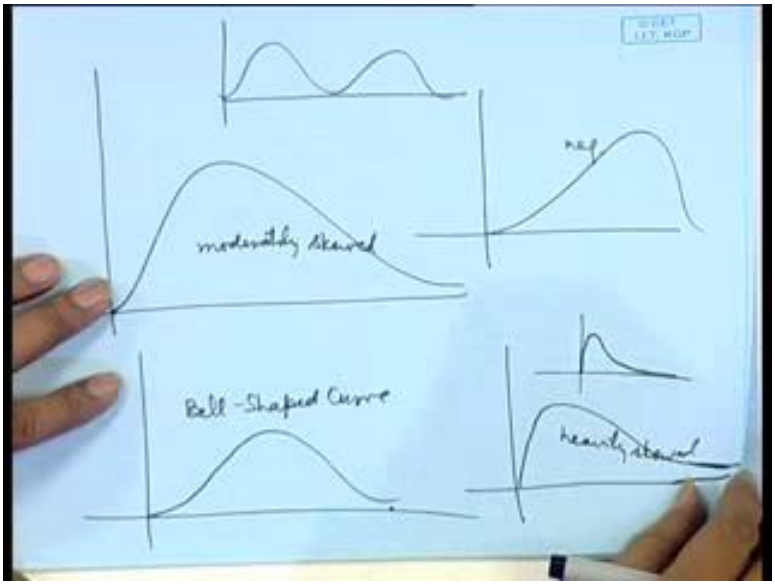
Probability and Statistics
Prof. Somesh Kumar
Department of Mathematics
Indian Institute of Technology, Kharagpur

Lecture - 52
Descriptive Statistics – IV

Make a histogram then the shape of the curves tell something about that.

(Refer Slide Time: 00:27)

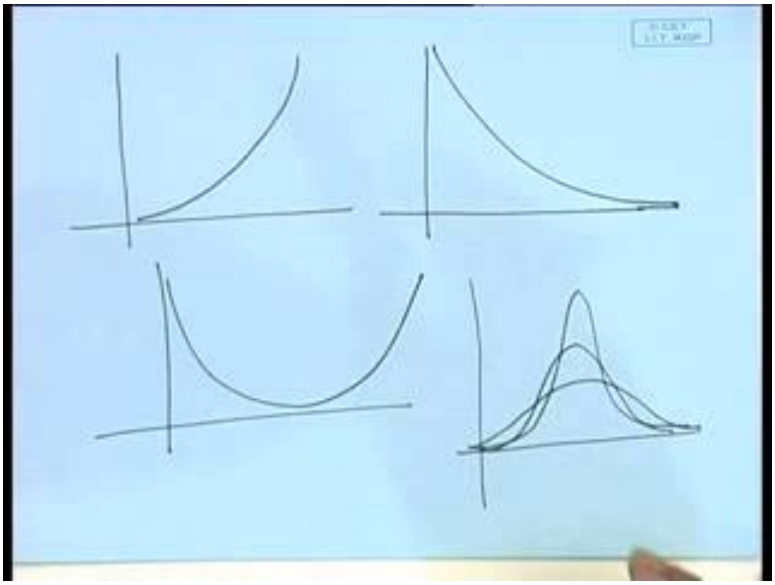


So, for example, if we draw a histogram and the curve is coming something like say this or a curve is coming like this or a curve is coming like this or a curve is coming like this or a curve is coming like etcetera. So let us look at these, if we observe this type of curve; it tells that the frequency distribution is quite symmetric in nature; that means, in the beginning the values are small then as the value increases then the value frequencies also increase, but after certain a stage; the frequencies again start to decrease finally coming towards 0.

This is known as bell shaped curve, which is more like a normal distribution. Or if you look at this; this is also bell shaped but slight is skew is there. So, this is moderately skewed; however, this one is heavily skewed. For example, we may consider, so this is heavily skewed distribution. Now here it is a skew is positive, here then a skew is negative. This shows that there is more than one peak of the distribution; that means, in the beginning it increases, then it decreases, then again it increases and then again it

decreases which is flowing somewhat unusual behavior; such curves are not observed much in practice; however, if it is observed then one should be careful.

(Refer Slide Time: 02:24)



You may also have curves of this nature; that means totally increasing or totally decreasing kind of thing. For example, this shows the frequency distribution of the families with certain income, so there are large number of families with very small income, there are very few families with a very high income; then there is a middle level; that means, a people with middle level incomes; the family with the middle level of income are also middle in number that is moderate in number, so this type of curve force this type of behavior. On the other hand, this type of curve may show some sort of age distribution for example, in a developed country like Australia or say France; there are less number of children and the number of people who are quite old because of the average life expectancy is high, so that number is d.

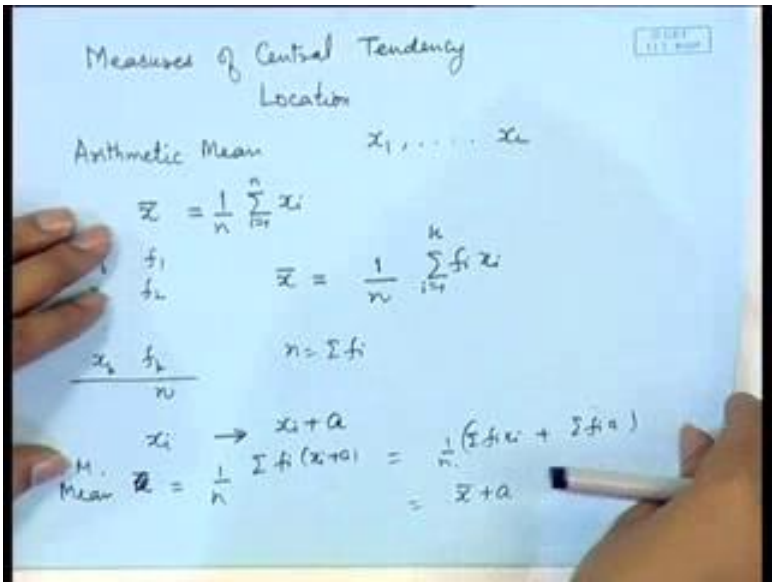
So, the curve shows an increasing trend these are somewhat unusual, but they also occur in practice. Another unusual type of distribution is something like a u shape; a u shape curve is quite uncommon because here it shows that in the middle the value becomes very low and in the ends the value become higher, but in some metrological data where you are looking at the say cloud in a (Refer Time: 04:19) or say rain falls, sometimes it may happen that in the beginning there is a more rain fall and in the end there is a more rain fall and in the middle there is less or the number of places where less number; less

amount of rain fall is there or more amount of rain fall is there is more and where moderate type of season is there that amount or you can say number is less.

So in that case you will in counter and unusual shape curve that is a u shape curve. So, you may also classify according to another characteristic something like you may have a distribution of this nature and you may have a curve of this nature, you may have a curve of this nature. So, this tells about the concentration of the values, this is partially a distributed, this is moderately is partly distributed and this is heavily concentrated distribution; large number of values are towards the center and very few values towards the end. Here the things are more or less equally distributed, so you can say a fact belly and tall belly. So, this type of curves are also common; now by looking at the different curves one gets a feeling about the kind of distribution that will be there.

So, for example if we observe heavily a skewed distribution then we know the nature of the data, if you know it is symmetric then we know the nature of the data, if it is a skewed moderately skewed or heavily skewed; we are able to make certain comment about the distribution for example, if it is increasing or decreasing or it is a u shape etcetera.

(Refer Slide Time: 06:10)



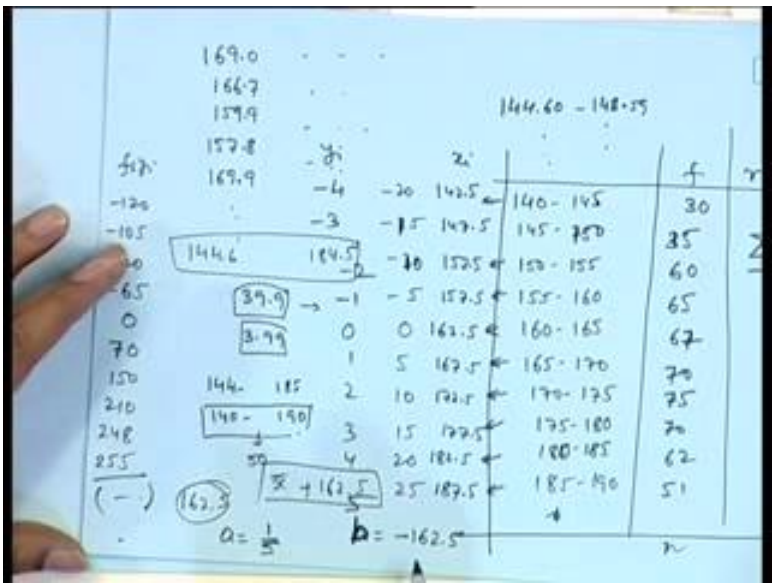
Now next is a measure; that means, from the given frequency distribution, we calculate certain measures which tell them about various characteristics of the distribution. So, the first of these are known as measures of central tendency or you can say measures of

location. Now this is something like even I have this one, we can say that most of the data is concentrated about the centre. Here the centre may be little bit shifted, the centre may be little bit on this side, here the centre may be little bit on this side of the more number of values because of the heavy that is given to this that there may be a long tail to the left and so on.

These are known as measures of central tendency or measures of location and there are methods of calculating that. So, here we will give example the first one is an average, so the easiest of the average is an arithmetic mean and arithmetic mean if I have the value say x 1, x 2, x n it is simply the 1 by n sigma x i; i is equal to 1 to n let us is call it x bar; that means, whatever observations are there, we simply take a plane average of that; this is known as the arithmetic mean. This is one of the most commonly used measures of location or measures of central tendency, now if I have a discrete frequency distribution; that means, x 1, f 1, x 2, f 2 and say x k, f k where total is n then the frequency distributions average value can be calculated using the formula 1 by n sigma f i, x i where n is sigma of f i.

If you have class intervals then; obviously, on the left side, we do not have single values of x i; the technique use this that one takes the mid value of the class interval and calls that as xi.

(Refer Slide Time: 08:42)



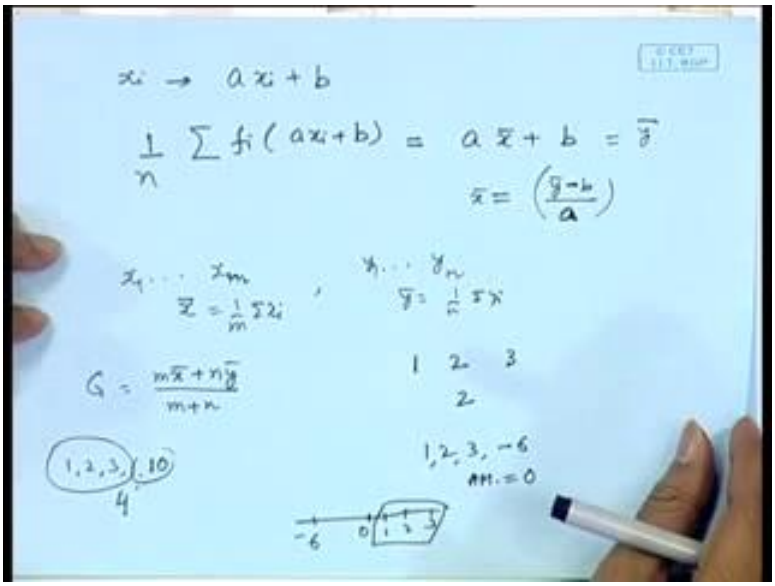
So, in the case of this particular example; you may take x i is as say 142.5, 147.5, 152.5, 157.5, 162.5, 167.5, 172.5, 177.5, 182.5 and 187.5. So, if these are the x i values and these are the f i values then again we can make use of the formula sigma f i, x i divided by n; where n is the total frequency here.

So this is the formula for the arithmetic mean; I must comment here that arithmetic mean is one of the most commonly used measures of central tendency; it is because of ease of calculation and also certain mathematical properties which it satisfies for example, if I look at say each value is shifted by suppose in place of x i, the values are shifted by a then what will happen to your average value x bar. So, b bar will become let me look at this formula 1 by n sigma f i; x i plus a; that means, mean of this which is also called A M. So, this is equal to 1 by n sigma f i; x i plus sigma f i, a. Now sigma f i is n, so this n cancels out and you are left with x bar plus a; that means, if each observation is shifted by a certain value then the mean is also shifted by that values.

Now this is actually giving a very nice method of calculation because many times we are dealing with either very large or very small values so we may shift all the values by a certain number to put them into a for example, if you look at these values. Now these values are quite inconvenient like 142.5, 147.5 and so on. What we can do is; suppose we shift all the values by say 162.5, then this will be minus 20, this value will be minus 25, minus 30; sorry this will be minus 15, minus 10, minus 5, 0, 5, 10, 15, 20, 25 and more over now if you see here I have shifted a as by taking minus 162.5. So, whatever average of this would have come actually it will come x bar minus 162.5.

Now using this; the calculations are much simpler because these numbers are quite nice round number multiples of pi. So, I can multiply and add. Another thing what has happened is that many minus values are coming. So, plus and minus values will cancel out each other and the total sum will become much smaller number. So, x bar can be calculated quite easily and then in that one you just add 162.5.

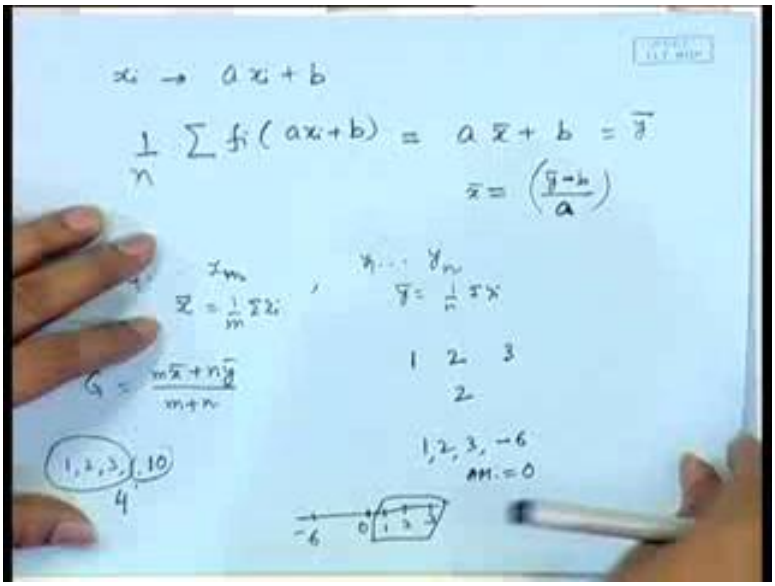
(Refer Slide Time: 12:19)



Sometimes we may scale also for example, I may shift x i by say a x i plus b, in that case your arithmetic mean will become 1 by n sigma f i; a x i plus b; this even if we expand then I get sigma x bar plus b. Now in the example of the heights, suppose I shift I take this let us say b is equal to 162.5 and a is equal to say I put 1 by 5 then the numbers are minus 4, minus 3, minus 2, minus 1, 0, 1, 2, 3, 4, 5.

Now, you see the calculations will become extremely simple; if I calculate suppose I call this y i and I make it able f i, y i then you see here the calculations minus 120. So, the numbers can be handled very quickly 105 then we have minus 120, we have minus 65, we have 0, we have 70, we have 150, we have 210 and we have 162 into 4; that is 248 and then 61 into 5 that is 255 and not only that there are minus values and there are positive values. So, if you add many of the values will automatically greater this step and this total will be become much smaller number.

(Refer Slide Time: 14:09)

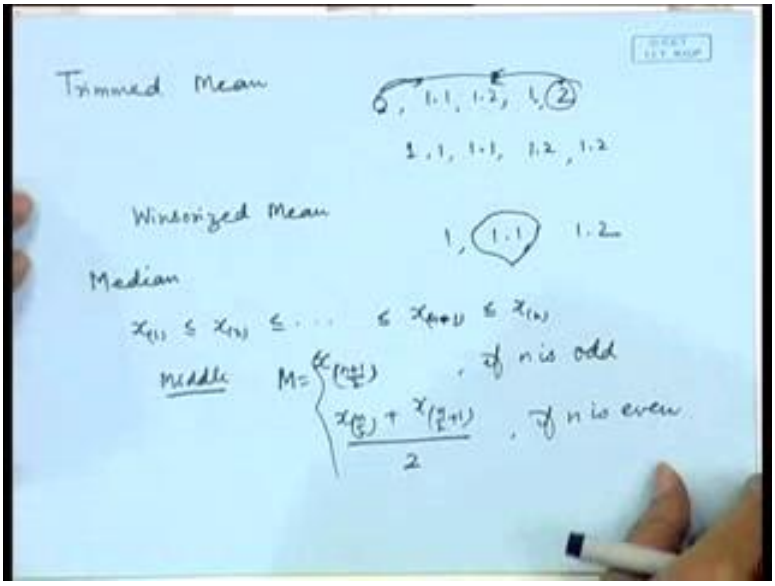


So, then in the answer what you do; if this is equal to y bar then x bar is equal to y bar minus b by a. So, whatever answer is coming we subtract b there and b divided by a and the arithmetic mean of the original sequence will come. Some other useful properties that the arithmetic mean satisfy, suppose I have two sequences of numbers. Suppose I say x 1, x 2, x n and say another sequence say- y 1, y 2, y n. So, let me put here say n numbers and here I put n numbers, so I can calculate the means of these sequences; this is x bar that is 1 by m; sigma x i, this is y bar say 1 by n sigma y i. So, if I am looking at the grand mean I can consider it as m x bar plus n y bar divided by m plus n; that is the grand mean of both the sequences taken together because m x bar gives the total sum of the first set of values and n y bar gives you the total sum of the second set of values.

So, if you add and then divide by the total number we get, so this arithmetic mean measure is amenable to various kind of manipulations and that is why it is quite useful; however, it has drawback, suppose I consider three numbers 1, 2 and 3 what is the arithmetic mean 1 plus 2 plus 3 is 6; 6 divided by 3 is 2. However, if I have 1, 2, 3 and another number is say minus 6; what will happen; the sum is 0 and therefore, the arithmetic will turn out to be 0. Now if you are looking at this sequence minus 6; 1, 2 and 3 then actually 0 is not occurring anywhere and in fact, 0 is less than 3 of the values. So, in some sense it is effected by the extreme values for example, in place of 1, 2 and minus 6; 1, 2, 3 and minus 6; I put say 10.

Then the arithmetic mean will become 4; which is again bigger then these three values and less than this one; much less than this. So, it gives (Refer Time: 16:49) to very high or very low values, so this is one drawback of this arithmetic mean. So, there are certain measures which are used here; for example, when arithmetic means are used for evaluation such as in a boxing competition of Olympic. So, there are 10 judges, who are evaluating at the end of each round they give certain marks out of 10 what the committee will do; they will discard the top value and the lowest value and then take the average of the remaining; this is known as trimmed mean. Another option is that the highest and the lowest values are brought to the next lowest or next highest.

(Refer Slide Time: 17:02)

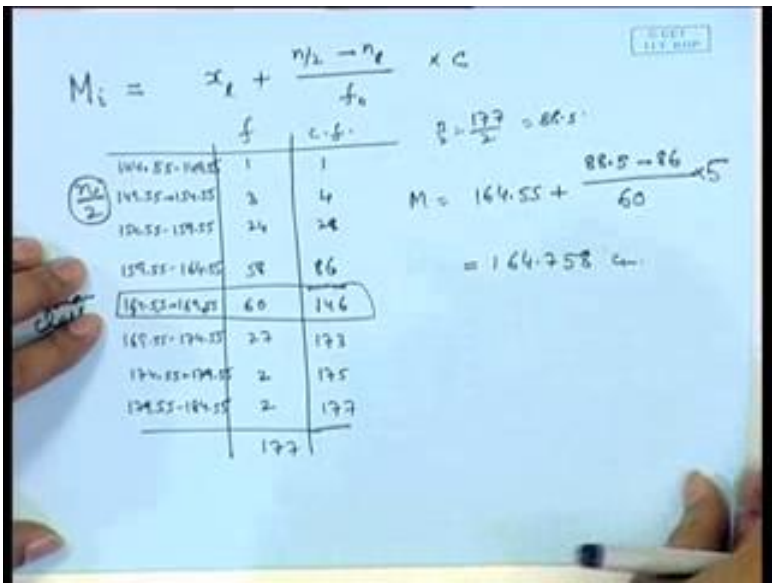


For example, if the values are say suppose I write five values. So, five values are 1.1, 1.2 and say 1 and another value is say 2 and one value is say 0. Then I convert this 0 also to 1.1 and this 0 to 1, so I will write the lowest value is 0. So, below that this is 0, so I will convert this as 1, so 1, 1, 1.1, 1.2 and this 2 I bring down to the next values that is 1.2 and then take the average; this is known as Winsorized mean. So, these trimmed mean and the Winsorized means; they neutralize the effect of the extreme values and many times in practice they are used especially when we are judging the people based on the averages etcetera because when there is a several judges are there who are judge evaluating a person, then there may be some bias in the form of certain prejudices and to discard that thing; we will avoid using the extreme values and so either we can totally discard them or we normalize them.

So, in place of 1 we may take 2 or 3 depending upon the situation because suppose there are in place of 10 there are 100 judges and in that case there may be a larger number which may have to be trimmed or beings raised. Some other quite useful measures of location are median, now median by the name it means it is the middle value of the observations for example, if I have written the sampler as 1, 1.1 and 1.2, so this is the middle value; in play suppose I have a even number then the middle two values take the average. So, what we do we can order the values x 1 less than or equal to x 2 less than or equal to say x n minus 1 less than or equal to x n. So, any sample of n values you order them and look at the middle; that means, if it is an even number then it will be x n plus 1 by 2 value or it will be x n by 2 plus x n by 2 plus 1 value divided by 2.

So, if n is odd and this is if n is even, so when we have a raw data; we can order them in a sequence and find out the median using this statement; however, if I have a frequency distribution then one may consider the following.

(Refer Slide Time: 20:41)



It is written as M i is equal to x l plus n by 2 minus n l divided by f naught into c. So, what we do? We firstly consider the total frequency that is n and we consider up to which class it is coming; that is n by 2 where it is coming. So, we look at the cumulative frequency table, so let me explain it through some example. So, we look at one example here the frequency distribution is given in this form 144.55 to 149.55, 149.55, to 154.55, 154.55 to 159.55, 159.55 to 164.55, 164.55 to 169.55, 169.55 to 174.55, 174.55 to

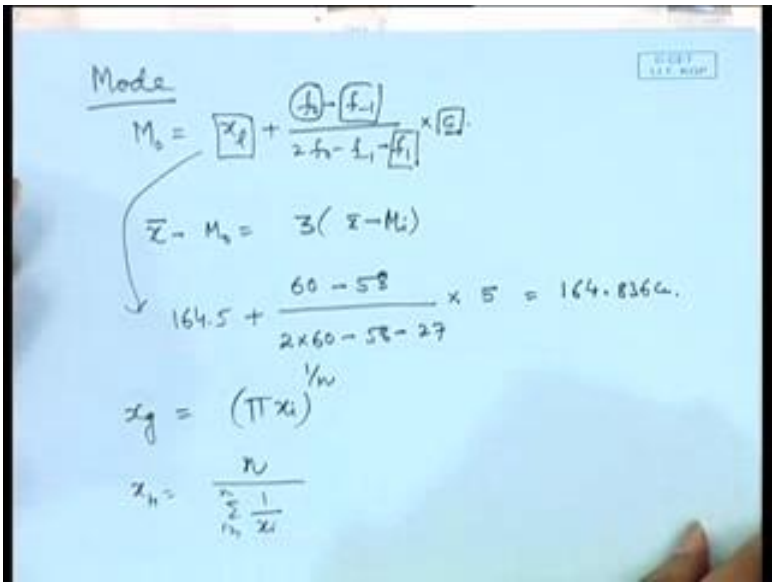
179.55, 179.55 to 184.55. The corresponding frequencies are 1, 3, 24, 58, 60, 27, 2 and 2 the total frequencies is 177. So, what we do we calculate the cumulative frequencies that is 1, 4, 28, 86, 146, 173, 175, 177. So, we look at what is 177 by 2 that is n by 2 that is 88.5. So, the value for which 86 is here, so 88.5 is above this, so this is called the median class.

Now, x l value is the lower value or the lower limit of the median class, so in this formula median will be 164.55 plus n by 2 that is 88.5 minus n l; n l is the cumulative frequency which is occurring just before that. So, that is 86 divided by f naught; f naught is the frequency of the median class that is 60 multiplied by c; c is the length of the class interval. So, you can see here the median value for a frequency classified data is given by x l plus n by 2 minus n l divided by f naught into c; how this is determined is by. Firstly, looking at what is n by 2; n is the total frequency find out what is n by 2. In the cumulative frequency table, you observe that this n by 2 where it will occur, so since 86 is coming here 88.5 is above this.

So; that means, that particular frequencies occurring in the next class that is from on 64.55 to 169.55, so this is known as the median class. Now when we classify the median class then x l denotes the lower limit of that class, n l denotes the cumulative frequency just before that class, f naught is the actual frequency of that class and c is the length of the class interval, so once we substitute this values; this value turns out to be 164.758 centimeter. Naturally this means this is an approximate value of the median, but median will lie in this class. So if I have the raw data, I can find out the middle value exactly, but when I have a frequency classified data, I can look at in this particular section.

In fact, if I consider a relative frequency distribution then m value will be coming somewhere in the middle, so that is the actually physical interpretation of the median.

(Refer Slide Time: 25:42)



Another measure of central tendency is mode; mode is that value which occurs most frequently in a given distribution. Therefore, when the raw data is given we can actually see where which value is occurring most frequently and if we are given a histogram, we can easily point out or if I have a frequency polygon then I can easily make out where is the mode. For example, here this is the mode, this is the mode or this is the mode or in these curve; this is the mode, this is the mode, this is the mode, this is the by model distribution etcetera. So, one can find out from the shape of the distribution where is the mode likely to lie.

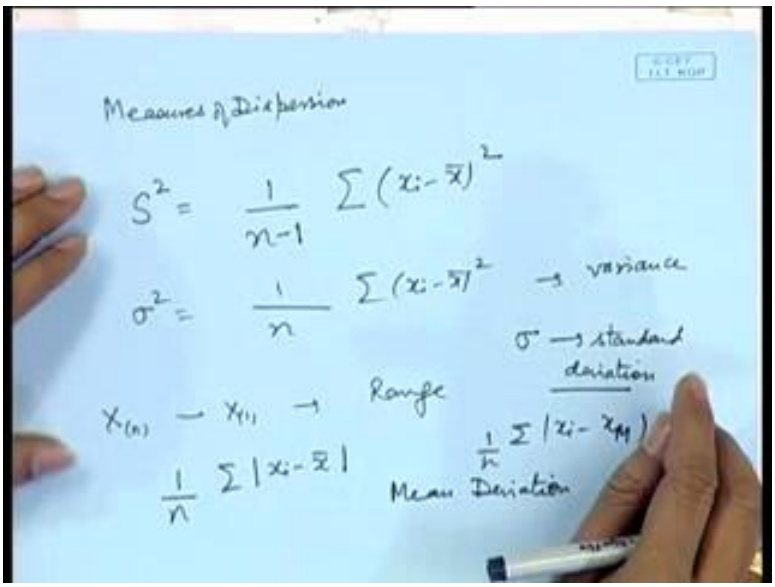
However, once again if I have a frequency classification data in the form of class intervals, in that case we need a particular formula that formula is given by that mode is equal to x l plus f naught minus f of minus 1 divided by twice f naught minus f of minus 1; minus f of 1 into c. So, what we do; we consider the class interval of the maximum frequency and f naught is the frequency of that class that is called the model class, f minus 1 is the frequency of the class prior to that and f 1 is the frequency after the class; x l is the lower limit of that class and c is the length of the class interval. So, this formula is used for calculating the mode when we have frequency distribution.

There is some relation between mean median and mode, so many times this is use like x bar minus m naught is equal to 3 times x bar minus M i; where M i denotes the median, m naught denotes the mode, x bar denotes the arithmetic mean, so many times this

formula is also used. If we consider the frequency classified data that is given in this particular table then you can easily see that the model class is 164.55 to 169.5 because this is having the highest frequency. So, for this particular thing x l is 164.5 plus 60 is the frequency of the model class, 58 is the frequency just before that then twice f naught minus f minus 1 minus f 1 that is the frequency of the next class in to the length of the class interval.

So, after simplification it turns out to be 164.836 centimeter one drawback of the mode is that relatively one may not be able to use that and it is, I mean it is not that convenient to calculate; that could be the thing. Some other measures of central tendency are like harmonic mean and by geometric mean, so in the raw form the geometric mean is defined as the product of all the values to the power 1 by n. Similarly a harmonic mean is defined as 1 by sigma; that is n by 1 by x i; i is equal to 1 to m that is the reciprocal of the arithmetic means of the reciprocals that is called the arithmetic mean.

(Refer Slide Time: 29:22)



Now apart from the measures of central tendency, we have measures of variation nor measures of dispersion or measures of variability. So, one of the popular measures of dispersion is the variance; we consider 1 by n sigma x i minus x bar whole square. Now there are different notations for these, we may use 1 by n minus 1 here sigma is square as 1 by n, sigma x i minus x bar whole square; this is known as variance and this one is sometimes called the sample variances. Another measure is that if I consider the

difference between the largest value and the smallest value that is called the range; that is the maximum minus the minimum.

We may also look at mean deviation that is 1 by n sigma x i minus x bar or 1 by n sigma x i minus form x m where x m is the median, so these are called mean deviations. All of this give information about the variability in the data, when we take sigma that is by square root of the variance this is known as the standard deviation. Apart from that we have measures of skewness and kurtosis, so that I will be telling in the next class.