing, I will practice this, I will make you to learn how to calculate this using excel sheets, how to actually calculate, how to actually make the use of computers to calculate these things. And then, we will also take it forward to the quality control, where we will plot x bar chart and R chart and another charts like this.

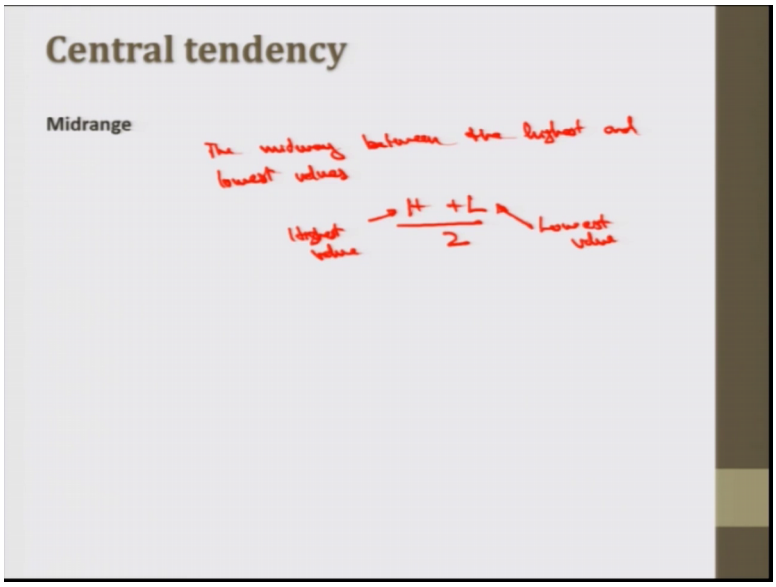
(Refer Slide Time: 08:58)



So, next value is now next central tendency is mode here, mode is denoted by M subscript o, it is the value that occurs with maximum frequency or the greatest frequency.

The value that occurs with maximum frequency, it may exist as a unique value or it may exist as a number of values. For instance, a 10 theta points and the value range varies from 5 to 8 and it is the number 6, that is repeating 3 of 4, 3 times and none of these other number is repeating, this many number of times 6 can be the mode, if 6 also repeating 3 times and 7 is also repeating 3 times, we can have 2 modes, but we cannot say, 6 plus 7 by 2, 6, 1, 5 is mode ok. So, mode is the value that occurs with maximum frequency.

(Refer Slide Time: 10:13)

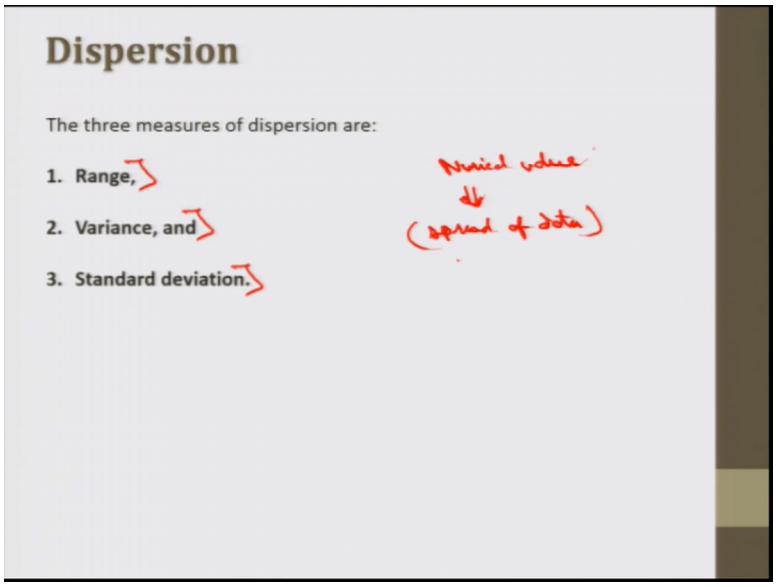


Next is mid range, mid range is the midway between the highest and lowest value, mid way between the highest and lowest values. So, if let me say, if I like I am talking about 10 values is from 7 to 8, the highest value is 8, the lowest value is 5, the midrange would be 8 plus 5 by 2, it is actually H plus L by 2, where H is the highest value and L is the lowest value ok.

Now, the 4 measures of central tendency represent, the 4 different methods of describing the middle value, these might be same in some case of data, data is behaving very symmetrically, these 4 values maybe same, but in general the values are different ok. Now where do we use mean, median, mode? That I will come when I will discuss about the normal distribution also and also I discussed when I discussed about this scales, the

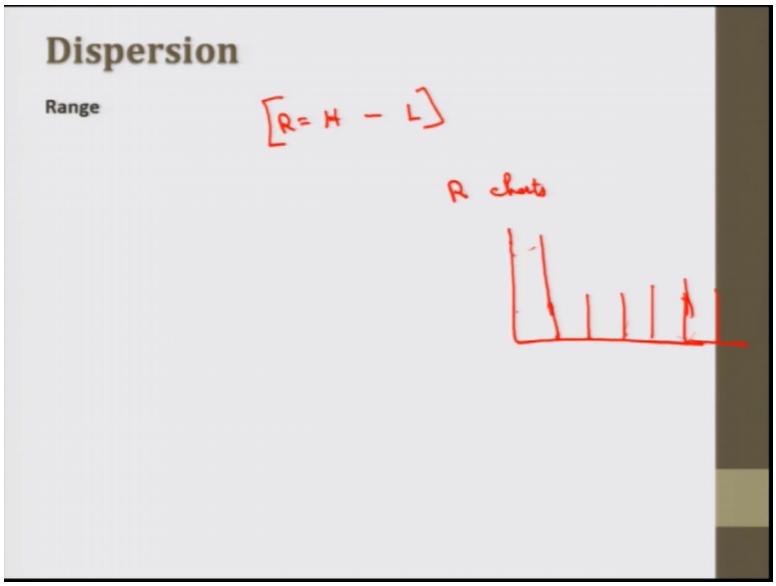
nominal, ordinal, interval and range scales, where can we use mean, median and mode like a nominal scale, we cannot use 3 and in case of the ratio scale, we can have all the 3 values ok.

(Refer Slide Time: 12:02)



Next is dispersion, dispersion means the spread of data and the spread of data can be found from 3 different ways. Number 1 is range, number 2 is variance, number 3 is standard deviation. Now these terms range, variance, standard deviation, they assign a numerical value to the amount of spread of data set. So, it is the spread of data and we assign a numerical value here ok.

(Refer Slide Time: 12:42)

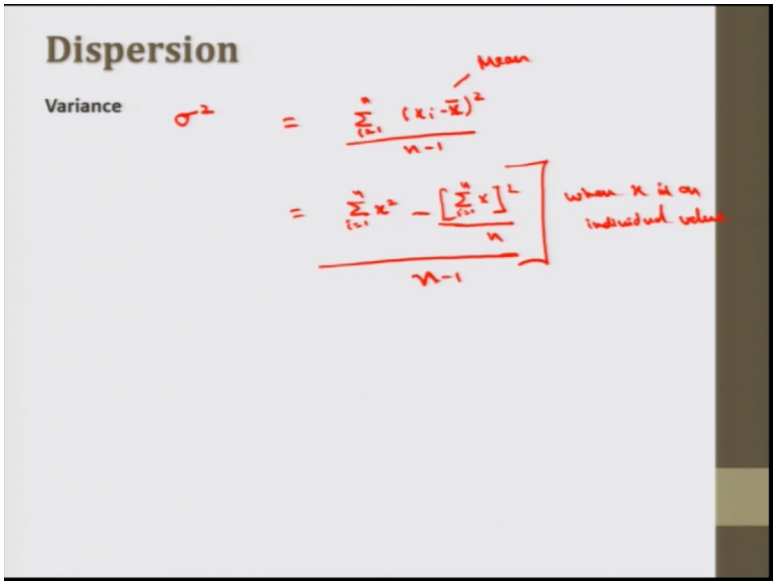


So, number 1 is range, what is range? Range is actually as, I said the highest value is there lowest value is there midway range was the average or the middle value of these range is nothing, but the difference between 2, R is equal to H minus L, where H is say the highest, L is the lowest value and the range is the simplest measure of the dispersion ok. So, we will draw the range or R charts and what are the significant within a subgroup and we have different samples and for each sample, we have different observations. For instance, you need to find the diameter of the shafts and you select the shaft 10 shafts after each hour in a production facility, after each hour and the in a day in a 8 hour working day, you set it 8 times. The sample size can be you can say, let me let me say you select shafts the sample size is 10 and in a n 8 hour day you select it 8 times ok. So, in that case within one group within one group that consists of one group is that is within one selection within one lot that is of 10 shafts of which, you are interested in finding diameter. What is the range of difference? What is actually range? Actually what is the range, which is the difference between the highest diameter and the lowest diameter and in the second group, what is the difference? So, we generally draw we will draw it actually in x bar and R charts, this is the R charts only we will draw we will see.

Where is the range varying for instance let me say, at in the start of the day, the range might vary like this and during the day range might be quite close. So, this means that while you make the setup, the very first value shaft that is manufactured or the when we

change the shift the start of the day, there might be some difference in the diameter that is occurring ok. So, this inference can be drawn from this.

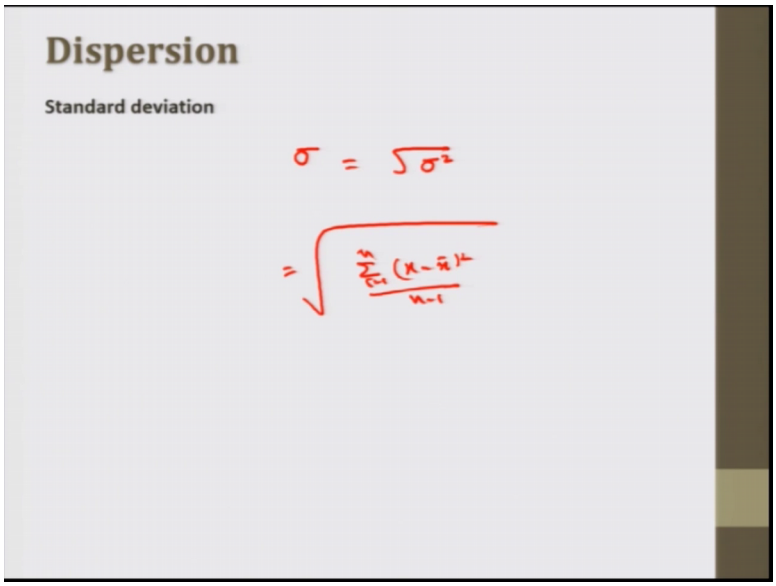
(Refer Slide Time: 15:00)



So, this is the application that will do in quality control. So, just this is the lecture to prepare for the next discussion that we are going to do, on the quality control on the normal distribution, etcetera. Next is variance, variance is denoted by sigma square. It is it is calculated such as summation i is equal to 1 2 n x i minus x bar whole square by n minus 1, where x a bar is actually mean ok.

So, another way to calculate x bar is mean, I can put the value here of x bar and open the bracket here, then it becomes summation i is equal to 1 2 n 4 x square minus summation x i is equal to 1 2 n whole square by n over n minus 1 ok. This is taken separately here. So, when x is an individual value, the mean and the n of in the sample size, the simple formula is used here when x is an individual value.

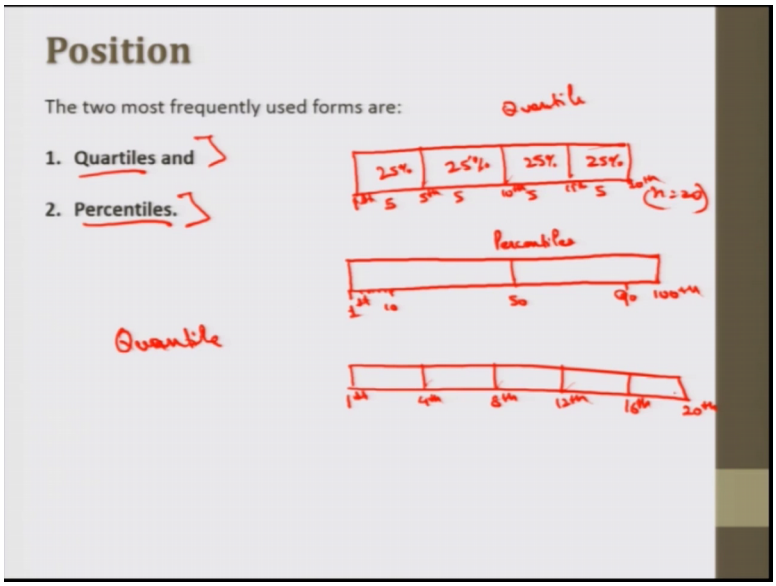
(Refer Slide Time: 16:43)



Next comes the standard deviation, it is denoted by sigma and it is actually the sigma is equal to under root of variance. So, if I take under root of these values, that is under root of summation i is equal 1 2 n x minus x bar whole square by n minus and you can derive the other relation for this as well, this is standard deviation.

So, standard deviation is one of the most used measures of dispersion, the reason for this is that there is square of the values the square of the differences ok, even when the differences lie above or below that is in towards the positive or the negative sides of the central tendency. It actually neutralize that effect, that is not taken into account in this, but range is simple linear or the arithmetic difference and standard deviation plays a vital role in normal distribution and in many distributions for the normal distribution, it is there are 2 statistic mean in standard deviation only.

(Refer Slide Time: 18:04)



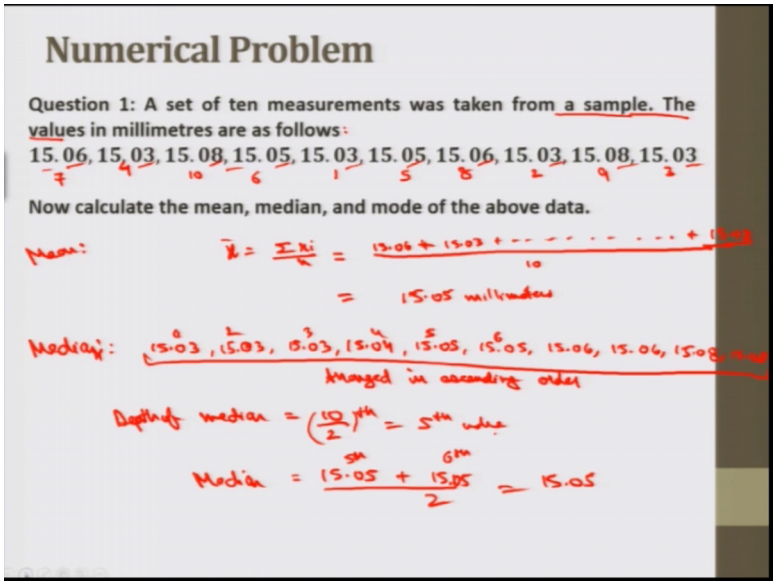
So, next I will discuss about the position. Position of the data is determined by the quartiles and percentiles actually, percentiles are derived from the quartiles only.

If I divide my data in the number of values, that we have into let me say 4 parts and I say, 25 percent of data is here, 25 percent here, 25 percent here, 25 percent here, that is if n is equal to 20, these values are 5, 5, 5 and 5 ok. You can say, this is first value, this is fifth, this is tenth, this is fifteenth and this is twentieth for n is equal to 20 ok. This is known as quartile similar to quartile, we can have percentiles if I divided into 100 divisions. If I divided into 100 divisions and I put 1 here and 100 here, I can put 10 here, 90 here and in between I have 50, these values can be call the percentiles, there is another term known as quantile. I just picked the quartile and percentile because, these are the most used otherwise, quantiles can be any number, we can divide the whole data, the whole data into any number. For instance, this is 20, I can divide it into 5 parts. So, this becomes 1 or I can call it first actually, the first number this is hundredth number, this is ninetieth number ok. So, it is first and twentieth here, what I will have? 1 to 4, fourth then eighth, then twelfth sixteenth and twentieth.; so, this is quantile, percentile is 100 quartile is 4 any, quartile this is twentieth quartile.

Quantile twentieth quantile, I can divided into any point this tells us the position of the data, this would be used when we will plot the box and (Refer Time: 20:41) plots or when we have when we also need to know the skewness ok. Skewness the coefficient of

the skewness or the strength of skewness can be calculated using the percentiles, using quartiles and using median and this can be utilized in that.

(Refer Slide Time: 20:58)



Now, I will try to solve an example here, for all these values. So, a set of 10 measurements were taken from a sample, the values in millimetres are as follows ok. These are the values mean, when we need to calculate the mean, the mean was x bar, which is equal to summation x i upon n these are 10 values. So, I will divide it by 10 and the 10 values, all these values 15.06 plus 15.03, this is very simple thing, you people know I will just put it here, all the 10 values it divided by 10 and I will calculate it is mean, that is equal to actually all the values are getting 16, you know the interesting thing here is that the values are 15.06, 15.03. If I take themean of the values after decimal places and then take 15 into a count that can also solve. So, I can also do it like a in way fifth I can subtract 15 from each, I can just take the mean of 6 plus 3 plus 8 plus 5 plus 3 plus 5 plus 6 plus 8 plus 3.

I can take mean of these as well only these values that can also do ok, but I have just put it the value, the mean here would be 6, 3, 8, 15, it is sum is actually 50 bar 5 50. So, it is 15.05 millimetres ok. This is mean, next is median for the median as I said, we need to first arrange the data in an order, the order that we have selected or that is generally selected is ascending order. So, I arrange data in ascending order, the smallest value here

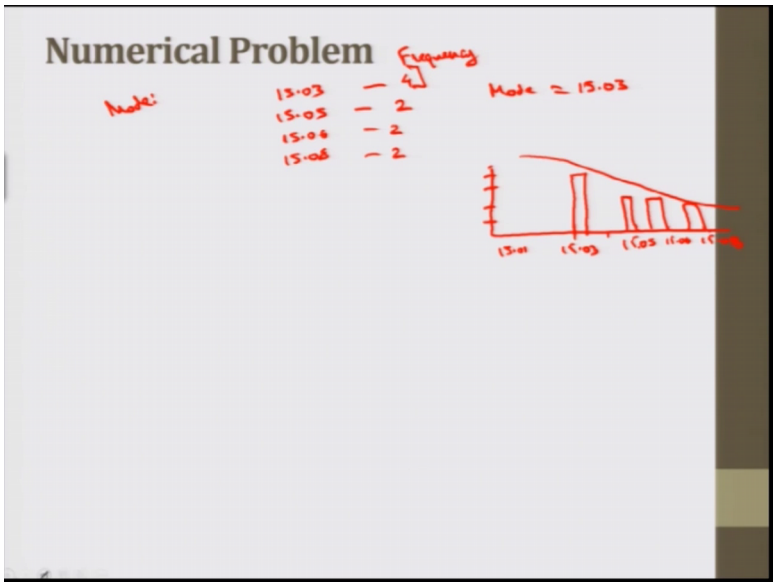
is 15.03 1 first value, second value is again 15.03, third value is 15.03, m is this one, this is fourth value 15.03.

Then fifth value here can be 5 15.05 then sixth value is another 5 is 15.05 then seventh value is 6 15.06, eighth value is another 6 15.06 and ninth value is 8 and tenth value is also 8, 15.08 and 15.08, when I arrange these into ascending order, these are actually arranged in ascending order ok. So now, we need to calculate depth of the median, since the number is even, when the number is even I would like to better explain it here only, for the median when the number is odd here it is when n is odd, but when n is even median is n by 2th value only, n is even when n is even. So, the number values are 10 here.

So, the depth of median I will calculate here, depth of median is equal to 10 by 2th 10 by 2 is equal to fifth value ok. So, when I say 1, 2, 3, 4, 5 and 6 ok. So, this is equal to depth of median is equal to fifth value and median is 15.05. So, in case of continuous data and in case of discrete data, there is a little difference, in case of continuous data in this case because, the number of values are even here, we can take the average of fifth and sixth value because, the average means something it can identify something, it is the length actually in millimetres ok.

So, we can say the median is equal to 15.05 plus 15.05, that is fifth value plus sixth value by 2 is equal to 15.05 ok. If the data is completely discrete, we can say there could 2 median values exist ok, we cannot just add the fifth and sixth value.

(Refer Slide Time: 26:39)

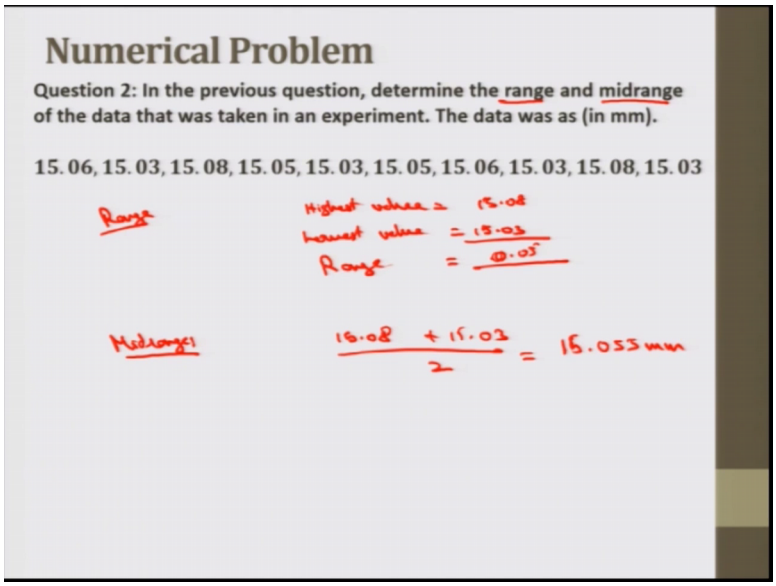


So, next we will calculate the mode, for this for mode let me see, how many times are these was repeated? 15.03 for mode, I need to see the frequency of 15.03, 15.05, 15.06 and 15.08, 15.03 is repeating, 1, 2, 3 and 4 times. I will put frequency here, it is repeating 4 times 15.05 is repeating 1 2, 2 times only, 2 times 15.06 is repeating 1 and 2, 2 times 15 is also 2 times.

So, this is 2, 2, 2 and 2. So, there exist a mode, mode is 15.03. So, mode is the maximum frequency that we have. So, therefore, mode is equal to 15.03. So, mode means the centre point in the frequency. If I say, this value is from 15.01 to 15.08 ok. These values are varying like this and this value 15.03 is repeating 3 number of times. If I call the frequency here, 1, 2, 3, 4 actually 15.04 is repeating 4 number of times and 15.05 is repeating 2 number of times, 15.06 again 2, this is again 2.

So, this kind of frequency distribution is here 15.05 and this is 15.06. So, the, we have mode here.

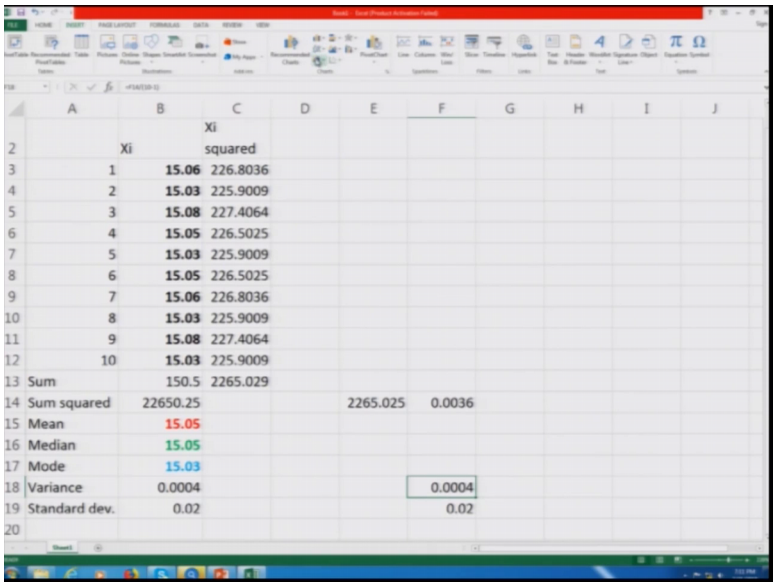
(Refer Slide Time: 28:38)



Next also we can calculate the measure of dispersion here ok. Again we have taken the same values and we need to find the range, midrange of the data that was taken in experiment, the range the maximum value here is for the range, range is highest value minus lowest value, highest value here is 15.08 and the lowest value is 15.03. So, this is the difference between 2 that is equal to 0.05, this is my range.

And midrange is midrange is 15.08 that is the highest value plus 15.03 by 2, this is equal to 11 15.055 mm this is midrange. So, as I said the purpose of this lecture is not to just recall the statistical concepts the fundamental [vocalised-noise] statistical concepts, that you have or you have to might have learnt in your matriculation on your on your senior secondary schools.

(Refer Slide Time: 30:20)



	A	B	C	D	E	F	G	H	I	J
2		Xi	Xi squared							
3		1	15.06	226.8036						
4		2	15.03	225.9009						
5		3	15.08	227.4064						
6		4	15.05	226.5025						
7		5	15.03	225.9009						
8		6	15.05	226.5025						
9		7	15.06	226.8036						
10		8	15.03	225.9009						
11		9	15.08	227.4064						
12		10	15.03	225.9009						
13	Sum		150.5	2265.029						
14	Sum squared		22650.25			2265.025	0.0036			
15	Mean		15.05							
16	Median		15.05							
17	Mode		15.03							
18	Variance		0.0004				0.0004			
19	Standard dev.		0.02				0.02			
20										

But to use the excel tool, we can also use R tool, if you like we can also provide you some notes on that to use the coding language or to calculate the simple observation of simple results, but for now, I will just use the excel sheet, I have put the data in the excel sheet here. This is length in millimetres, the data that was given first of all I will put numbers here, 1 2 you know if I select 1 and 2 together first of all if I select a single cell, if I select single it is selected using this green box is occurring here.

If I come at this if I come and the centre of the cell, this kind of plus is there, which is actually select tool if I drag it here keep on dragging, it will just select the column, it is it may select the row, it may select this column, it may select the whole matrix here, whatever you like ok. So, this is select tool and this is arrow type tool, this tool is the move tool, I can move it this value in any cell I like I will I will move it back.

Another tool is there, which is actually the formula drag tool, the simple plus is there if I select the single cell here, which is cell A 2, this is cell A 2, if you see this is column A, row 2, column A row 2. If I select this cell A 2 and drag it here, the number 1 will appear drag means the [vocalised-noise] that formula is actually dragging down here, but if I select these 2 cells, it will drag down and put the serial numbers from 1 to 10 ok.

So, I have just put it, put a serial number here because, I will need to I will also do the same calculations that we did the simple calculation with pen, I will also try those calculations in this excel sheet ok. How to find the median? How to find the mean in

excel?; however, they are direct formulas to find variance, standard deviation excel is a very versatile tool, we can do all the regression and many things like if we have to really exploit or explore actually the excel tool, it has many things in it has macro, it has a certain annotations and many things in design and all those things.

But the simple calculation if you like to do if I need to find the average of this data, I will just put I just select this cell, put this is equal to AVE, you know selection is they are options are coming here this is average of these cells, I have select this cells. So, this arrow this dots are moving; that means, this is being selected not selected being selected.

So, if I put enter here, it will give me average here, I will put it put average here or I will better put mean here. I have put mean here, the mean value is 15.05, I will color it red ok. So, mean value is there I can also find median here, median this is equal to median of these values, median is again 15.05. So, let me try to found find the mode.

For mode I need to put here, mode is also coming, if you try mode is 15.02 03. So, if I see these values these were exact, these are the exact values that I calculated using my pen only mean, median and mode. I will put different colours here, another way is if we need to arrange this in ascending order what I can do? I can select these values and then sort here sort and filter. Sort from smallest to largest or I can sort I can sort it from here sort from smallest to largest, we just sort the column A which is always this is already sorted from smallest to large.

If I do from largest to smallest from largest to smallest, you know the values are changed here, it has sorted from the 10 to 1. The value corresponding to 10 is 15.03 only we can pick any value, value corresponding to 5 is again 15.03 value corresponding 3 is 15.08, if undo this control plus, these undo you can see value corresponding to 3 is 15.08.

So, I like to sort the column B, these values I like to solve these values in ascending order. So, I will do custom sort here, custom sort here as you select sort by not column A. I will select sort by length in millimetres, which is column B value is smallest to largest ok. Now these values are arranged in ascending order, if we need to calculate the median in a way that, we said if we need to find out with a 10 values, we can find the medial value, this is at sorted in an ascending order.

But, I am more interested in finding the standard deviation here. So, I am going to pick this second formula, it is x square summation of x square minus summation of x whole square by n actually, this is summation of the values the squares of the values, this is sum of the values and the square of that by n and by this is computed by n minus 1.

So, I will pick this formula to do my calculations here. So, this is x i, the column B contains x i here from B 2 to B 11, I will better put some in between here, I am just moving it down I will put sum the summations sum this is equal to sum of if I put is equal to sum this is equal to sum, if I put sum here this is equal to sum of these values enter, sum of these values is 150.5 ok.

So, now I will square these values, this is equal to 15.06 of value B 2 hat 2, the symbol hat here means the power, the power of the number B 2 hat means B 2 power 2 B 2 power 2 would be square B 2 B 2 power 3 would cube, would be the cube of B 2 B 2 to the degree of 4, maybe B 2 power 4, I can put the values here, but I just I am not interested in the square here. So, B 2 square, this will square, this value, this is actually 15.06 square is equal to 226.8036. So, I will drag this formula here, let me see is this formula correct yes this is B 3 square ok. I will drag this formula to all this 10 values.

Now, I have x i square this is x I will put another row here, I can put just select an insert it will insert a row above this is x i, this is x i squared I will wrap the text wrap, text will bring it in like this x i squared. So, this is x i squared summation of x i square sum of this if you like I drag the formula in the column, I can also drag the formula in the row, this is sum from the values B 3 to B 12 these values which are in a blue box in a big blue box.

If I drag this formula here, it will also bring the formula to the C column. Now if I see this is sum of C 3 to C 12, this big blue box sum of C 3 to 12. So, I have got this summation of x i squared and this is summation of x i. So, the formula was summation of x i summation of x i squared, that is 2265, this is equal to I need to square this as well summation of x i whole square. So, let me put it here I will drag it further down, I will drag it down and I will put here sum squared ok.

Now, this value is equal to square of this value. So, the formula is. So, we call summation of x square minus summation of x whole square, this is summation of x squared actually, summation of x squared minus summation of x whole square by n and this whole the difference of this divided by n minus 1. So, using this formula summation

of x squared is 2265 and summation of x square summation of x square by n, this is equal to I will put here this is equal to summation of x square divided by n, n is 10 this is this value this is actually, summation of x square minus divided by summation of x square by n.

Now, difference of these values summation of x squared minus this. So, I will put this value here, this is equal to summation of x squared minus summation of x square ok. This value whatever this value is now, variance is equal to I will put variance here. Variance is equal to this value divided by n minus 1 10 minus 1, this is my variance 0.004.

So, what if I will put variance here, what if I just calculate the variance using sum formula in excel? So, variance I will just pick the variance, variance of these values just to check just to cross check this thing, variance of these values yes, the variance is same ok. So, we calculated the variance using the given relation or the formula also, we see that the variance is same like you can just find the variance, you can just put the command is equal to variance, it will just put the formula and calculate the variance. So, these are only 10 readings, we can have thousands of readings, we can if I have thousands of reading, we need to divide into subgroups. We had to do multiple calculations for that pen would not work, we have to use this excel ok.

So, variance also we can calculate standard deviation, standard deviation standard deviation is square root of variance this I will first calculate my standard deviation, this is equal to square root SQRT square root of this value, enter this is 0.02 what if I calculate standard deviation, if I have try to find standard deviation using the excel formula? Standard deviation of the values from B 3 to B 12 enter this value is same.

So, I pick the formula of standard deviation here and found the standard deviation value as 0.02. I calculate the standard deviation, it is 0.02, I put the formula for the variance it is 0.0004 calculated using the relationship it is same ok. So, this is actually, the use of excel we will keep using the excel tool for further calculation, we will also see when I will discuss in the next part of this lecture. The variance plots, I can see how can we plot variance graphics using excel sheet. We have insert here and we have many charts here.

if I open the chart, this is actually kind of bar charts, we have stacked bar, the standing bars, we have scatter diagram here and line diagram pie chart ok. So, I will discuss about data visualization, in the next part of this lecture then I will use excel to plot something.

Thank you.