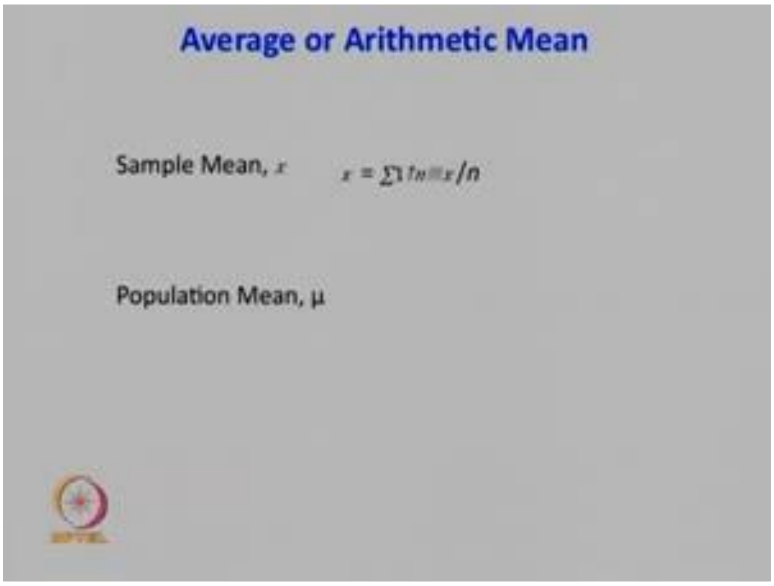
Introduction to Biostatistics
Prof. Shamik Sen
Department of Bioscience and Bioengineering
Indian Institution of Technology, Bombay

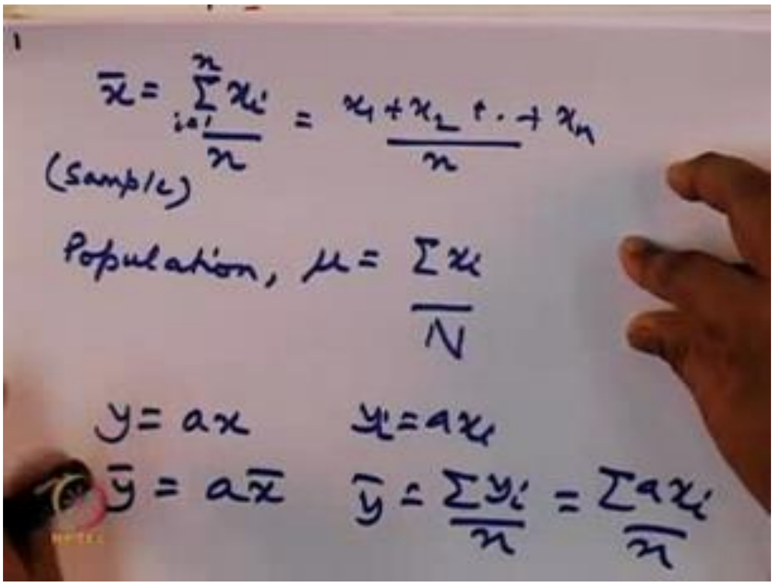
Lecture – 05
Measures of Variability, Standard deviation

Hello and welcome you all to today's lecture. Hope you have had the time to go through our discussion in previous lecture where we focused on three matrixes of quantifying data its mean, median and mode. Today we will begin with recapping what we are discussed briefly and then go on to other another important aspect of quantifying data which is to quantify the variation in data. So, let us begin by starting with our discussion of arithmetic mean.

(Refer Slide Time: 00:47)

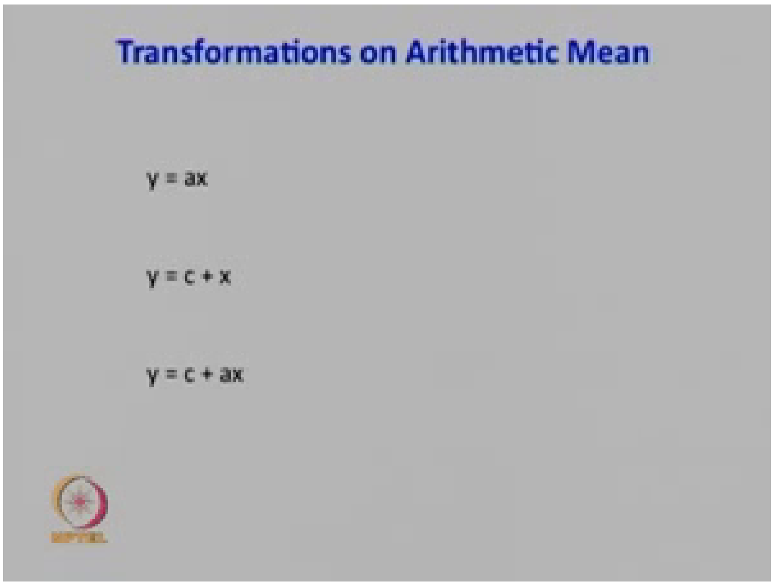


(Refer Slide Time: 00:55)



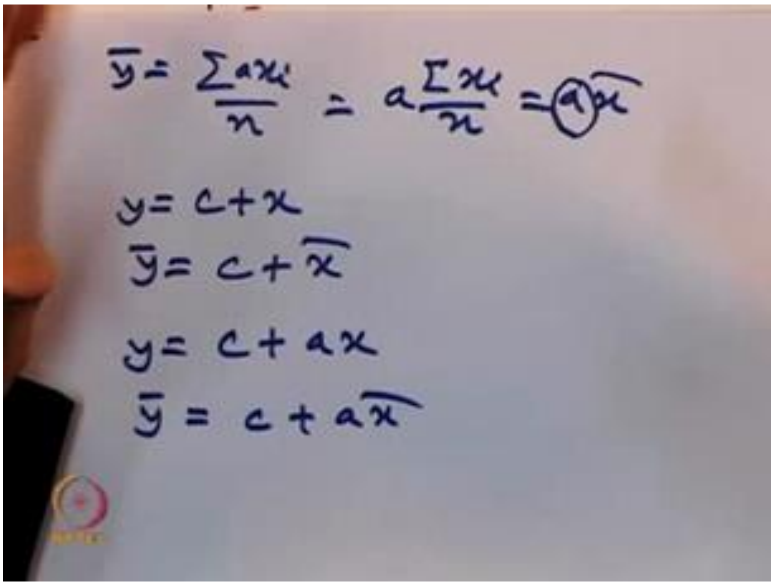
So, we had shown we are discussed that you could do arithmetic mean simply by writing x bar is equal to summation of xi by n where n is the number of observations in the sample. So, the summation means from i is equal to 1 to i equal to n. So, it is simply x 1 plus x 2 plus up to x n by n and similarly as per jargon if, this x bar is for a sample if you are doing for a population then you replace x bar by mu and it is summation xi by capital N. So now, we had then discussed about the different kinds of transformations which can which can be done on arithmetic mean and seen how you would vary the different values.

(Refer Slide Time: 01:35)



For example, if you have y is equal to a of x then we come to the conclusion y bar is equal to ax bar and it is easy to prove because if you have y is equal to a xi then y bar is defined by summation of y i by n is equal to summation of a xi by n. So, from then on, if y bar is equal to summation of a xi by n since a is a constant we can take it out and write summation xi by n is equal to ax bar.

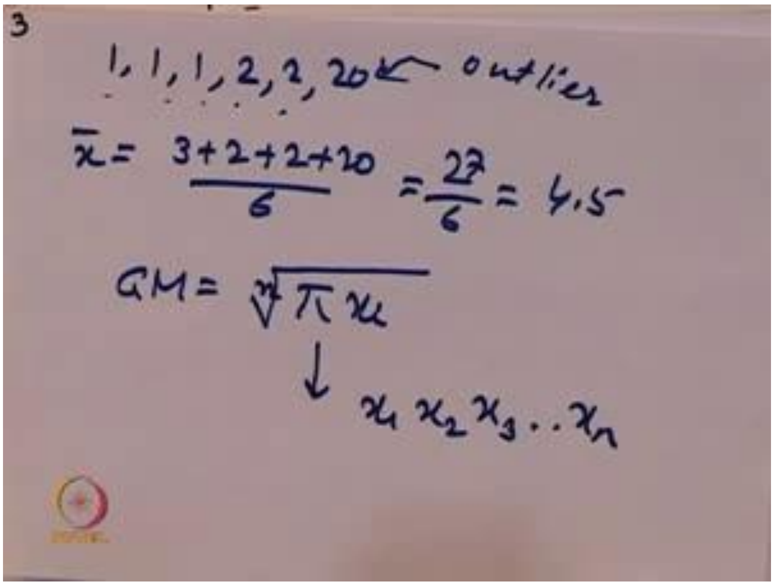
(Refer Slide Time: 02:10)



In other words when you have a p factor multiplied you know p factor a operated on x and that is how you calculate y you simply multiply a with x bar to obtain the value of y bar in the case of y is equal to c plus x then y bar is nothing since average of a constant is a constant my c plus x bar the third case in the general case where y equal to c plus a x then this would give me the formula y bar is equal to c plus ax bar. So, these transformations are particularly helpful when we are you know doing things manually by hand.

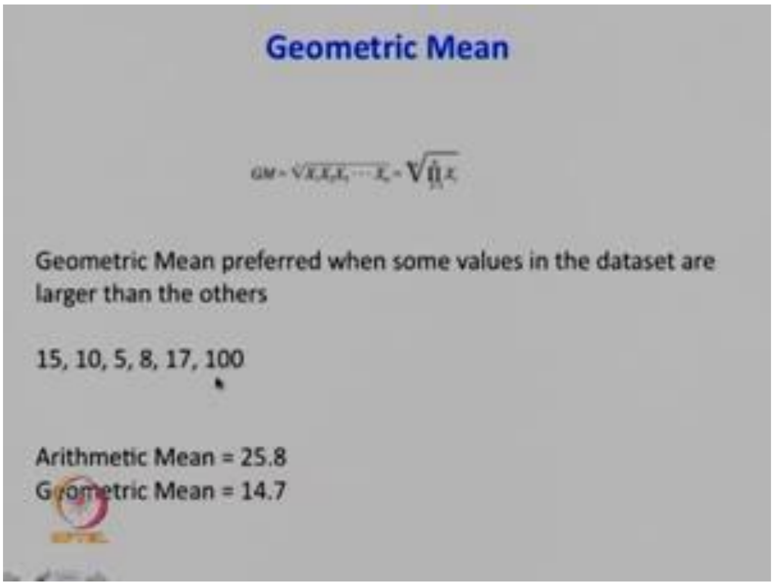
So, this is arithmetic mean and we found that one of the main caveats of the arithmetic mean is if you have a wide variation in your values.

(Refer Slide Time: 03:16)



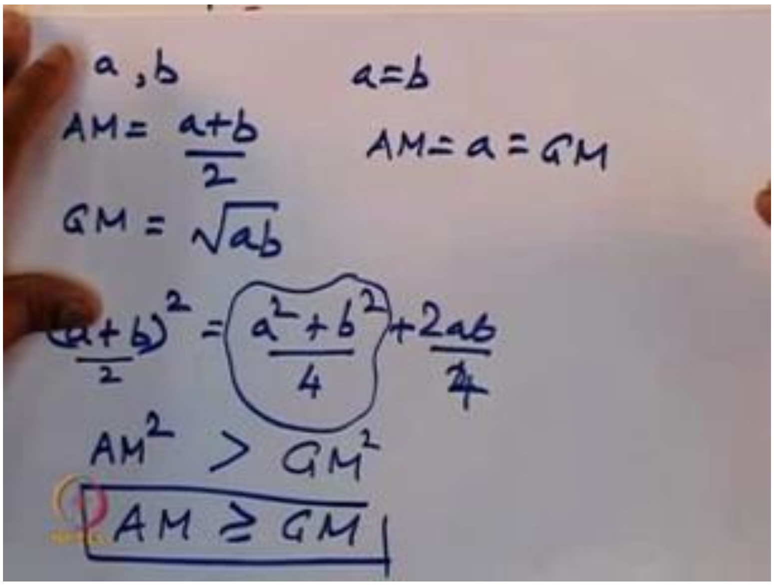
So, let us just take one particular example let say if I have my values as 1, 1, 1, 2, 2, 20. So, I have 6 values my x bar is equal to 3 plus 2 plus 2 plus 20 by 6; 7. So, 27 by 6, it will be 4.5. So, we can clearly see my numbers are 1, 1, 1, 2, 2, and I have kind of this outlier which is 20 this completely shifts my average to a value of 4.5 which really does not have any relevance to how my data looks in other words. This is one of the main deficiencies of arithmetic mean it is sensitive to outliers. So, one of the alternates to arithmetic mean this use this particular concepts of geometric mean. So, where we calculate geometric mean by. So, geometric mean is nothing but root of pi of x i and pi of x i. So, this is the nth root of pi of x i pi of x i means x 1 dot x 2 dot x 3 dot, dot, dot, x n.

(Refer Slide Time: 04:02)



So, now what we find is for these particular values that I have chosen 15, 10, 5, 8, 17, and 100. So, you can clearly see in that most of the values live within 17 except for this one number which is 100. So, if I calculate the arithmetic mean for this particular sample the arithmetic mean turns out to be 25.8 and as with the previously discussed case, we can see that 25.8 is much bigger than the number 17. So, the alternative if we calculate the geometric mean it gives me the value of 14.7 which is much closer to this population. So, this is an example which shows that arithmetic mean a. So, geometric mean is much less sensitive to variations in outliers compared to the arithmetic mean.

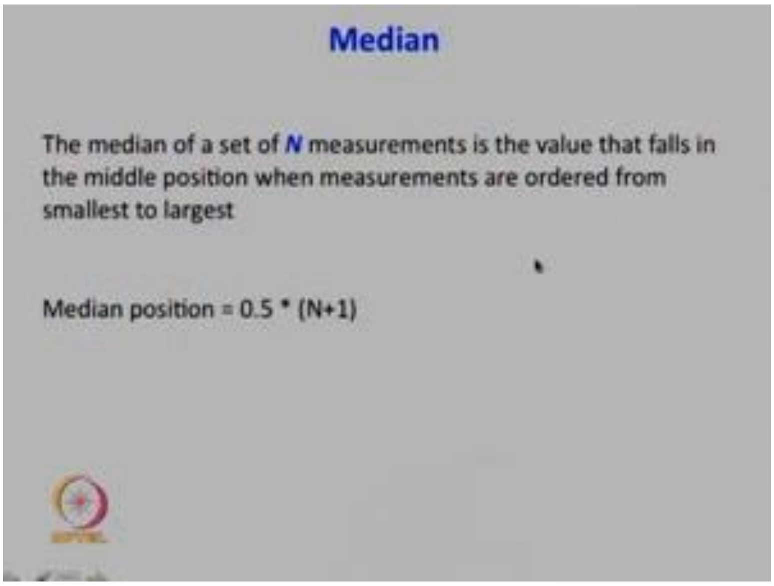
(Refer Slide Time: 05:28)



And in the generic case, in this particular case that we work out you find that the geometric that the geometric mean is less than the arithmetic mean and is it true for any data set. So, it can be shown. So, let us take 2 numbers let us take 2 numbers a and b; a and b. So, my arithmetic mean will be defined by a plus b by 2 and geometric mean will be square root of a b. So, is can I say anything us to how geometric mean and arithmetic mean relate to each other. So, I can write, if my a plus b a plus b whole square a plus b whole square is a square plus b square plus 2 a b. So, in other words you can see that a plus b whole square if I take, if I divide by you know by 2 a plus b by 2 whole square is a square by 4 by 2 by 4. So, a b by 2, since this is always positive since this quantity is always positive I can clearly say that, this term is nothing but arithmetic mean whole square and is, it is basically has to be greater than geometric mean.

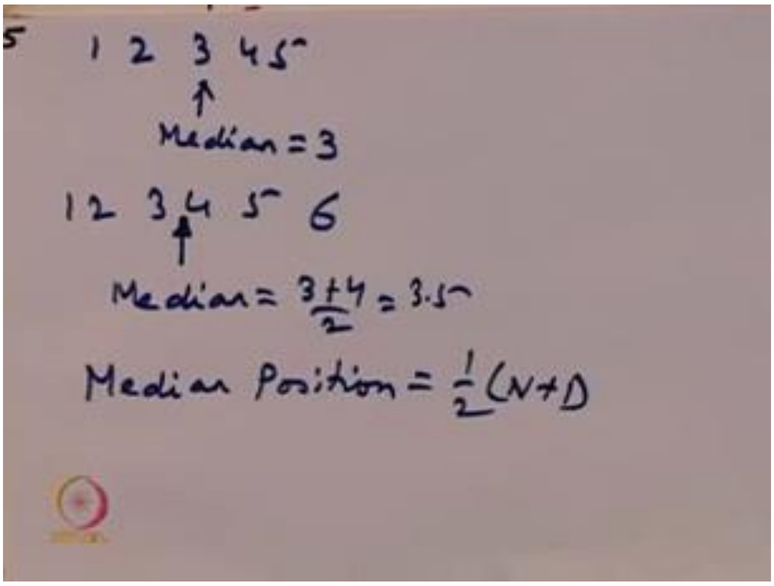
So, if geometric mean a square of a b. So, I can clearly say that 8 whole square in general case arithmetic mean is greater or equal to geometric mean. So, as we can see if a is equal to b for the case a equal to b; arithmetic mean is equal to a and equal to geometric mean. So, I can have this particular equation which says that arithmetic mean is always greater or equal to geometric mean. So, this is one of the reasons why your geometric mean is much less sensitive to extreme values.

(Refer Slide Time: 07:15)



Now, we have the next concept of median. So, the median the median of a set of n measurements is the value that falls in the middle position when measurements are ordered from smallest to largest. So, in other words the median position is 0.5 slash N plus 1. So, if you have 5 numbers let say 1, 2, 3, 4, 5, then your median number and this is a sorted data set. So, you can see that this is my median.

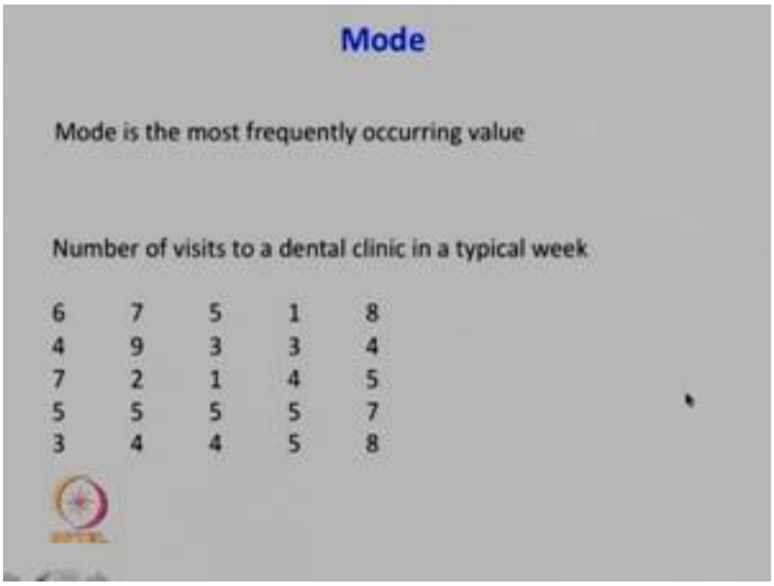
(Refer Slide Time: 07:29)



But if you have 6 numbers let say you have 1, 2, 3, 4, 5, 6, then the median position comes in between here and this is why your median is going to be? So, in this case median is equal to 3. In this case my median is equal to 3 plus 4 by 2 is equal to 3.5.

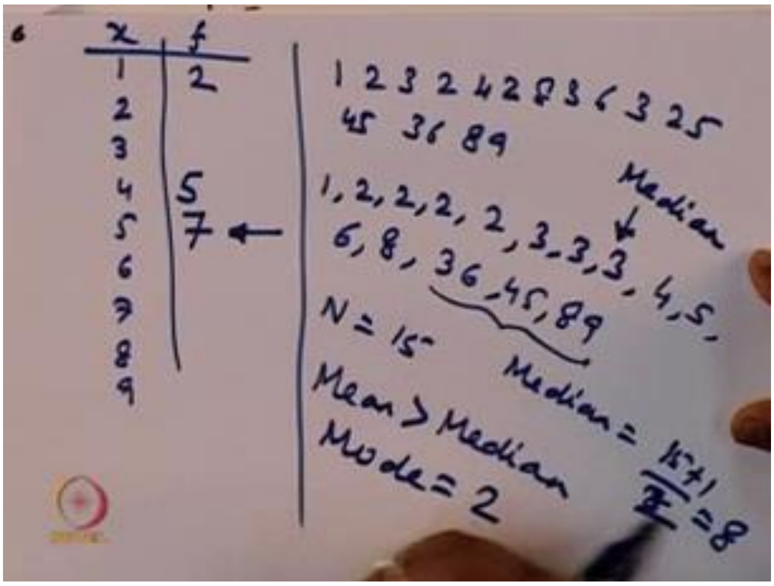
So, you have to find the position median position as half into n plus 1 and then find out whether you have to average between 2 numbers if your data set is even or you have an unique value if your data set is odd.

(Refer Slide Time: 08:20)



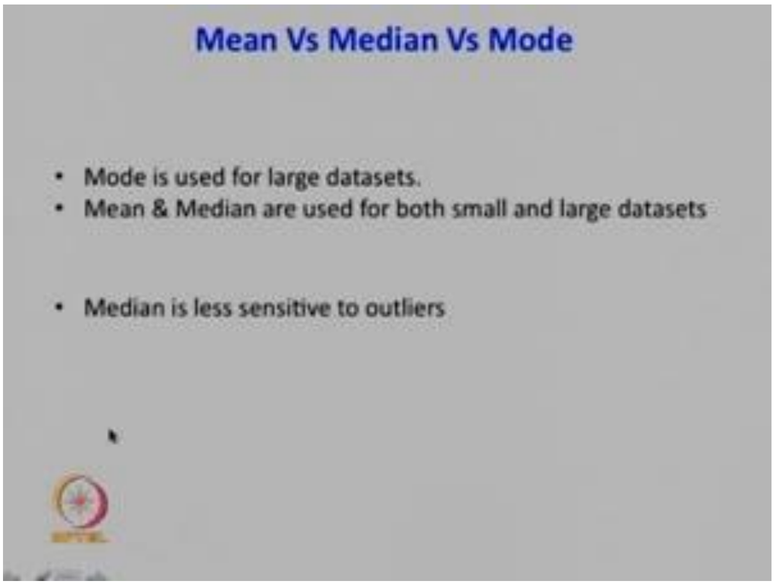
The third metric we had discussed was the mode. So, mode is the most frequently occurring value and this is for example, the number of visits to a dental clinic in a typical week this is the data. So, how do you calculate the mode? You would not first find out the frequency distribution. So, we can see that you have 1, 2, 3, 4, 5, 6, 7, 8, and 9, as the number of values I can correspond. So, this is x and this is my frequency f. So, I can see that for 1, I have 2.

(Refer Slide Time: 08:33)



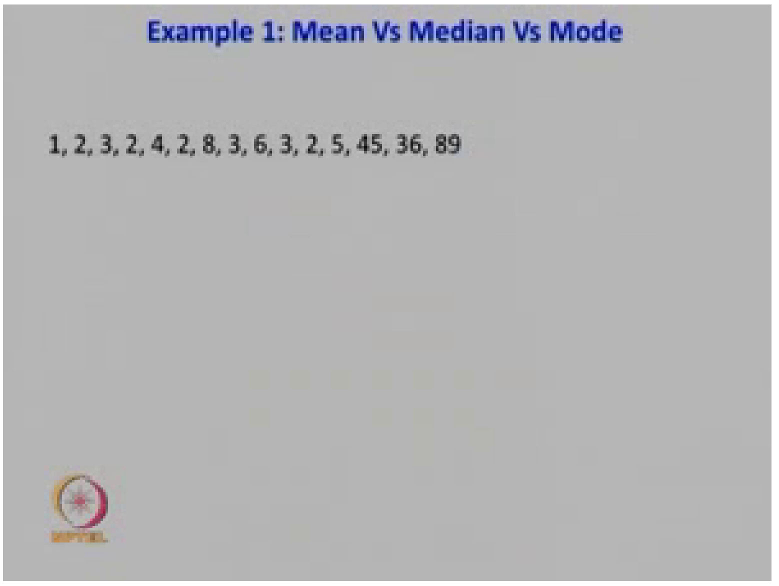
So, let us you know without going through the entire list I think my median values. So, for 5 for example, 1, 2, 3, 4, 5, 6, 7, there are 7 values for the number 5 and 6 is of course, much small 4 values is 1, 2, 3, 4, 5, 5, for 4 and if I am not mistaken, 7 is the value which is maximal occurring 8 is or so. So 7 occurs for the most number of time 5 the number 5 is the most frequent in other words 5 is your mode.

(Refer Slide Time: 09:24)



So, this brings us to the question as to which of these 3 values should you really you know consider making mean a median or mode and it is clear you know generally mode is used when you describe large data sets mean and median can be used interchangeably for both small and large datasets and as we discuss. So, again you know the arithmetic mean is of course; sensitive to its outliers, but the median is less sensitive to outliers. So, let us just do some few examples.

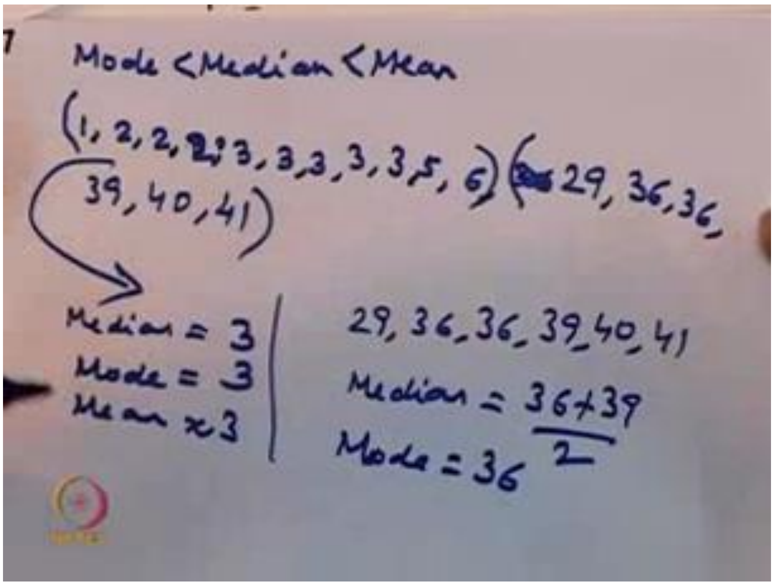
(Refer Slide Time: 09:53)



So, in this particular example, let say you have the numbers 1, 2, 3, 2, 4, 2, 8, 3, 6, 3, 2, 5, 45, 36, 89. So, if I were to arrange them in the proper order 1, 2, 3, 4, 4, 2s are there; there is 3 3s, 3 3s, 4 is just 1, then you have 5 then you have 6 then you have 8 then 36, 45, 89. So, I have my total n is 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 and 13, 14, 15, 15. So, my median position is going to be 15 plus 1 by 2 is equal to by 2 is equal to 8. So, my median is 4, 5, 6, 7, and 8. So, this happens to be my median right my average when I take the average because of the presence of these 3 numbers my mean is of course, going to be much greater than the median and mode is the maximum occurring value which is 2.

So, mode is equal to 2 median is equal to 3 median is equal to 3 and mean is of course, greater than median and is greater than mode.

(Refer Slide Time: 11:29)



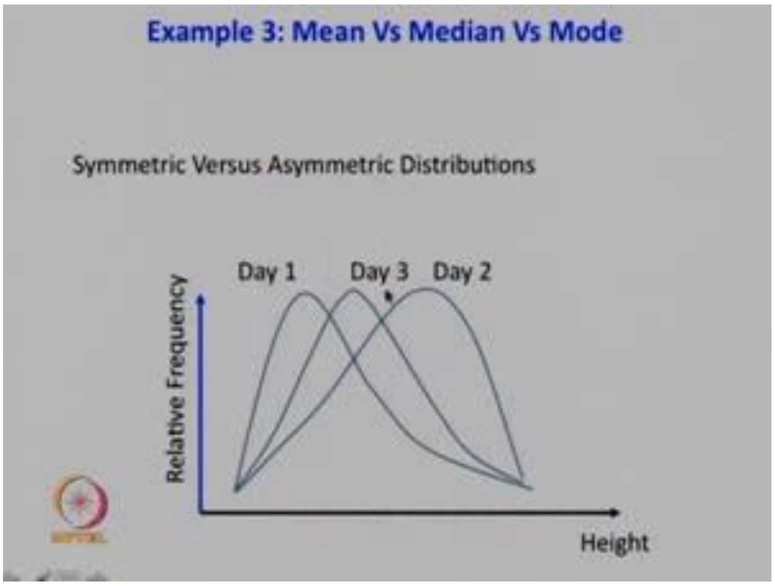
So, in this particular example, in this example that we worked out we came to the conclusion that mode was less than the median was less than the mean.

So, let us consider the next example this is a next example as again we can, if you look at the data set. So, and if I order arrange them in order to then 3. So, there is another 2, 3, 3, 3, 3, 3, there are 5 3s, there is no 4, there is 1 5, 1 6, then you have 36; no sorry, 29, 36, 36, 39, 40, 41. So, 39, 40, 41, what you clearly see in this data set if I can partition this data into 2 groups 1. So, there is, as compared to the previous case, where there was only 3 numbers which were huge here you have 1, 2, 3, 4, 5, 6, numbers which are reasonably huge.

So, it gives us the idea that you really have 2 subpopulations in this whole set. So, in this case neither the mean not the median not the mode would make sense. In fact, if you group if I were to group them separately for this group I can have work out my median 4, 5, 6, 7, 8, 9, 10, 11, 12, 6. So, median is 3, mode is 3 and mean will of course, will slightly greater than 3 because no it is you know; it will be approximately 3 only slightly higher than 3 and for this set you have numbers from 29, 36, 36, 39, 40, 41.

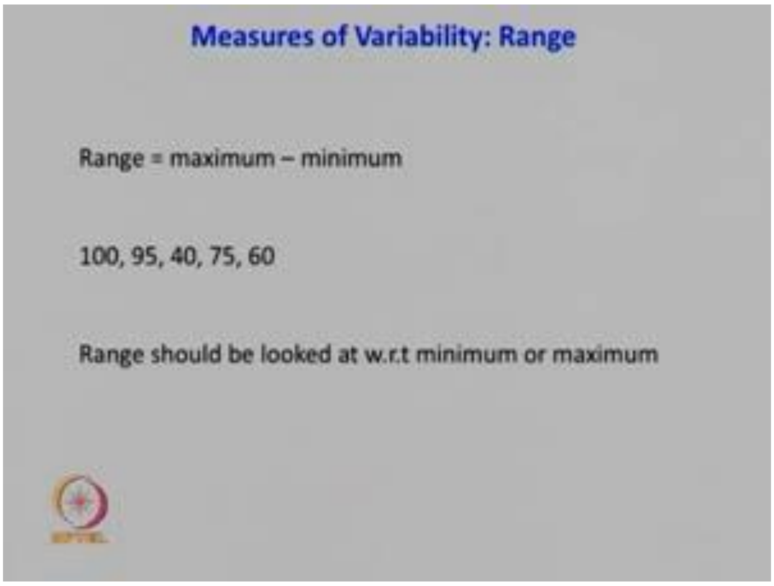
So, for these numbers my median 4, 5, 6, 7, median is going to be 36 plus 1, 2, 3, 4, 39, by 2 my mode there is no mode no mode is equal to 36 and median and will also be somewhere in between. So, what you see here is from the previous case where there were it seemed that there were only 3 outliers in this particular case they are clearly 2 different sets. So, it begs the question that what would be the best way of you know quantifying this kind of a data.

(Refer Slide Time: 14:27)



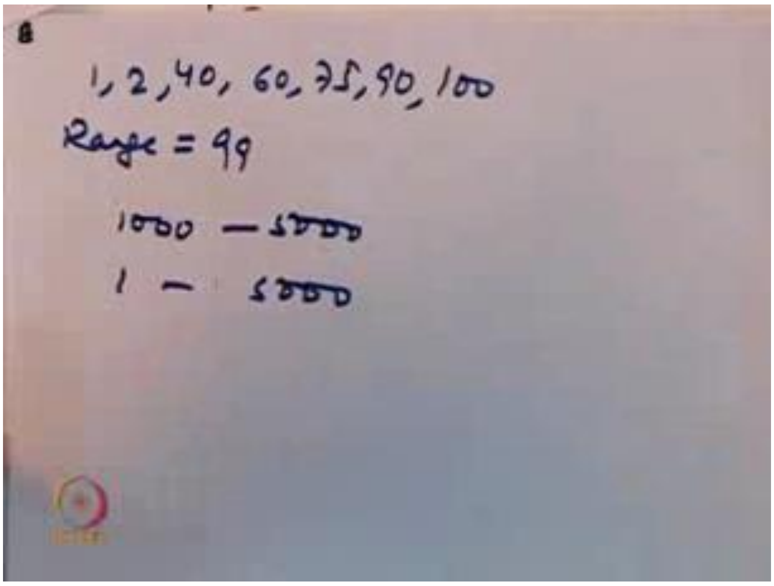
So, again let us take this particular example where you have symmetric versus an asymmetric distribution. So, in 3 different days you can have in terms of high profiles you can have 3 different distributions what you see is on day 3 the data is very symmetric on day 1, it is q to the left, on day 3, day 2, it is q to the right. So, it tells us that in addition to quantifying mean median and mode there must be other way of capturing this variation in this data and one of the measures which is very frequently used is this measure of range which is nothing but maximum minus minimum. So, I can define the range as maximum minus minimum.

(Refer Slide Time: 15:10)



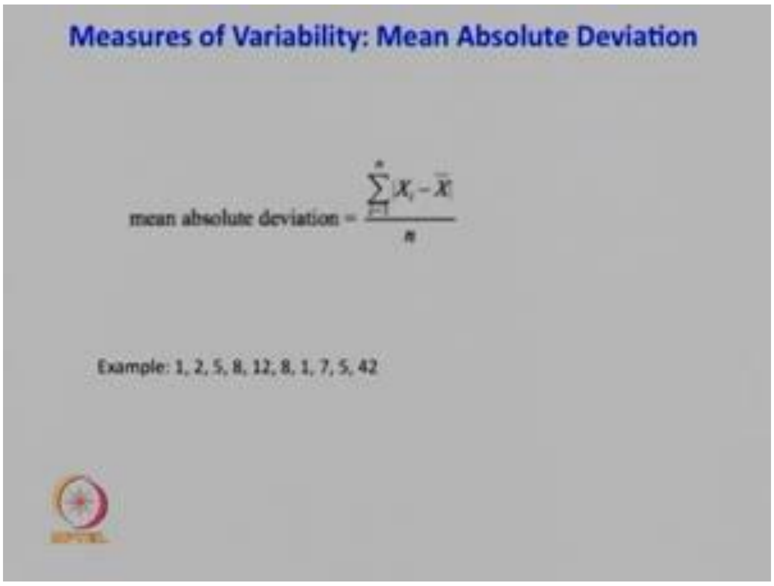
So, in this particular case, my minimum is 40, my maximum is 100. So, that brings us range is equal to 100 minus 40 equal to 60, but as you can see is the range by itself would not have any meaning unless the values are also put in context.

(Refer Slide Time: 15:39)



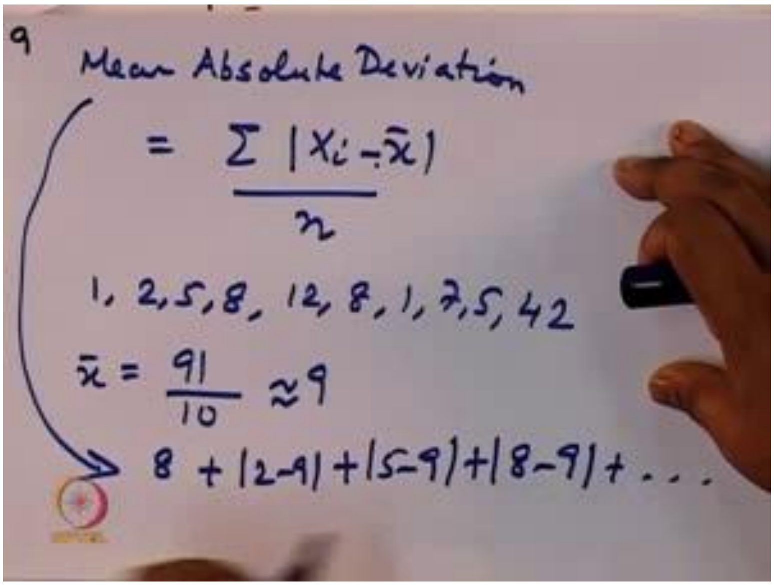
So, for example, if I have numbers as 1 2 and then I add these other numbers which is 40, 60, 75, 90 and 100 then my range is 99 versus in the case it was 60. So, the concept of range has to be thought about in the with respect to the minimum or the maximum. Similarly for example, if you can have data going from 1,000, all the way to 5,000 or 1 to 5,000. So, it does not. So, your range has to be you know thought about in the concept of your maximum or minimum. So, if you have again outliers then the range is too broad it does not particularly give a clear data as to where bulk of the data is situated.

(Refer Slide Time: 16:30)



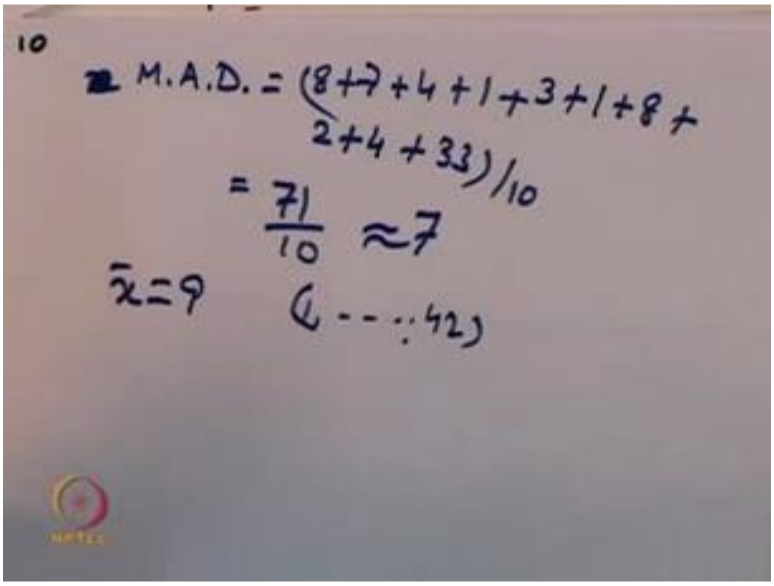
So another way of measuring variability is using the mean absolute deviation. So, we can work out this particular example.

(Refer Slide Time: 16:38)



If you; your mean absolute. So, mean absolute deviation is summation of mod of x i minus x bar by n. So, in this particular case, let us say if I have a data set as 1, 2, 5, 8, 12, 8, 1, 7, 5, 42, I have to calculate my x bar. So, my x bar becomes 3 plus 5, 8, 8 plus 8, 16, 28, 36, 37, 44 plus 5, 49, 91 by 1, 2, 3, 4, 5, 6, 7, 8, 9, and 10. So, approximately 9 let us say so, my mean absolute deviation is nothing. So, this value becomes 8 which is plus 2 minus 9 mod of 2 minus 9 is 7 plus mod of 5 minus 9 plus mod of 8 minus 9 and so on and so forth.

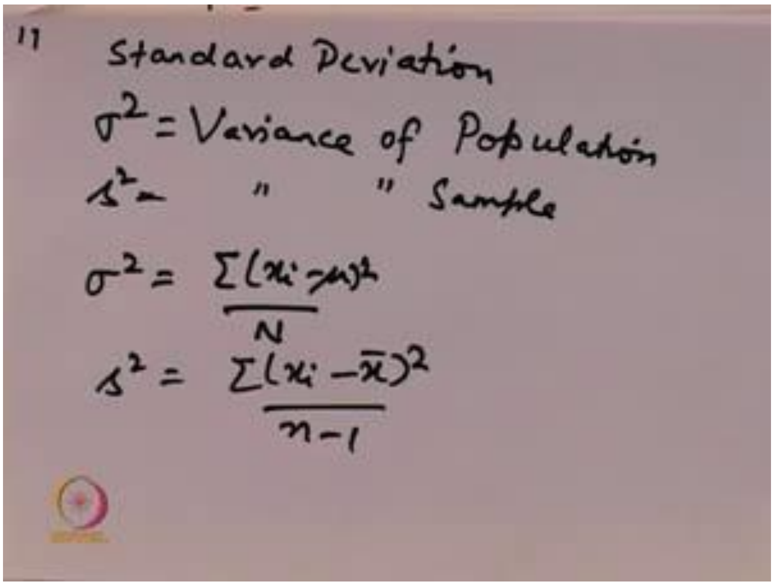
(Refer Slide Time: 17:55)



So, I can calculate this exact value, I can calculate this exact value as x bar as mean absolute deviation equal to 8 plus 7 plus 4 plus 1 plus 3 plus 1 plus 8 plus 2 plus 4 plus 33 whole divided by 10. So, it roughly comes to 8 plus 7, 15, 20, 24, 32 and 6, 38, 71 by 10 is roughly 7. So, as you can clearly see in your values that your x bar was now. So, x bar is 9 and this mean absolute deviation is 7. So, this reason because your value is ranged across a wide range from one all the way to 42. So, when your x bar this mean absolute deviation is comparable; that means that you have a wide heterogeneity in your data.

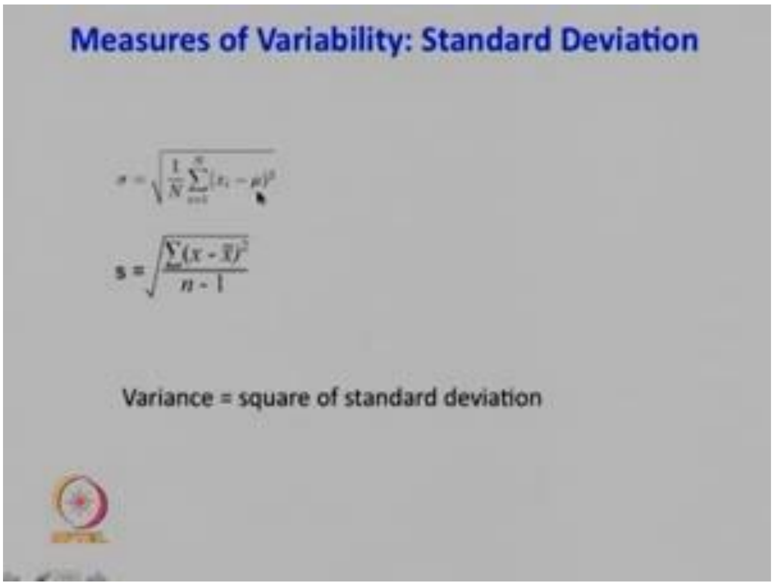
So, the most the most widely used metric the most widely used metric as a sign of deviation as a mark of variance is standard deviation.

(Refer Slide Time: 19:13)



So, I can let us come to the formula of standard deviation what you can see here you have, so instead of doing just the mean absolute deviation, you square the differences.

(Refer Slide Time: 19:22)

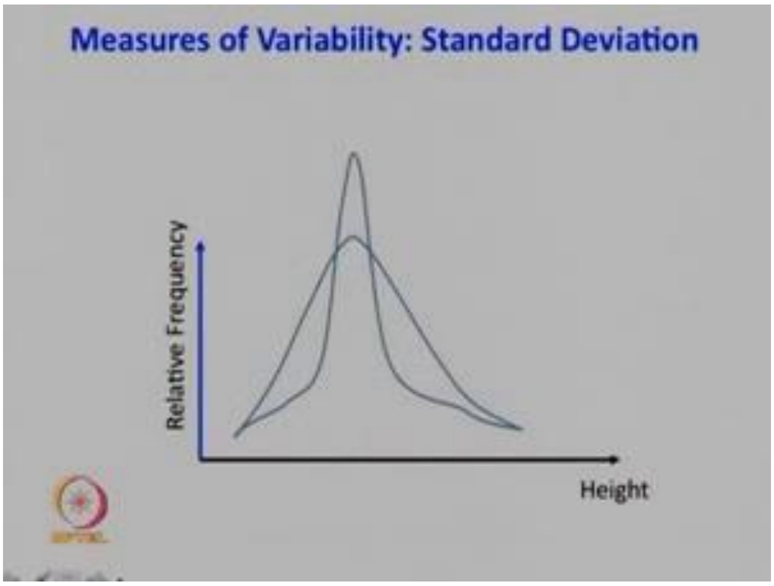


So, whether or not it is positive or negative whether you are you know your x values is less than the population mean or the greater than the population mean this square is always positive you add them up and then you divide by the total number of observation and you take a square because you had squared them up while adding. So, this is your definition of standard deviation for a population for standard deviation of a sample it is pretty much the same except there is a notable difference instead of dividing by capital N you divide by n minus 1.

So, this is the small difference in how you define the standard deviation between a population and between a sample and this the deviation of n minus 1 for a sample is simply to take into account that when your sample size is small when you divide by n minus 1 it gives the better estimate of the standard deviation of the whole population and variance. So, you can either. So, sigma square is equal to variance for the population and s square. So, variance of population and s square is variance of sample. So, sigma square is nothing but summation x i minus mu whole square by capital N and sigma square. So, s square is nothing but summation x i minus x bar whole square by n minus 1. So, these are called variances.

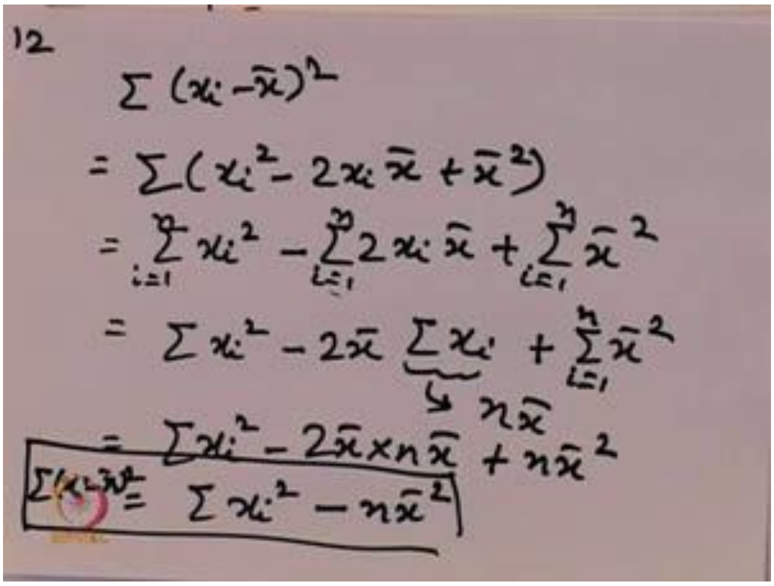
So, what you can clearly see is variance is just you know it is always positive and it is square of the standard deviation.

(Refer Slide Time: 21:13)



So, how do we you know go about computing the variation, but you can clearly see these are 2 distributions you can see that in one of them it has a much you know prominent peak in the middle and then these other values are less prevalent verses the second distribution of you know is much more broader. In other words if we calculate the standard deviation it will turn out that my standard deviation for this population is going to be smaller than the standard deviation from this population. So, this is what the variability will convey now I can there is just one small mathematical trick.

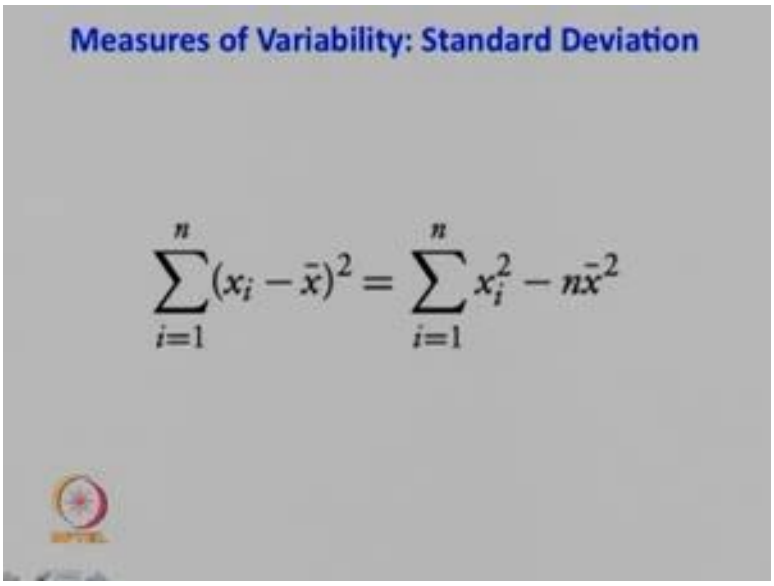
(Refer Slide Time: 21:52)



So, when I talk of you know summation of x i minus x bar whole square. So, I can let us. So, the, I can expand it. So, this I can write it as x i square minus 2 x i x bar plus x bar square. So, I can then bring it out I can write summation x i square minus summation 2 x i x bar plus summation x bar square.

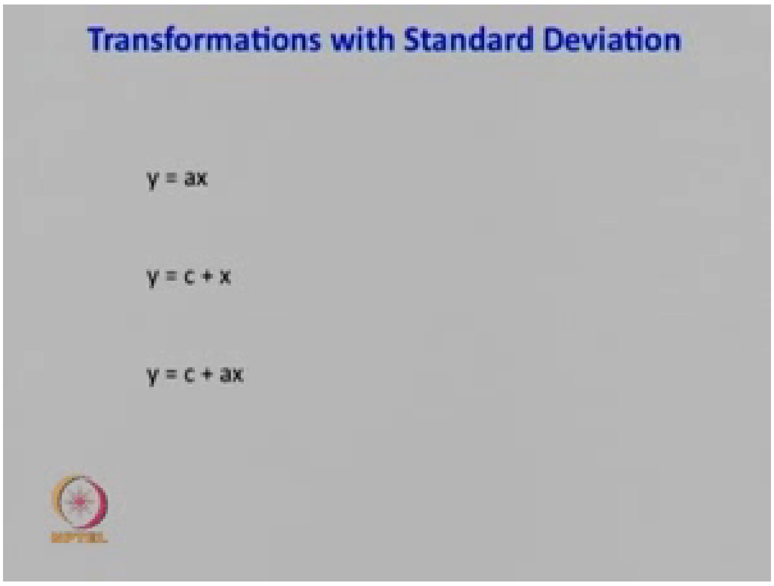
So, each of them is i is equal to 1 to n i is equal to 1 to n i is equal to 1 to n. So, this remains as summation x i square, but in this particular term since x bar is the mean I can take it out. So, I can take out 2 x bar summation x i and I can write plus summation x bar square. So, summation x i is nothing but n times x bar. So, this equation then becomes summation x i square minus 2 x bar into n x bar and summation x bar square summed up n times this is also i is equal to 1 to n this will be n x bar square. So, this final expression is summation x i square minus n x bar square. So, this is a useful formula when we are doing it.

(Refer Slide Time: 23:25)



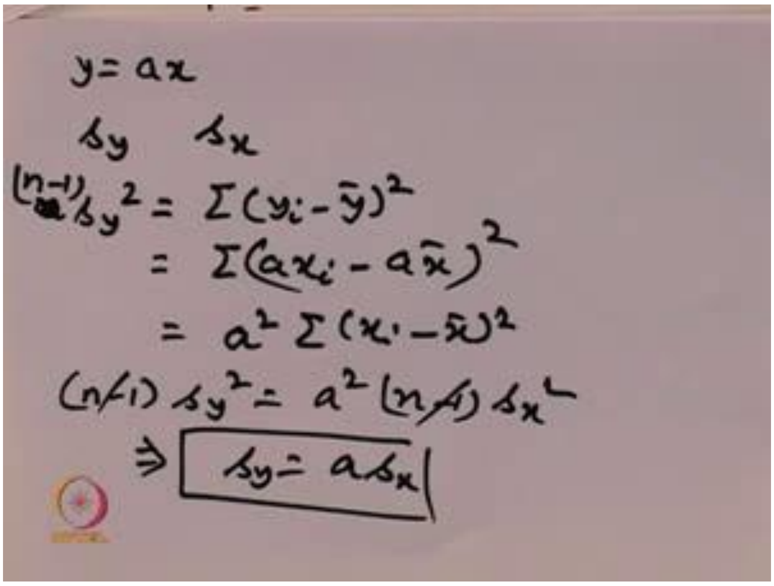
So, this is what I have written here that your measures of variability this thing can be simplified to form this. So, as opposed to taking the difference from mean if you have x i you can just add them up and then you know, you are calculating sample mean or population mean and you just my subtract is n x bar square to obtain this particular value.

(Refer Slide Time: 23:55)



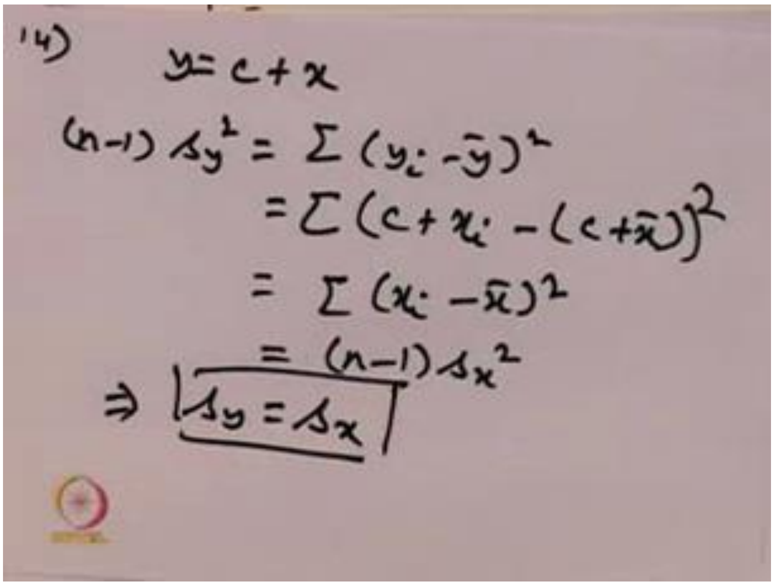
Now, let us do some transformations of standard deviation transformations with standard deviation. So, we again come to this particular you know term where you have 3 particular cases y is equal to a x.

(Refer Slide Time: 24:00)



Let us say, if this was my sigma y if s y the question is how is s y and s x related. So, what is s y n s x what is the relationship between s y and s x. So, the way to do it so, I know my s y. So, let us say if I were to do s y square or let us say n is n minus 1 s y square is nothing but summation y i minus y bar whole square now y i. So, I can put it as a xi minus ax bar whole square is nothing but a e can be taken common s square into summation x i minus x bar whole square. So, I can write n minus 1 into s y square is equal to a square into this term is n minus 1 into s x square. So, this would give to me that s y is equal to a into s x. So, I can cancel this terms out and this is the final formula which remains a. So, s y is nothing but a into s x for this particular case.

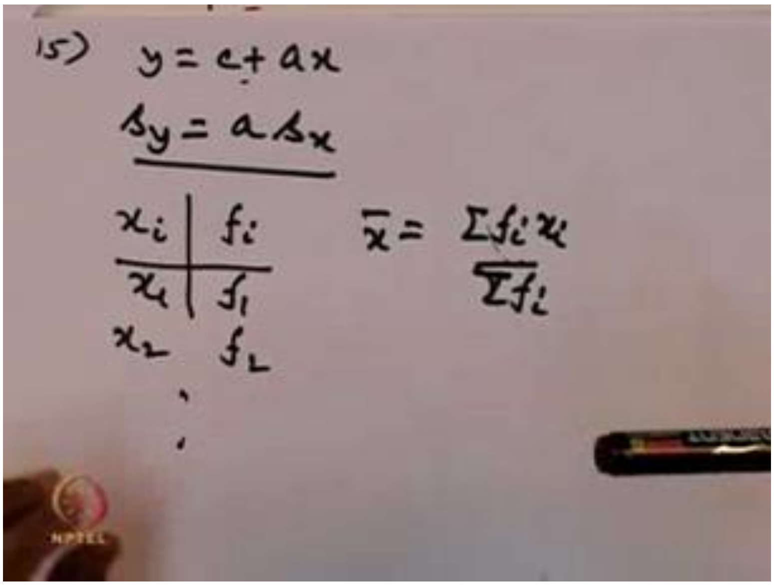
(Refer Slide Time: 25:25)



So, let say if my y is defined by c plus x then I can think of writing similarly I can write. So, you see constant. So, n minus 1 into s y square is equal to summation y i minus y bar whole square, but you see in this case y i is c plus x i and minus y bar is c plus x bar whole square. So, I can deduct c from each other which is nothing but x i minus x bar whole square.

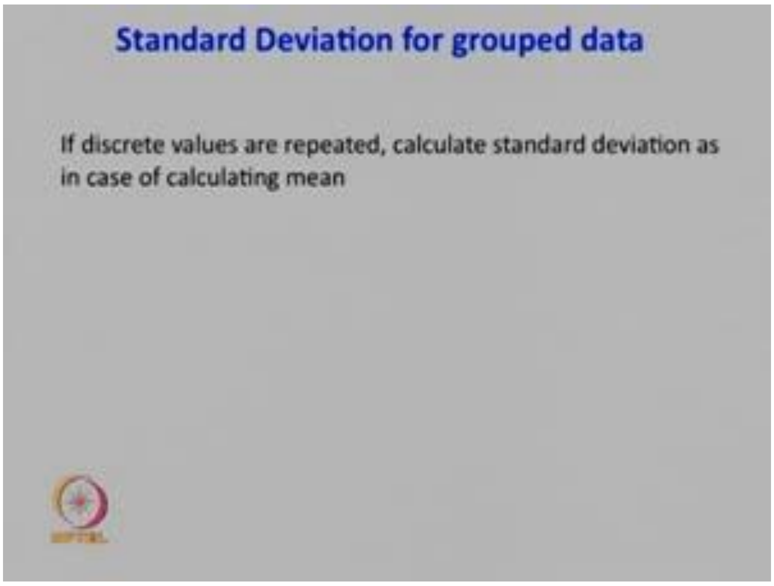
So, this is nothing but n minus 1 into s x square. So, this would give me that s y square s y is equal to s x. So, when you have a constant mean when you have a constant mean added to this value it does not change the final standard deviation. So, in other words, standard deviation is insensitive to any constant mean added the in the most general case when y is equal to c plus a x then by combining the previous concepts.

(Refer Slide Time: 26:26)



We can arrive at the equation s y should be simply is equal to s a s x because the c does not come into play while computing the standard deviation now this x this thing can be extended to find out standard deviation for grouped data for grouped data I mean that if you have x x i an you have an f i for the corresponding value.

(Refer Slide Time: 26:45)



So, x 1 is f 1, x 2 is f 2 so on and so forth, then I know my x bar is summation f i xi by capital N or by summation f I, all I need to do is to compute these frequencies and put them in place to get the final value.

So, that is all about the basics of standard deviation. So, I hope you understand. So, into ins to summarize we saw how from mean and median and mode how they can compare and what kind of values they are arithmetic mean is of course, sensitive to outliers median is not sensitive to outliers at all mod in the case when you have a bimodal distribution then neither measure makes any value it is better to split the data into 2 different distributions and then separately calculate their either mean median or mode for that from there we went on to discussing what is standard deviation and I hope you of you are convinced that standard deviation is a very important metric of quantifying how your values are dispersed across. So, mean itself by itself does not convey the picture of how dispersed your data is.

So, outliers will indeed have an effect in the standard deviation and one important point to note is when you have for the population you divide by n capital N when you calculate the standard deviation for the sample you divide simply by n minus 1 and this is because when your sample size is small then dividing by n minus 1 gives a much better estimate of the population standard deviation and we ended up by doing some few transformations just like calculating the mean for different transformations. I hope we have seen that how if you have a preterm or you know a constant it does not in have any impact on the standard deviation of the population, but when you have a pre factor a in front of x then your s 5 is simply multiplied you know s x multiplied by u a.

With that I would like to thank you for your attention. And will meet again in next lecture.

Thank you.