Probability and Statistics

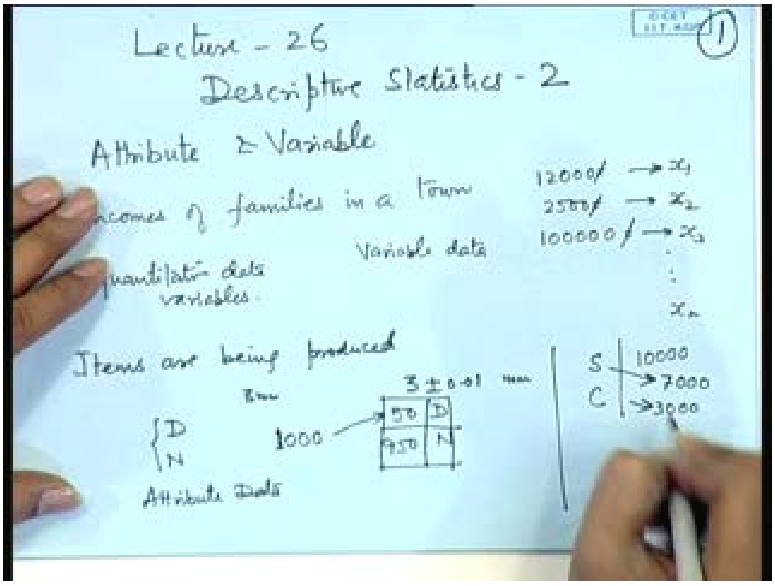
Prof. Dr. Somesh Kumar
Department of Mathematics
Indian Institute of Technology, Kharagpur

Module No. #01
Lecture No. #26
Descriptive Statistics II

In the previous lecture, we have discussed the variety of data that is available in the Statistical Sciences. And, we saw that we can represent it graphically to make it appropriate for analysis or for understanding the kind of data that we are having, what it represents, etcetera. However, only graphical or tabular representation may not yield too much. One may have to present it in a very systematic way, so that one can draw various characteristics of that population or the data.

So, we will look at how to consider the frequency distribution of the data. The first thing that we observe is that, in what way the data may arise. So, the data may arise in one of the two possible ways. One way could be that the to start with the data is numerical.

(Refer Slide Time: 01:26)



For example, if we are writing down say, incomes of families in a town, so in that case, corresponding to each family you are writing down certain numerical value. Suppose, we are recording the incomes per month, so the incomes could be say 12000 rupees per month, it could be 2500 rupees per month or it could be 100000 rupees per month etcetera.

So, here the primary recording of the data itself is in the numerical form. This data is said to be variable data or frequency variable data. So, we can give a nomenclature here like 12000 we can write as x 1 value, 21500 we can write as x 2 value, 100000 as you can write as x 3 value and so on. So, the data is recorded in the form of variables x 1, x 2, x 3, x n. These are called quantitative variables. Because corresponding to each variable, we are having a numerical value associated with this. So, this is termed as quantitative data or quantitative variables. But, there are certain other places where the data is not recorded in the beginning as a numerical data.

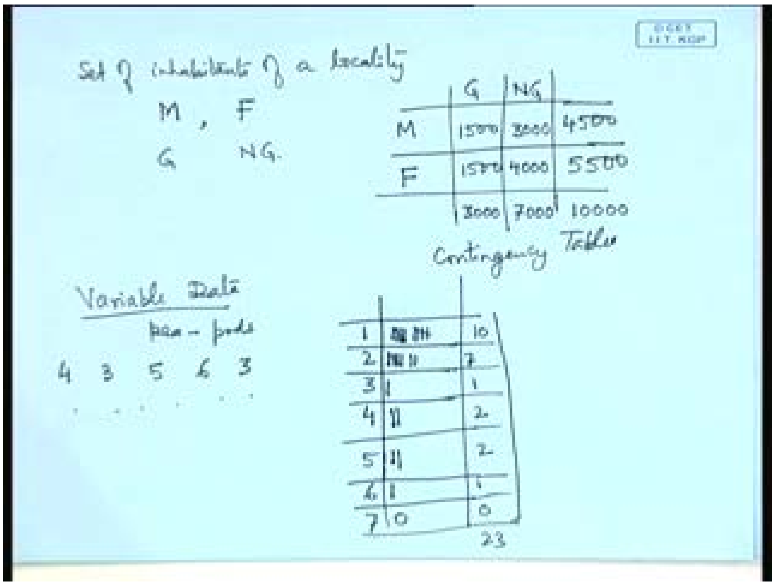
For example, if a manufacturing process is on and certain items are being produced. A quality control instructor, he measures certain characteristics of that product. For example, if these are, say golds produced by a manufacturing process, so they measure the diameters of the golds, and if the golds are having diameter less than a certain value or more than a certain value, then they are considered to be defective. And, if the diameters are within a certain range, we call them to be alright. For example, if the average diameter is say 3 millimeter and we may specify the limits as 3 plus minus 0.01millimeter. So, if the say, if a gold produced has a diameter between 2.99 millimeters to 3.01millimeters. It is considered to be alright. Otherwise, it is said to be defective. However, the instructor reports an item to be defective or non-defective. So, corresponding to each item, the value recorded is either D or N.

Now, if we record for thousand golds D or N, then out of this, may be 50 are defective and 950 are non defective. So, at a second stage the data becomes quantitative. This is known as attribute data because here the initial recording is in terms of the character that is or you can say the type, for example, if we are looking at a population of students and we may then, classify them as in the school students or a college students.

So, each student may be a school student or a college student. So, if we have say 10,000 students, then out of this 10,000 may be you have 7000 as school student and 3000 as a college student. So, whenever we classify the data according to certain property, then it is said to be attribute data. It is the technique for analysis of attribute data. And, the quantitative data are somewhat different.

In this particular section, we will explain how to make frequency distributions for the quantitative data. In the case of attribute data, it is relatively easy to make a frequency distribution. For example, one may have two characteristics according to which we classify the data.

(Refer Slide Time: 06:04)



Set of inhabitants of a locality

M , F

G NG.

	G	NG	
M	1500	3000	4500
F	1500	4000	5500
	3000	7000	10000

Contingency Table

Variable Data

[illegible] - [illegible]

4 3 5 6 3

. . . . . .

1	卌 卌	10
2	卌 ||	7
3	|	1
4	||	2
5	||	2
6	|	1
7	0	0
		23

Suppose, we look at a set of say inhabitants of a locality, now when we do the survey we may record their gender; that means, male or female. We may also record whether they are graduates or non-graduates; that means, below graduates. So, now, suppose we have a population of 10,000 in the locality, we may classify firstly according to male, female.

So, maybe we have 4,500 male and 5,500 females. Now among this, we look at how many graduates and non graduates. So, suppose the number of graduates is 3000 and there is 7000 non-graduates are there, now out of the 3000 graduates, how many of them are males and how many of them are females? Suppose it is 1,500 males, 1,500 females. Out of 7000 non-graduates, how many of the males? So, 3000 are males and 4000 are females.

So, this is known as contingency table. So, an attribute data can be represented in the form of a contingency table. Here, we are classifying according to two characteristics. One may classify according to 3, 4 etcetera. And, accordingly the representations can be made in the form of contingency tables.

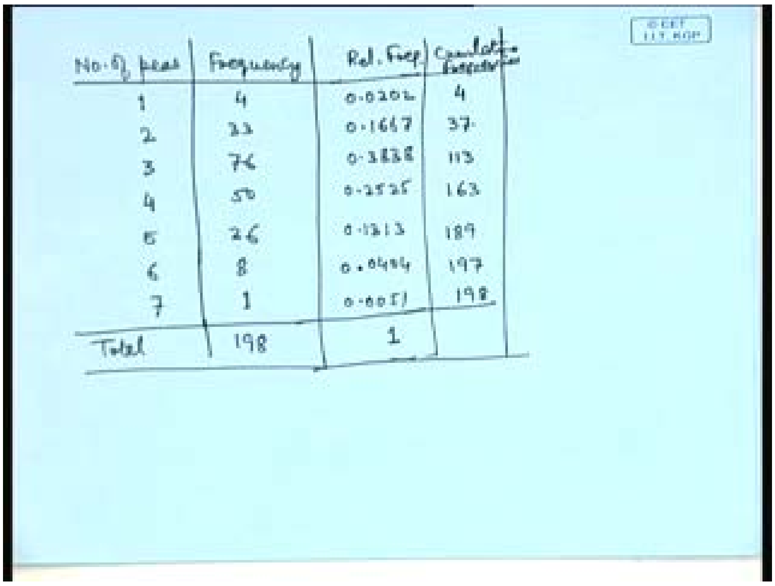
Now, we consider variable data. Suppose, the data is recorded on the number of peas in certain pea-parts pea pea parts, so in each pea-part how many pea peas will be there? So, there may be no peas. There may be 1 pea, 2 peas, and so on. So, finally we classify, if it is 0 we neglect it. Out of 198, how many are having 1? How many are having 2? How many are having 3? How many are having 4? How many are having 5? How many are having 6? And, how many are having 7 peas? So, the primary data would be in the form of certain numbers like, we take one pea-part and see how many peas are there, say 3.

Then, we take another one. How many are there say 3? Next, how many are there suppose we say 5? Next, how many are there suppose we say 6? Next, how many are there suppose we again have 3? So, we record the data. Now from here, we make the frequency distribution. Now, there are certain elementary techniques for creating the frequency distribution. That is called counting techniques.

So, if one occurs once, we may say tally mark; if 2 occurs once, we say make a tally mark; suppose 4 occurs, we make a tally mark; suppose next one is again observed, we make a tally mark; then, suppose next 6 is observed, we make a tally mark; suppose, then next 5 is observed, we make a tally mark; suppose again one is observed, we make a tally mark; next suppose 4 is observed, we make a tally mark; next suppose one is made, suppose we have 5, suppose we have 3. Now, suppose 1 is again observed, then these are already 4 symbols. We cross it. This is showing a block of five frequencies.

So, in this elementary way, one can make a distribution and see how many frequencies have occurred for each of them. So, suppose if I have observed only this much, we can consider the frequency distribution as corresponding to 1, it is 10; corresponding 2, it is 7; corresponding to 3, it is 1; corresponding to 4, it is 2; corresponding to 5, it is 2; corresponding to 6, it is 1; corresponding to 7, it is 0. So, if I have taken only these many peas pea-parts, so this is 23. Out of 23 pea-parts, the frequency distribution of the number of the peas is given by this. Let me take a rather detailed example, where 198 pea parts were recorded corresponding to their number of peas per pea-part and the frequency distribution turns out to be…

(Refer Slide Time: 11:25)



No. of peas	Frequency	Rel. Freq	Cumulative frequency
1	4	0.0202	4
2	33	0.1667	37
3	76	0.3838	113
4	50	0.2525	163
5	26	0.1313	189
6	8	0.0404	197
7	1	0.0051	198
Total	198	1	

So, we write here number of peas and here, we write frequency; that is, how many times each occurrence is there. So, 1 is occurring 4 times, 2 is occurring 33 times, 3 is occurring say 76 times, 4 is occurring 50 times, 5 is occurring 26 times, 6 is occurring 8 times and 7 is occurring once. So, this total sum is 198. This is called basic frequency distribution.

Now, one may present this data in various forms. One may look at relative frequencies. Here. The relative frequency means, we mean the percentages or the proportion of each number in the total. So, for example, how much is 4 of 198? It is nearly 0.0202, how much is 33? It is point 0.1667, 76 is 0.3838, 50 is 0.2525, 26 is 0.1313, 8 is 0.0404 and 1 is 0.0051. The sum is 1. If we look at the last column, it tells the relative occurrence of each number in the sample.

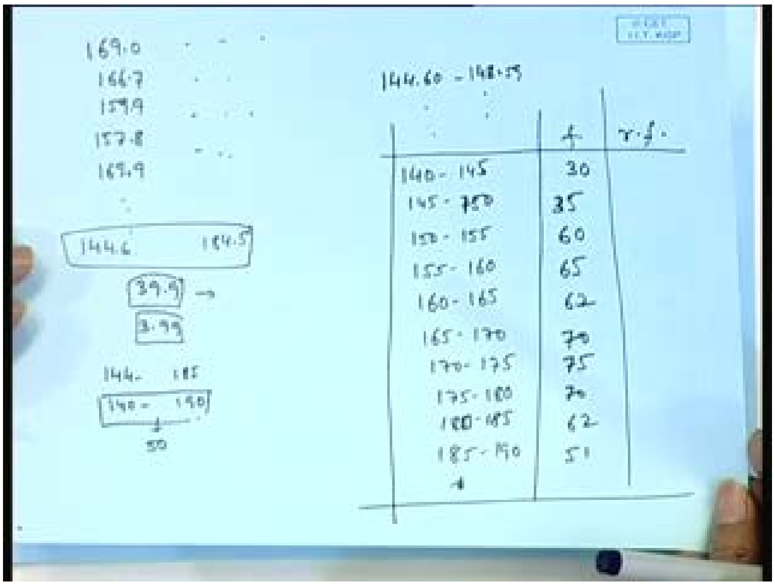
For example, 3 occurs nearly 38 percentage times, say 2 occurs nearly 16.6 percentage of the times, 4 occurs nearly 25 percentage of the time, whereas 1or 7 they occur very less. So, 1occurs only 2 percentage and 7 occurs only 0.5 percentage of the times. So, this information is also useful. Sometimes we look at cumulative frequencies Cumulative frequencies. When the data is arranged in a tabular form and we have the frequencies, we may add them successively.

So, for example, here it is 4, then 37, then 113, then 163, then 189, 197, 198. If we look at this cumulative frequency, it tells how the frequencies are adding up. So, this also gives lot of useful information. This type of a representation is called a discrete table or discrete variable table. Because here corresponding to each value, certain values are given. However, sometimes the data is too numerous.

For example, if we are recording the heights of students or incomes of the families, then the data is too numerous; in the sense that if we record height, we are not only going to record it in terms of inches, it may be in the terms of point, the fractions of inches or centimeters and then the fractions of the centimeters and the fractions they are of. So, the total numbers of values that we are considering are too many. This is called continuous data and in that case, it is more appropriate to split the data into intervals.

For example, we may consider how many persons are having height from 140 centimeter to 150 centimeter, how many of them are having height from 150 centimeter to 160 centimeter, how many of them are falling it into 160, 170 centimeter, etcetera. So, this type of classified data, once again if we have the raw data, so raw data will be of the form say…, I will just give few values here. The heights are recorded in centimeters of say in Indian nails.

(Refer Slide Time: 15:56)



169.0
166.7
159.9
157.8
169.9

$\boxed{144.6 \quad 184.5}$

$\boxed{39.9}$ →

$\boxed{3.99}$

144- 185

$\boxed{140 - 190}$

50

144.60 - 148.59

	f	r.f.
140- 145	30	
145 - 150	35	
150 - 155	60	
155- 160	65	
160- 165	62	
165 - 170	70	
170- 175	75	
175- 180	70	
180- 185	62	
185 - 190	51	

So, heights may be in the form 169 centimeter, 166.7 centimeters, 159.9 centimeters, 157.8 centimeters, and 169.9 centimeters and so on. So, now from here, we make classes. Now, we may make classes of various kinds. We may make from, say we may record what the lowest value is and what the highest value is. Suppose, the lowest value is say 144.6 and the highest value is say 184.5, then we look at this difference. The difference is 39.9. Now, this is not a very convenient number. However, if we are very strict we can look at only this difference. And, divide it into say six classes or five classes. Now, if I divide it into say ten classes, then each class will be of the length 3.99 centimeters; that means, my class intervals will be like 144.60 to 148.59 and so on. Now, this looks slightly inconvenient from mathematical point of view.

So, a better option could be that I can consider the length form 144 to 185 or even better we can consider something like 140 to 190, which is a more convenient representation.
So, this is the total length is 50 centimeter. And, if we divide into intervals of length say ten intervals, then the interval length will be like 140-145, 145-150, 150-155, 155-160, 160-165, 165-170, 170-175, 175-180, 185-1 180-185 , 185-190 and that is all. And, in the next column, we will put the number of people occurring in each. So, the numbers may be like say 30, say 35, 60, 65, may be say 62, 70-75, may be again 70, 62, 51, etcetera.

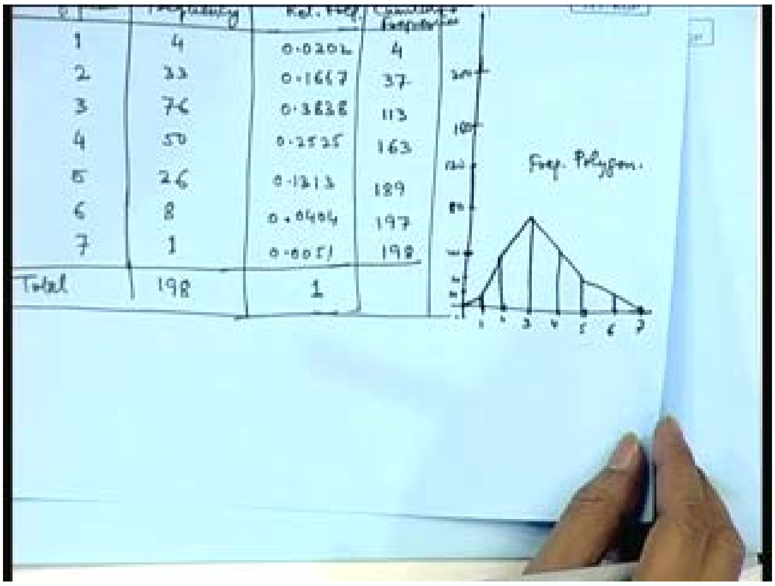
Again, here we may consider frequency as well as relative frequency. Now, when we present these class intervals like this, the rules for the counting have to be made for example, 145 is here, 145 is here.

So, we have to make a rule that, this will be considered less than or equal to. That means, if a person is having height equal to 145, then he will be considered in the preceding class or we may make a rule of the reverse side that is, if it is for 145 it will be considered in this class. But, we have made a rule prior to preparing the table. The next thing is the number of the class intervals. How many class intervals are sufficient to represent a given data into a frequency distribution? So, the general principle is that the number of intervals should not be too small; it should not be too large also.

If you keep too small, then lot of information is hidden. For example, if I made the intervals only say two intervals, say 145 to say 165 and 165 to 190, and then we are not able to distinguish the persons who are say between 140 to 150 and between 150 to 160 etcetera. Because everything we are putting into the same. At the same time, if we are making too many intervals; suppose in place of 10, I make here 25 intervals of length 2 each. So, 140-142, 142-144, and then it will be difficult to analyze. It is same thing like in the discrete case that if I have too many values here, then it is not easy to handle that value from a statistical methodology point of view. There are some guidelines for making the class intervals and the number of intervals. And, one should follow them before making the thing.

Now, the next thing is the graphical representation of frequency distribution. So, if we have a frequency distribution of this nature, then we may simply make a bar diagram in the form.

(Refer Slide Time: 21:26)



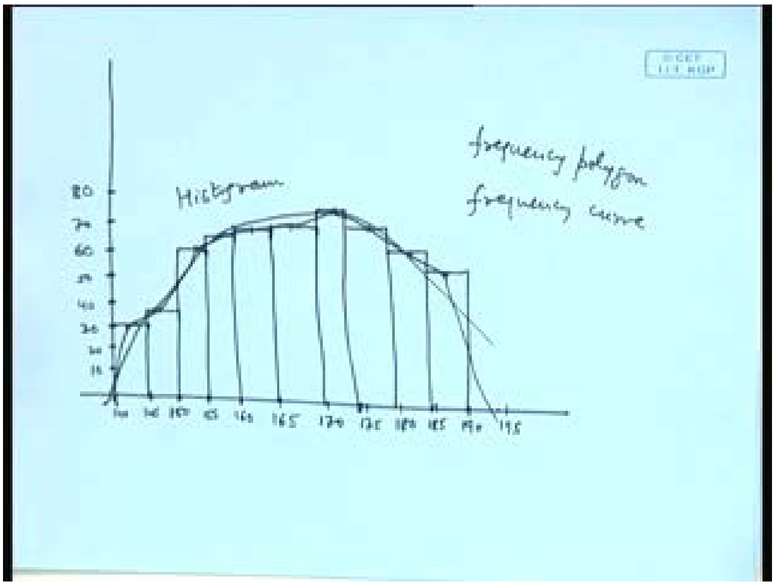
[illegible]	Frequency	Rel. Freq.	Cumulative Frequency
1	4	0·0202	4
2	33	0·1667	37
3	76	0·3838	113
4	50	0·2525	163
5	26	0·1313	189
6	8	0·0404	197
7	1	0·0051	198
Total	198	1	

So, on the x axis I plot this numbers, so 1, say 1, 2, 3, 4, 5, 6 and 7. And, we make a bar of the height, which is the scale into the frequency is given here.

So, now we have to see the relative position. So, if I have this as say 5, then 10, so now, you see that I need a large gap here. So, this is 20, then this is 40, then this is say 60, then this is 80. And, so the table has to be very large. So, in place of this, I can make this as say 10, then this is 20 and then this is say 40, and then this is say 80, and this is 120, this is 160 and say, this is 200.

Now, if you see the numbers that we are going to present here, 4 that will come somewhere here, then 33 will be coming up to this height, then 76 will be coming up to this height, 50 is coming up to say this height, 5 the 26 is coming up to this height, 8 is coming up to this height and 1 is this one. So, this tells the relative importance of or you can say relative occurrence of each term in the frequency distribution. So, bar diagrams are quite useful. One may join these values. This tells us about the shape of the curve and this is known as frequency polygon. However, if one has a classified data of the continuous variable, it is more appropriate to draw a histogram.

(Refer Slide Time: 23:42)

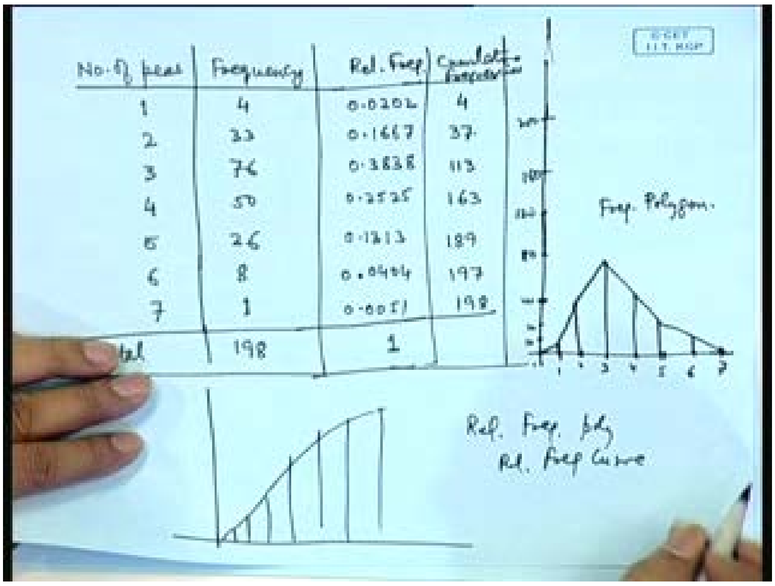


So, for this data one may consider. So, we have say 140 to 145, 150, 155, 160, 165, 170, 175, 180, 185, 190,195, say. So, here we have to make the scale first. We look at the values, which are there 30, 35, 60, and so on. So, if we make the scale like 10, 20, 30, 40, 50, 60, 70, 70, 80. So, the maximum value is 80. So, we may draw like this. 10, 20, 30, 40, 50, 60, 70 and 80.

Now, we observe the frequency for the class interval 140 to 145 is 30. So, we draw a rectangle of the height 30 in the interval 140 to 145, then for 145 to 150 the frequency is 35, so we add the height up to 35 here. And, for the interval 145 to 150, we consider the histogram of this length. Similarly, the next is 60 from the interval 160 to 155 to 165 it is 65, then next it is 67, 165 to 170 is 70, 170 to 175 is 75, 175 to 80 it is 70 and 180 to 85 it is 62, 185 to 90 190 it is 51.

This tells that, from 150 to 190 the distribution is almost uniform. The frequencies are almost same. They do not change too much. So, this graphical representation of the frequency distribution in a histogram form gives lot of useful information. For example, we may take the mid points and join them using a straight lines. So, this is again a frequency curve for the continuous case. We may also join them using a free hand curve. Then, this is known as frequency curve. So, you have frequency polygon, then we join the points using straight lines. If we join using free hand, then it is frequency curve. And, this is known as histogram.

(Refer Slide Time: 27:29)

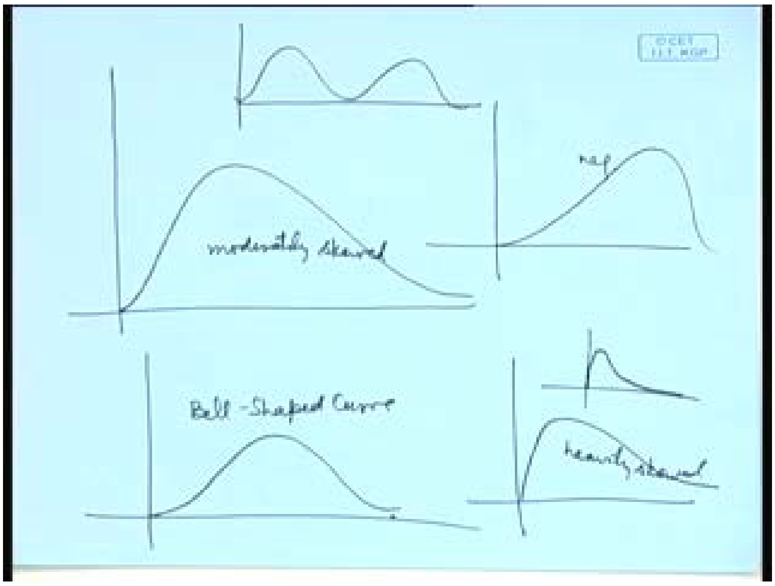


No. of peas	Frequency	Rel. Freq	Cumulative Frequency
1	4	0.0202	4
2	33	0.1667	37
3	76	0.3838	113
4	50	0.2525	163
5	26	0.1313	189
6	8	0.0404	197
7	1	0.0051	198
tal	198	1	

Freq. Polygon.
1
2
3
4
5
6
7
Rel. Freq. poly
Rel. Freq Curve

We also have a so-called cumulative frequency curve. Because once we have made the cumulative frequencies we may join, we may plot the cumulative frequencies here in place of the ordinary frequencies. So, naturally this will be increasing. Take down the values here and if you join that, this is known as relative frequency polygon or relative frequency curve, depending upon how we join them. So, the slow part you can say the speed by which it is increasing tells the relative importance of the each values of the variable; because that will tell, how much frequency is allotted to that one. Now, when we make a histogram, then the shape of the curves tell something about that.

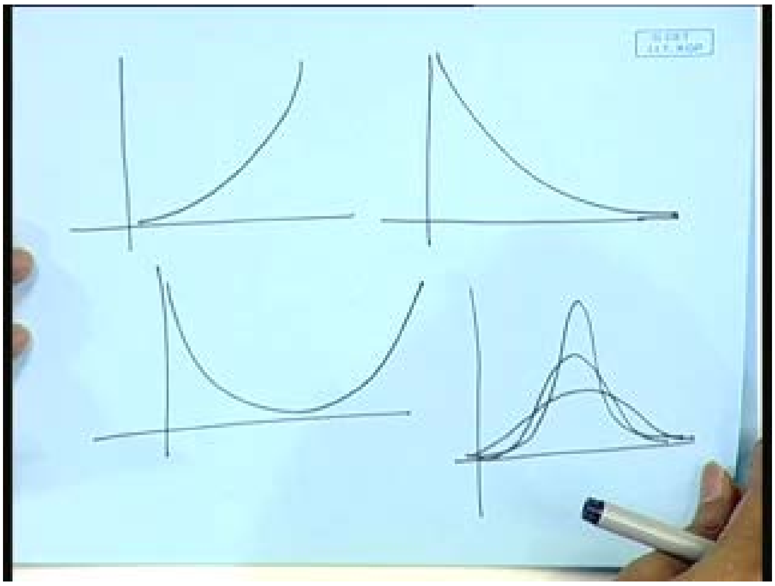
(Refer Slide Time: 28:29)



So, for example, if we draw a histogram and the curve is coming something like, say this or a curve 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 frequencies are also increase. But, after certain 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 quite 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 the skew is positive, here the skew is negative. This shows that there are 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, showing somewhat unusual behavior. Such curves are not observed much in practice. However, if it is observed, then one should be careful.

(Refer Slide Time: 30:25)



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 incomes. 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, the people of the with middle level income, the family with the middle level of income are also middle in number that is, moderate in number.

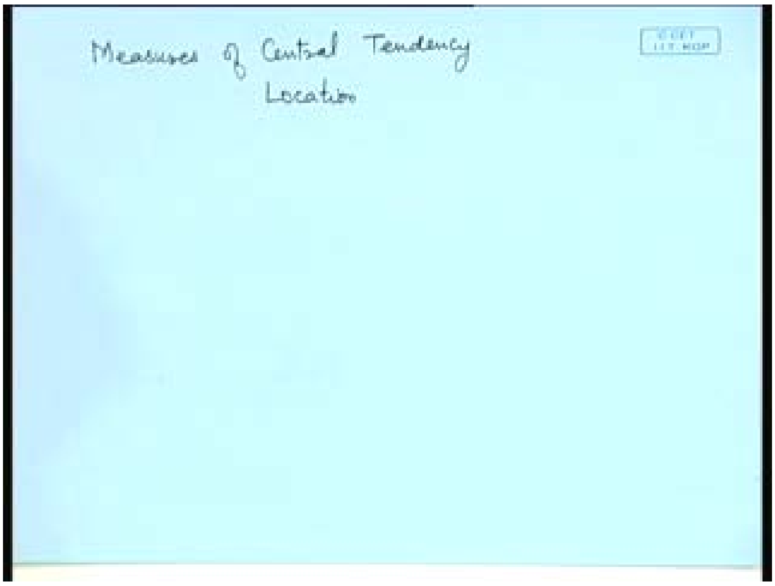
So, this shows this type of curve shows this type of behavior. On the other hand, this type of curve may show some sort of age distribution. For example, the persons with a For example, in a developed country like Australia or say France, there are less number of children. And, the number of people who are quite old because of the average life expectancy is high, so that number is big. 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. 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 becomes higher. But, in some meteorological data, where you are looking at the say cloudiness or the burstiness or say rainfalls, sometimes it may happen that in the beginning there is a more rainfall and in the end there is more rainfall and in the middle there is less. Or, the number of places where less number of less amount of rainfall is here or more amount of rainfall is here, 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 encounter an unusual shape curve. That is a u shaped 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 sparsely distributed. This is moderately, sparsely distributed and this is heavily concentrated distribution. Large number of values is towards the center and very few values towards the end. Here the things are more or less equally distributed. So, you can say a flat belly, a tall belly. So, this type 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 a heavily skewed distribution, then we know the nature of the data. If we know it is symmetric, then we know the nature of the data. If it is skewed, moderately skewed or heavily skewed, we are able to make certain comments about the distribution. For example, if it is increasing or decreasing or it is a u shape etcetera.

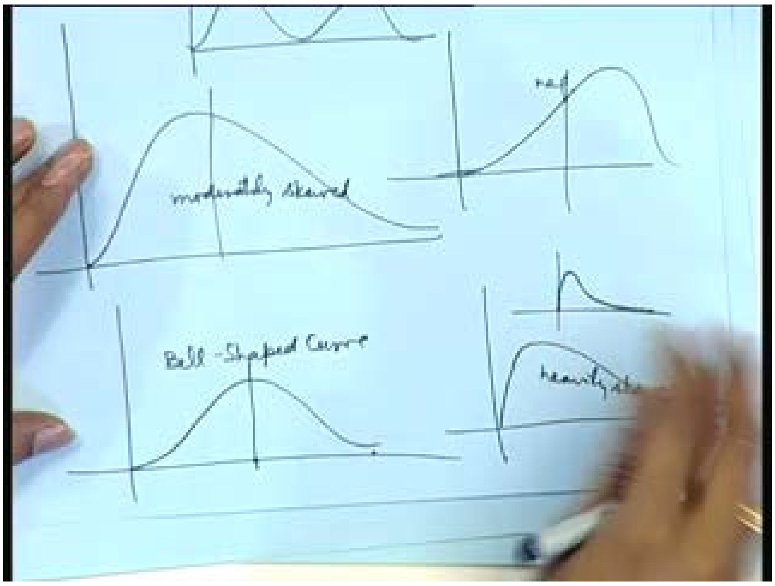
(Refer Slide Time: 34:11)



Measures of Central Tendency
Location

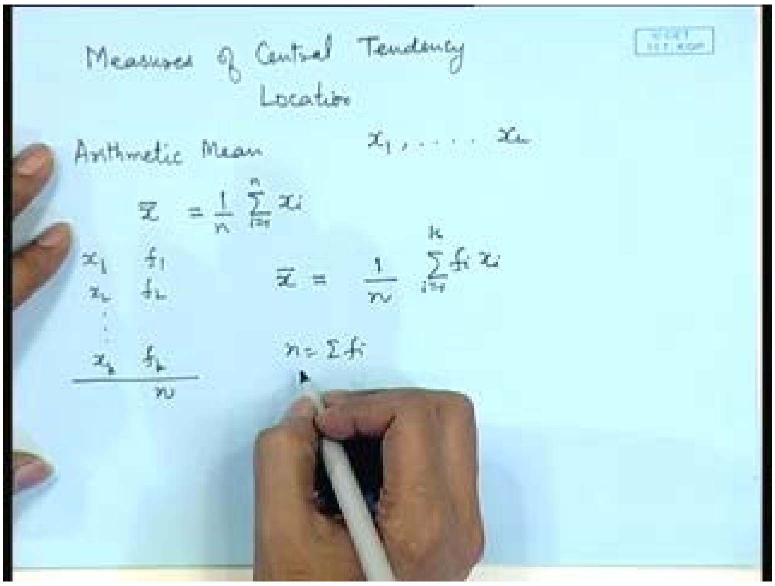
Now, next is measures. That means, from the given frequency distribution we calculate certain measures, which tell them about various characteristics of the distribution. So, the first of these are known as measures of central tendency or you can say measures of location.

(Refer Slide Time: 34:42)



Now, this is something like, even I have this one. You can say that most of the data are concentrated about the center. Here, the center may be little bit shifted, the center may be little bit on this side, here the center may be little bit on this side of the more number of values because of the heavy weighted given to this. But, 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.

(Refer Slide Time: 35:19)



Measures of Central Tendency

Location

Arithmetic Mean $x_1, \ldots \ldots x_n$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

x_1 f_1

x_2 f_2

$\vdots$

x_k f_k

n

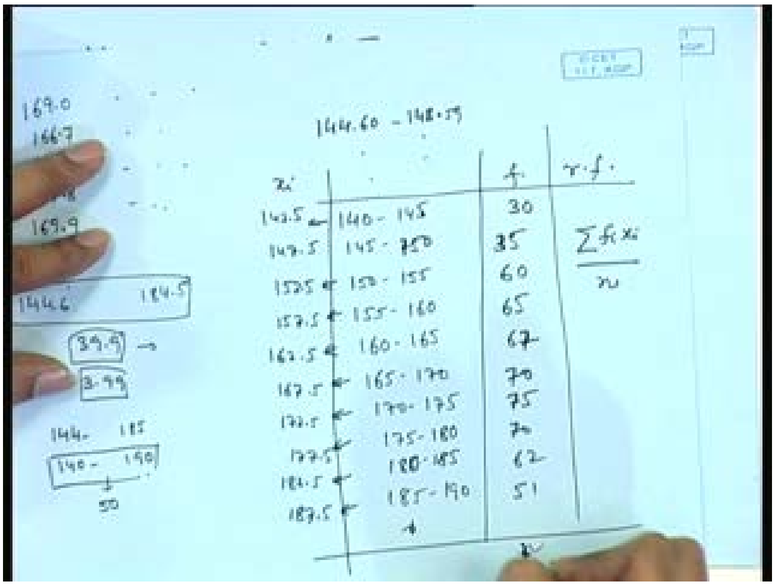
$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} f_i x_i$$

$n = \sum f_i$

So, here we will give example. The first one is an average. So, the easiest of the average is an arithmetic mean. An arithmetic mean is if I have the value say x 1, x 2, x n, is simply the 1 by n sigma x i, i is equal to 1to n. Let us call it x bar; that means, whatever observations are there, we simply take a plain 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 distribution’s 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, you do not have single values of x i. The technique used is that one takes the mid values of the class interval and calls that as x i.

(Refer Slide Time: 36:45)



169.0
166.7
169.9

$\boxed{144.6 \qquad 184.5}$

$\boxed{39.9}$ →

$\boxed{3.99}$

144- 185

$\boxed{140 - 190}$

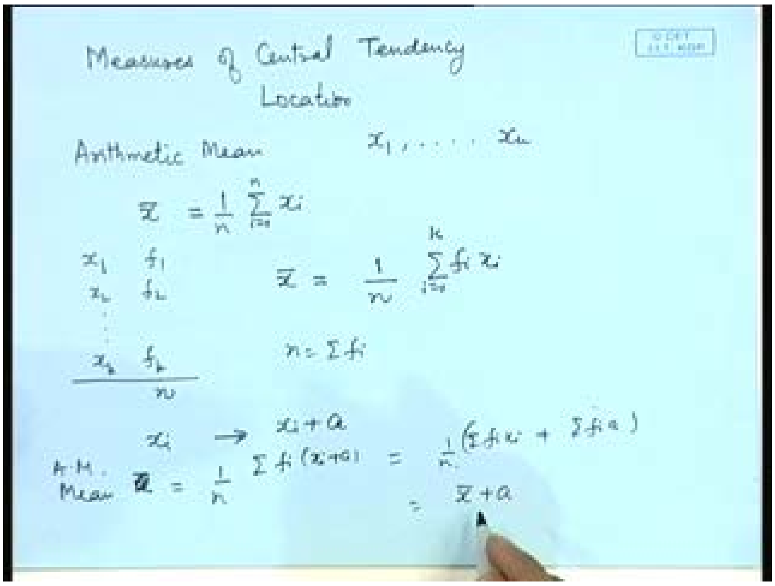
50

144.60 - 148.59

x_i		f	r.f.
142.5 ←	140 - 145	30	$\frac{\sum f_i x_i}{n}$
147.5	145 - 150	35	
152.5 ←	150 - 155	60	
157.5 ←	155 - 160	65	
162.5 ←	160 - 165	67	
167.5 ←	165 - 170	70	
172.5 ←	170 - 175	75	
177.5 ←	175 - 180	70	
181.5 ←	180 - 185	62	
187.5 ←	185 - 190	51	
	+	n	

So, in the case of this particular example, you may take x i as say 142.5, 147.5, 152.5, 157.5, 162.5, 167.5, 172.5, 177.5, 182.5 and 18 7.5. 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.

(Refer Slide Time: 37:51)



Measures of Central Tendency

Location

Arithmetic Mean $\quad x_1, \ldots, x_n$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

x_1	f_1
x_2	f_2
$\vdots$	
x_k	f_k
	n

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} f_i x_i$$

$$n = \sum f_i$$

$$x_i \rightarrow x_i + a$$

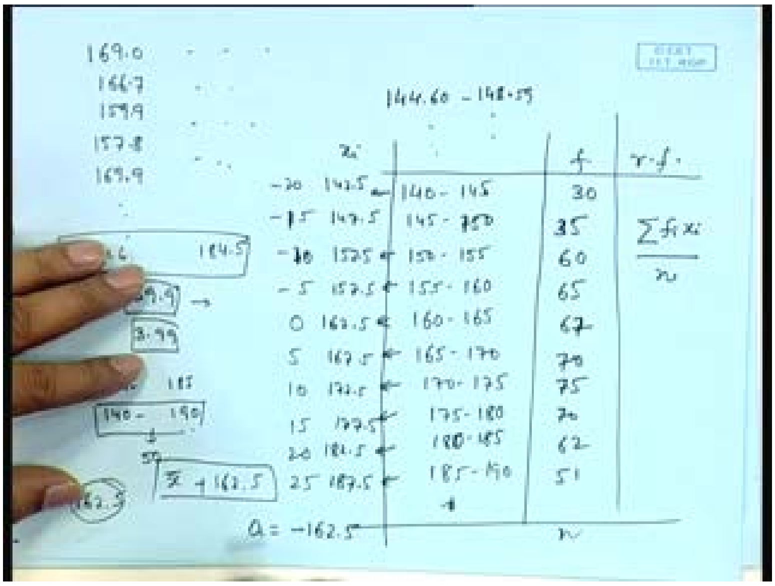
A.M. Mean $\bar{x} = \frac{1}{n}\sum f_i (x_i + a) = \frac{1}{n}\left(\sum f_i x_i + \sum f_i a\right)$

$$= \bar{x} + a$$

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, x bar will become, let me look at this formula 1 by n sigma f i x i plus a; that means, the mean of this, which is also called a m. ok.

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 value. 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.

(Refer Slide Time: 39:02)



169.0
166.7
159.9
157.8
169.9

184.5

3.99

185

140 - 190

$\bar{x} + 162.5$

162.5

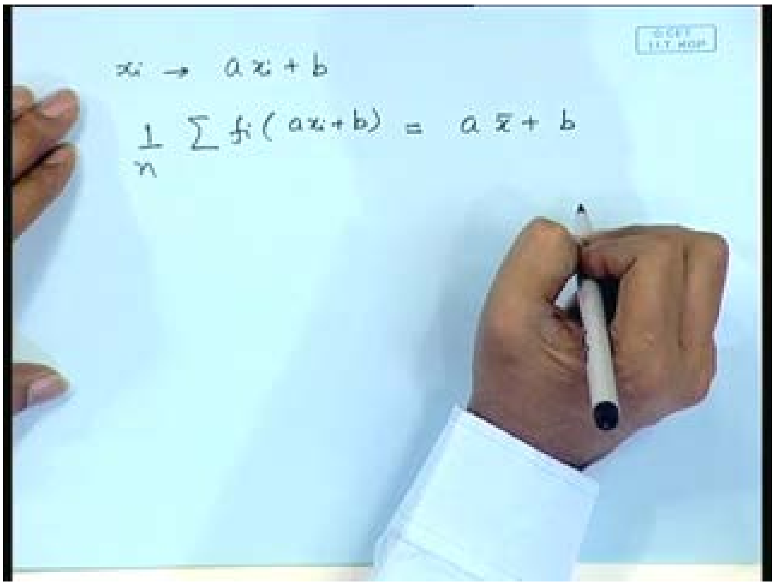
144.60 - 148.59

	x_i		f	x·f
-20	142.5	140 - 145	30	$\frac{\sum f_i x_i}{n}$
-15	147.5	145 - 150	35	
-10	152.5	150 - 155	60	
-5	157.5	155 - 160	65	
0	162.5	160 - 165	67	
5	167.5	165 - 170	70	
10	172.5	170 - 175	75	
15	177.5	175 - 180	70	
20	182.5	180 - 185	62	
25	187.5	185 - 190	51	
	a = -162.5		n	

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. So, and moreover now if you can see here I have shifted a as, by taking minus 162.5.

So, whatever average of this would have come, actually it would come x bar minus 162.5. Now using this, the calculations are much simpler because these numbers are quite nice round numbers, multiples of 5, so that I can multiply and add another thing. What has happened? 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. Sometimes you may scale also.

(Refer Slide Time: 40:20)

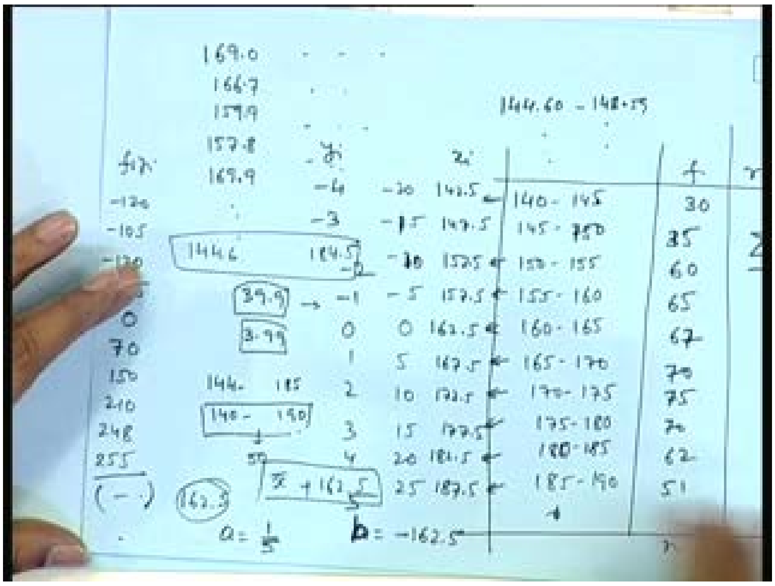


$$x_i \to a x_i + b$$

$$\frac{1}{n}\sum f_i(ax_i+b) = a\bar{x} + b$$

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 one if we expand, then I get sigma x bar plus b.

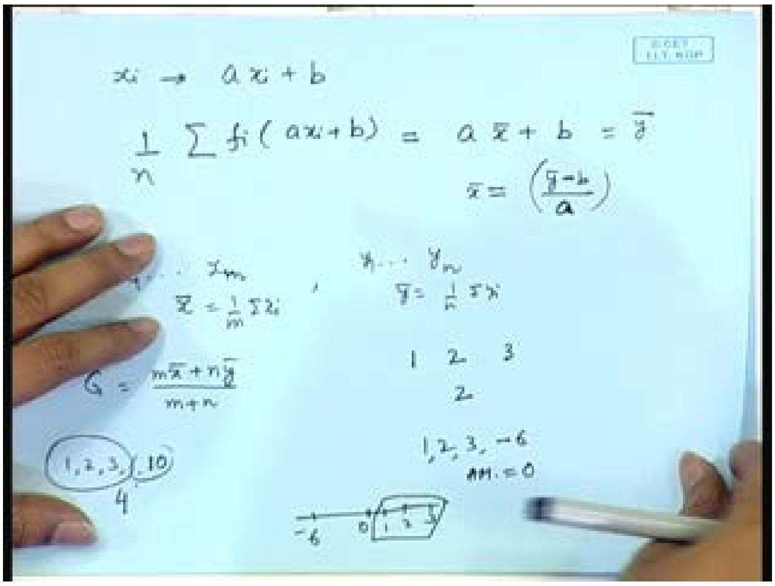
(Refer Slide Time: 40:52)



Now, in the example of the heights, suppose I shift, I take that b is equal to minus 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 a table f i y i, then you see here the calculations minus 120. So, the numbers can be handled very quickly; 10, 5. Then we have minus 120, we have minus 65, we have 0, we have 70, we have 150, we have 210, and we have 62 into 4 that is 248 and then 51 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 get adjusted and this total will become much smaller number.

(Refer Slide Time: 42:11)


$x_i \to a x_i + b$
$\frac{1}{n} \sum f_i (a x_i + b) = a\bar{x} + b = \bar{y}$
$\bar{x} = \left(\frac{\bar{y}-b}{a}\right)$
x_m
$\bar{x} = \frac{1}{m} \sum x_i$
$y_1 \dots y_n$
$\bar{y} = \frac{1}{n} \sum y_i$
$G = \frac{m\bar{x} + n\bar{y}}{m+n}$
1 2 3
2
1, 2, 3, 10
4
1, 2, 3, −6
AM. = 0
−6 0 1 2 3

So, then in the answer what you do is, 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 we divide by a and the arithmetic mean of the original frequency will come.

Some other useful properties that the arithmetic mean satisfies, suppose I have two sequences of numbers, suppose I say x 1, x 2, x n and say another sequence is say y 1,y 2 y. So, let me put here, say m numbers and here I put n number. So, I can calculate the means of these frequencies. This is x bar; that is 1 by m sigma x i. This is y bar, say 1 by n sigma y i.

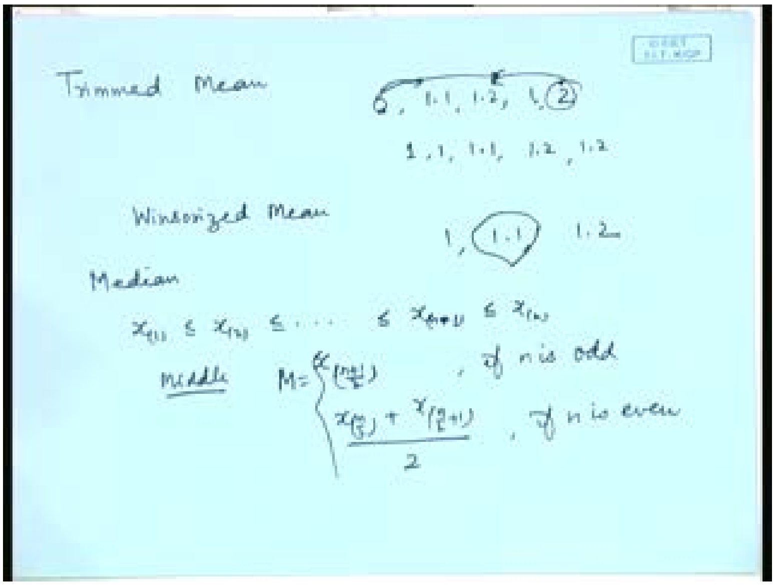
So, if I am looking at the grand mean, I can consider it as m x bar plus n y bar divided by m plus n; that is, the grand mean of both the sequences taken together because m x bar gives the total sum of the first set of values and n y bar gives you the total sum of second set of values. So, if we add and divide by the total number we get. So, this arithmetic mean measure is an enable to various kinds of manipulations. And, that is why it is quite useful.

However, it has one drawback. Suppose, if 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 mean 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 three of the values. So, in some sense, it is affected 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 than these three values and less than this one, much less than this. So, it gives undue attach 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 Olympics, so there are ten judges, who are evaluating at the end of each round. They give certain marks, say out of 10. What the committee will do? They will discard the top value and the lowest value and then they take the average of the remaining. This is known as trimmed mean trimmed mean.

(Refer Slide Time: 45:33)



Trimmed Mean

0, 1.1, 1.2, 1, 2

1, 1, 1.1, 1.2, 1.2

Winsorized Mean

1, 1.1 1.2

Median

$x_{(1)} \le x_{(2)} \le \cdots \le x_{(n-1)} \le x_{(n)}$

$$\underline{\text{middle}} \quad M = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & \text{, if } n \text{ is odd} \\ \dfrac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2} & \text{, if } n \text{ is even} \end{cases}$$

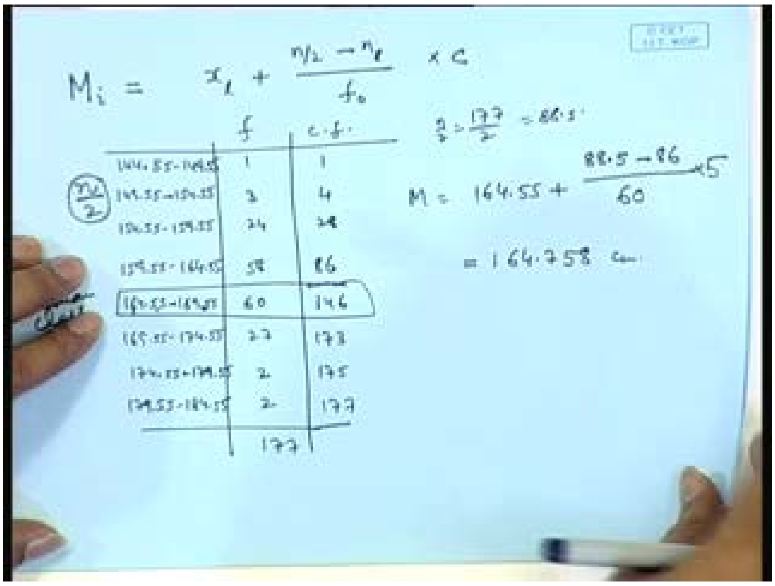
Another option is that the highest and the lowest values are brought to the next lowest or next highest. For example, if the values are say, suppose I write five values, so five values are 1.1 and 1.2 and say 1 and another value is say 2 and one value is say 0. Then, I convert this 0 also to 1.1 and this 0 to 1. So, I will write the lowest value is 1. 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 value 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 this 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 larger number which may have to be trimmed or be winsorized.

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 sample as 1, 1.1 and 1.2, so this is the middle value. If, suppose I have the even numbers, then the middle two values are used to take the average. So, what we do is, we can order the values x 1 less than or equal to x 2, less than or equal to say x n plus 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 two value or it will be x n by 2 plus x n by 2 plus one 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 technique.

(Refer Slide Time: 48:41)



$$M_i = x_l + \frac{n/2 - n_l}{f_0} \times c$$

	f	c.f.
144.55-149.55	1	1
149.55-154.55	3	4
154.55-159.55	24	28
159.55-164.55	58	86
164.55-169.55	60	146
169.55-174.55	27	173
174.55-179.55	2	175
179.55-184.55	2	177
	177	

$\frac{n}{2}$

$$\frac{n}{2} = \frac{177}{2} = 88.5$$

$$M = 164.55 + \frac{88.5 - 86}{60} \times 5$$

$$= 164.758 \text{ cm}$$

However, if I have a frequency distribution, then one may consider the following. 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 this through some example.

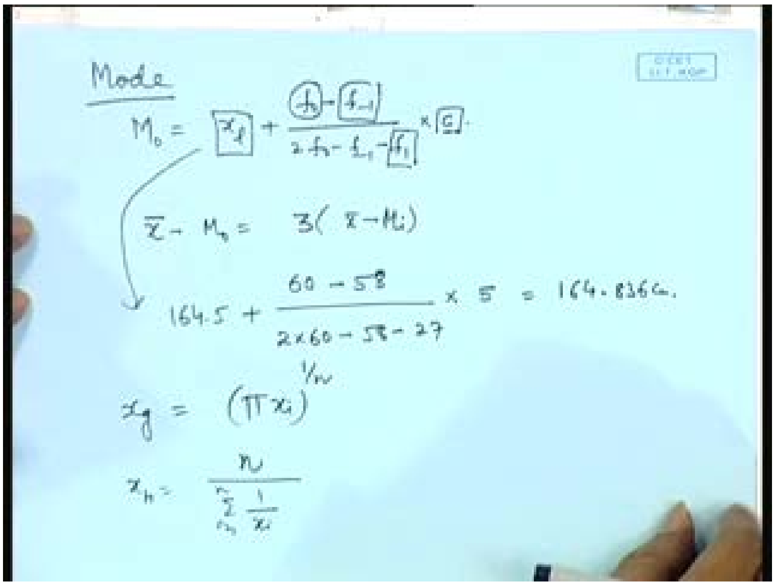
So, we look at one example here. The frequency distribution is given in this form 144.55 to 149.55, 149.55 to 154.55, 154.55 to 159.55, 159.55 to 164.55, 164.55 to 169.55, 169.55 to 174.55, 174.55 to 179.55, 179.55 to 184.55.

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 frequency is occurring in the next class. That is from 164.55 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 value, 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 fashion. In fact, if I consider a frequency 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: 53:43)



Mode

$$M_0 = \boxed{x_l} + \frac{\enclose{circle}{f_0} - \boxed{f_{-1}}}{2f_0 - f_{-1} - \boxed{f_1}} \times \boxed{c}$$

$$\bar{x} - M_0 = 3(\bar{x} - M_i)$$

$$164.5 + \frac{60 - 58}{2 \times 60 - 58 - 27} \times 5 = 164.836$$

$$x_g = (\Pi x_i)^{1/n}$$

$$x_h = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

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 a mode or this is a mode or in this curve this is a mode, this is a mode, this is a mode, this is a bimodal 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 1divided by twice f naught minus f of minus 1minus f of 1into c.

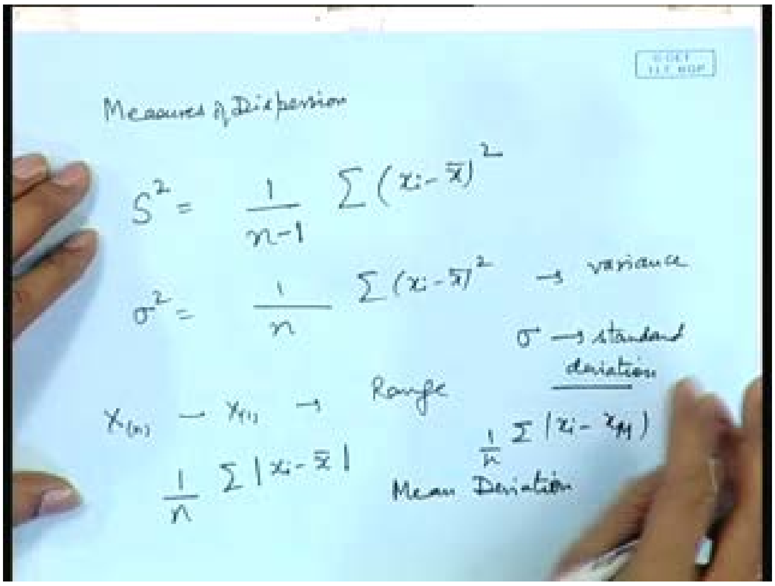
So, what we do is we consider the class interval of the maximum frequency and f naught is the frequency of that class; that is called the modal class, f minus 1 is the frequency of the class prior to that and f 1 is the frequency after the class, x l is the lower limit of that class and c is the length of the class interval. So, this formula is used for calculating the mode when we have a frequency distribution. There is some relation between mean, median and mode. So, many times this is used like x bar minus M naught is equal to three 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 modal 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 modal 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 into 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 1. 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 the 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 is equal to 1 to n that is, the reciprocal of the arithmetic means of the reciprocals. That is called the arithmetic mean.

(Refer Slide Time: 57:23)



Measures of Dispersion

$$S^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2$$

$$\sigma^2 = \frac{1}{n} \sum (x_i - \bar{x})^2 \rightarrow \text{variance}$$

$\sigma \rightarrow$ standard deviation

$X_{(n)} - X_{(1)} \rightarrow$ Range

$$\frac{1}{n} \sum |x_i - \bar{x}| \qquad \frac{1}{n} \sum |x_i - x_M|$$

Mean Deviation

Now, apart from the measures of central tendency, we have measures of variation or 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 this. We may use 1 by n minus 1 here, sigma 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 variance.

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 minimum. We may also look at mean deviation; that is 1 by n sigma xi minus x bar or 1 by n sigma x i minus some x m, where x m is the median.

So, these are called mean deviations. All of this gives information about the variability in the data. Then, when we take sigma that is the 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 you in the next class.