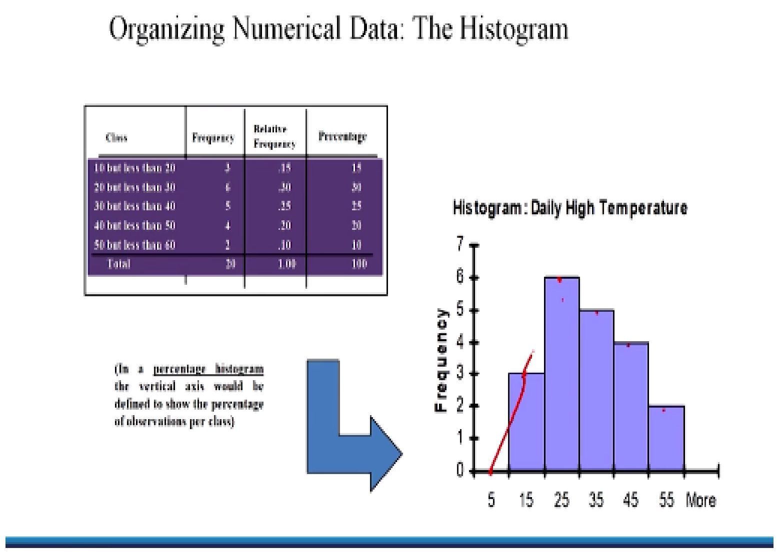
Business Statistics
Prof. M. K. Barua
Department of Management Studies
Indian Institute of Technology - Roorkee

Lecture – 05
Data Representation Techniques - Part II
and Measures of Central Tendency - Part I

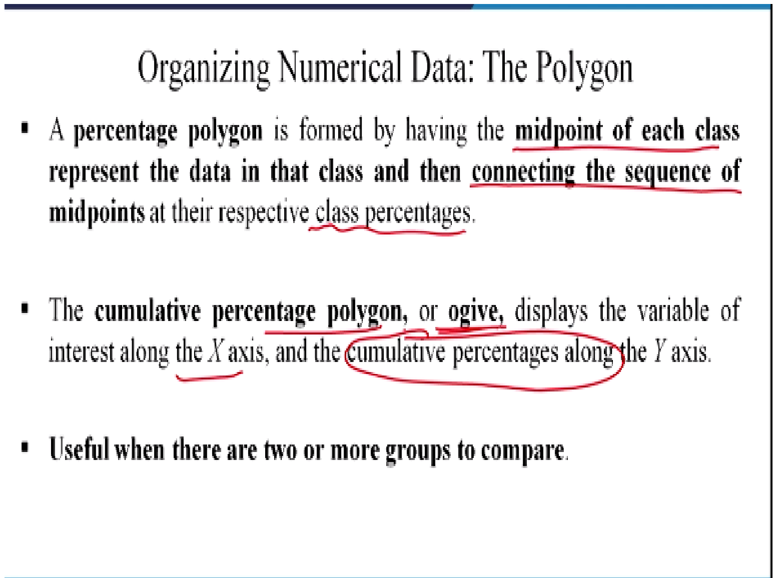
Good morning friends, I welcome you all in the session as you are aware in previous class, we were discussing how to present categorical and non-categorical data or matric or non-matric data. Today, we will see some of the other methods of presenting numerical data, so let us look at polygon. As you have already seen previous session that histogram is one of the methods of data presentation.

(Refer Slide Time: 01:02)



Just look at this, now the polygon is similar to histogram only the point is you need to connect midpoints of these classes, then it becomes a polygon, right, this you can see in next slide, right.

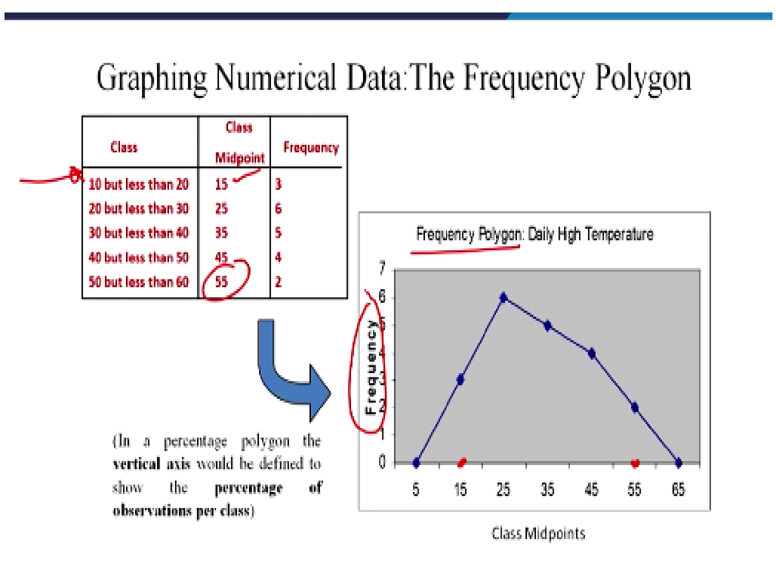
(Refer Slide Time: 01:17)



So, what is a polygon; is formed by collecting the mid points together, right now, midpoints I would you get; get each class will have a midpoint right, let us say if is let us say the class interval is let us say 1 to 10, so midpoint would be 5, right. A cumulative percentage polygon, you can also call it ogive, it displays the variable of interest along X axis and the cumulative percentage rate along Y axis, okay.

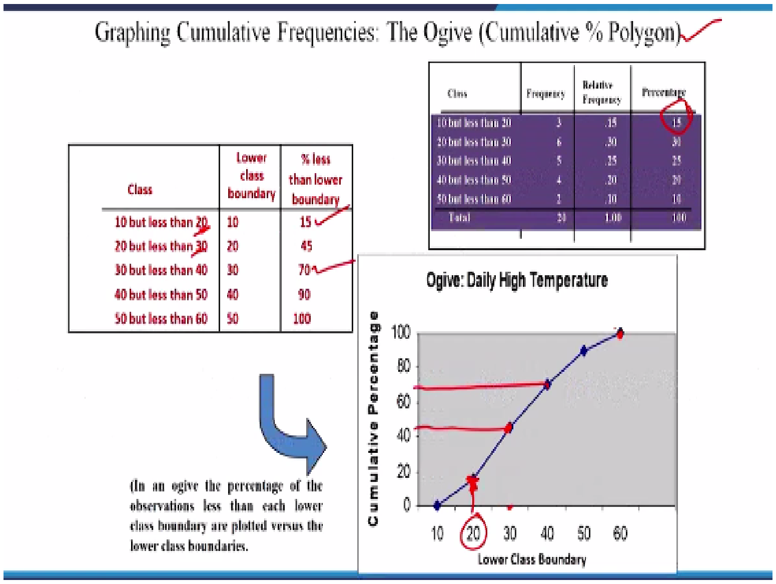
So that is the difference between percentage polygon and a polygon, a polygon will connect all the midpoints but it would be accumulative in case of ogive, it would be a cumulative percentage, useful when you want to compare 2 groups of data, right.

(Refer Slide Time: 02:19)



So, this is a frequency polygon, we will take the same example, so the midpoint for first interval is; for first class is 15 and for the last one it is 55, right, so this is 55, so you just connect all these midpoints, now this is your frequency polygon, if you write over here percentage then it would become percentage polygon.

(Refer Slide Time: 02:49)

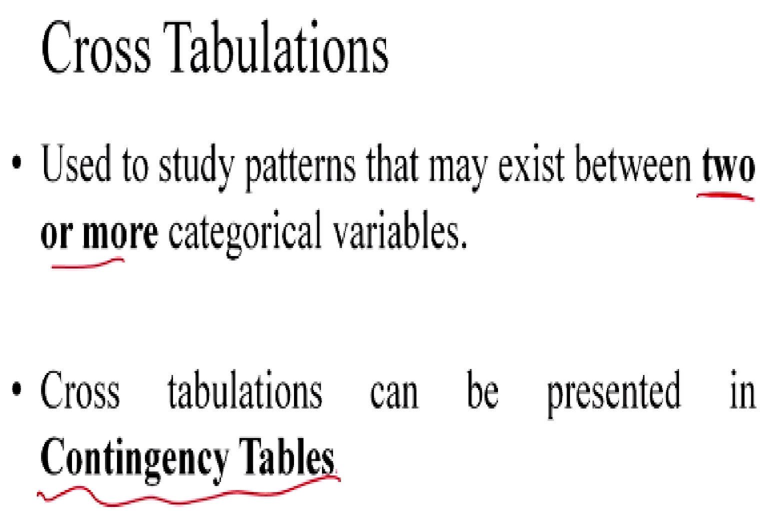


Ogive was I said is nothing but a cumulative percentage polygon, so rather than connecting midpoints of histogram here you have got a cumulative percentage right, so let us say the first one is 15, right, so this is 15, so for ogive you have got the upper limit is the base, right, so at this point you need to have this point similarly, next is 30, so this is 30 and this is the; next one is next cumulative percentage okay.

And similarly, you can have for example, the first one is 15 right, so this 15, next is 45, this 45, right, next is 70, so this 70, is not it and so on, so finally this one is 100, so this is ogive, now so for what we have done; we have represented numerical and non-numerical data in different forms be tabular form, chart, bar chart or pie chart or histogram or whatever it is but there are certain limitations of those charts.

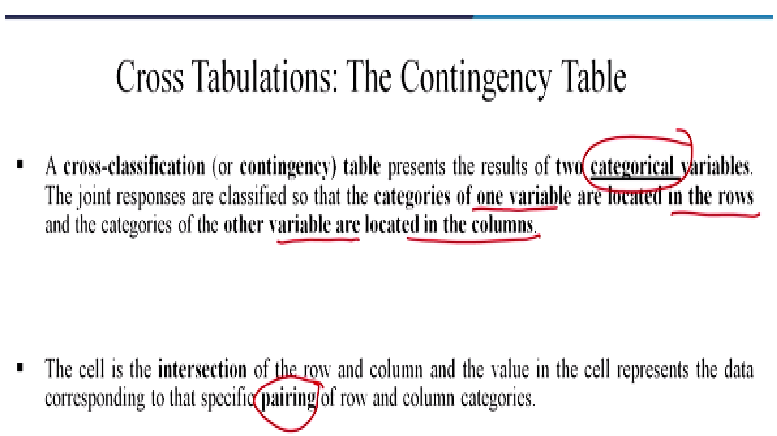
And the limitation is that we can use all of them only for one particular variable of interest now, many times what happens; you have got more than one variables and you need to find out some relationship between those 2 variables or you may have more than 2 variables as well is not it, so when you have got 2 variables, then frequency distribution will not work or polygon will not work, then what you need?

(Refer Slide Time: 04:46)



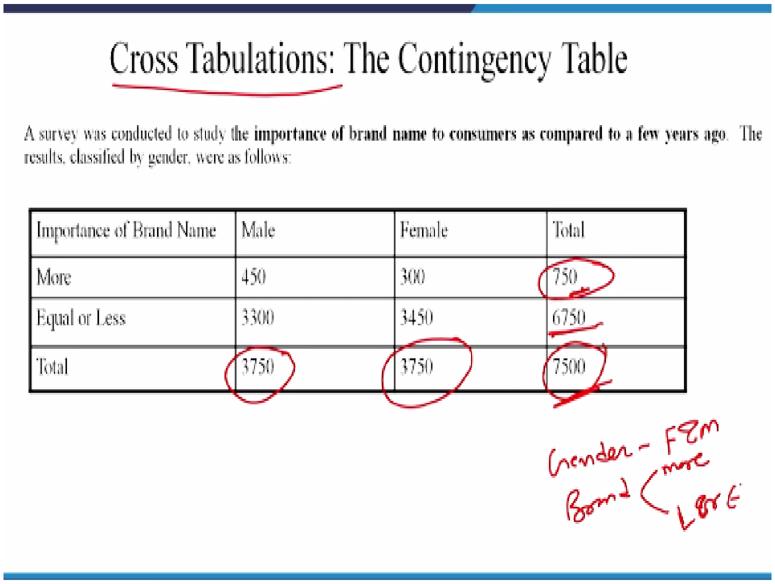
You need something called cross tabulations, so cross tabulations we generally use for 2 or more variables, specially categorical variables okay and cross tabulations can be presented in terms of contingency tables, right, so we will have a table in next slide that would be called contingency table, right.

(Refer Slide Time: 05:03)



So, a contingency stable presents the result of two categorical variables as I have already mentioned, the joint responses are classified, so that the categories of one variable are located in the rows and the other variable are located in column, right, so you will have some rows and some columns in contingency table. The cell is intersection of row and column and the values in the cell represent the data corresponding to that specific pairing of row and column list.

(Refer Slide Time: 05:44)



Importance of Brand Name	Male	Female	Total
More	450	300	750
Equal or Less	3300	3450	6750
Total	3750	3750	7500

So, let us look at an example, so this is your cross tabulation, so let us say you have done a study and you have asked to let say 7500 respondent and the question you have asked them is whether the brand which they are using is more important than what it used to be earlier, right, so the question was whether the brand which you are using is now more important than what it was earlier and data were collected from 7500 respondents.

Now, there are as I said cross tabs are for 2 variables, so which are 2 variables, so we collected data let us say from 3750 males and the same number of females in your survey, so you have got males and females, right, now 750 respondent have said that the importance of brand is now more than what it used to be earlier and 6750 set; no, it is not more but it is equal or less than that so, there are 2 variables, which are 2 variables, which are 2 categorical variables, can you identify from this table?

The first one is gender right, so gender you have female and males right and the second one is; is the brand, is not it, so brand has again got 2 categories whether it is more preferred than earlier or it is less than or equally preferred, right, so this how you can have use of cross tabulation.

Now, as I said there are thousands of data representation tools or charts, graphs or plots whatever you call them.

(Refer Slide Time: 08:08)

Scatter Plots

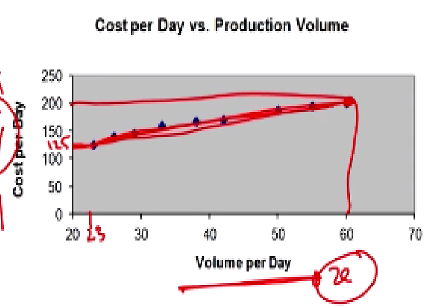
- **Scatter plots** are used for numerical data consisting of paired observations taken **from** two *numerical variables*
- One variable is measured on the **vertical** axis and the other variable is measured on the **horizontal** axis
- Scatter plots are used to examine possible **relationships** between two numerical variables

Now, there is something called scatterplot, now scatterplot is again a type of plot wherein you are using 2 numerical variables right, so can you may find out; you can find out actually a relationship between these 2 variables whether they are related or not first of all, if related then whether they are positively related or not related, so you have got 2 numerical variables on one axis you have got one variable and the other one would be on horizontal axis. As I said it used to examine possible relationship between 2 numerical variables, so let us look at an example.

(Refer Slide Time: 08:56)

Scatter Plot Example

Volume per day	Cost per day
23	125
26	140
29	146
33	160
38	167
42	170
50	188
55	195
60	200

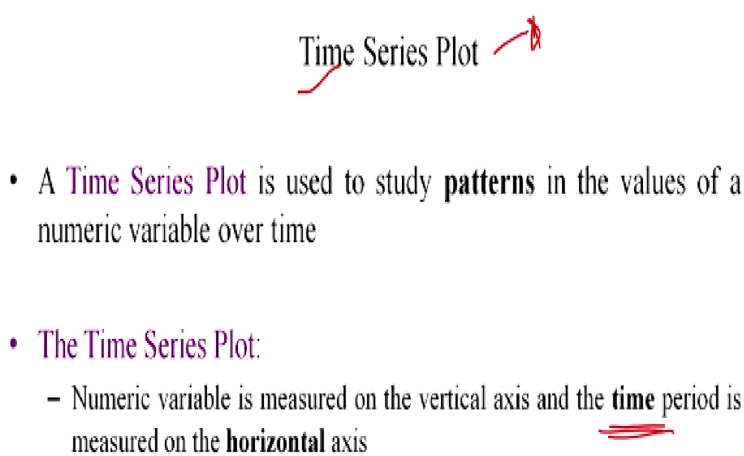


So, volume per day and cost per day, let us say you are manufacturing some let us say liquid, right liquid, let us say beverages, right, so let us say from the assembly line, you are making let us say 23 litre of beverage and the cost is 125 rupees per day similarly, for 60 litres of beverage the cost is 200 and there are other values in between, so you can have these 2 axis, so on x axis you have got volume per day and on y axis you have got cost per day, right.

So, we want to find out is there any relationship, so this is how you can draw a scatter plot like this, so the first one is 23, so this is 23 and this value is 125, okay and so on, so finally this one is 60 and 200, right, so this is your 200 point, right, so if you connect all these; you know dots and this becomes a scatterplot. So, do you see is there any relationship? Yes, there is a relationship, you can say that is the volume per day increases the cost also increases slightly, is it not.

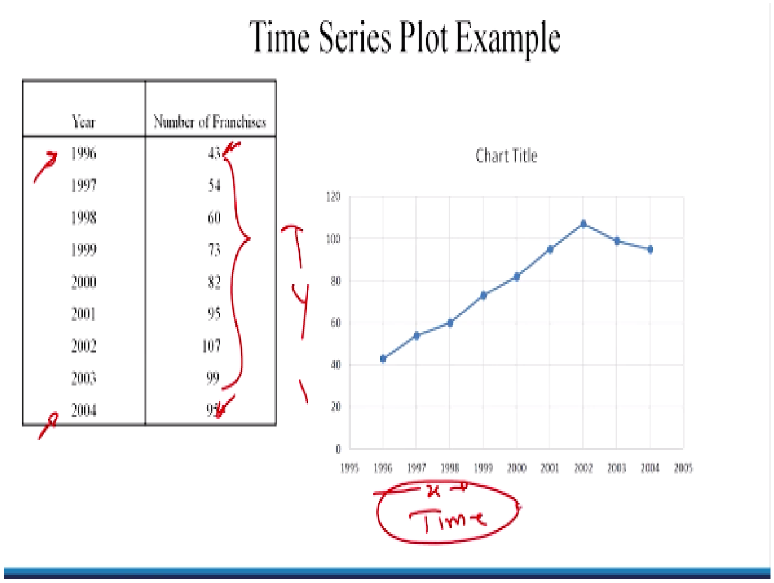
So, there is a positive relationship, when the variable on x axis increases the variable on y axis also increases, so this is the case of positive relationship, so this is scatterplot.

(Refer Slide Time: 10:41)



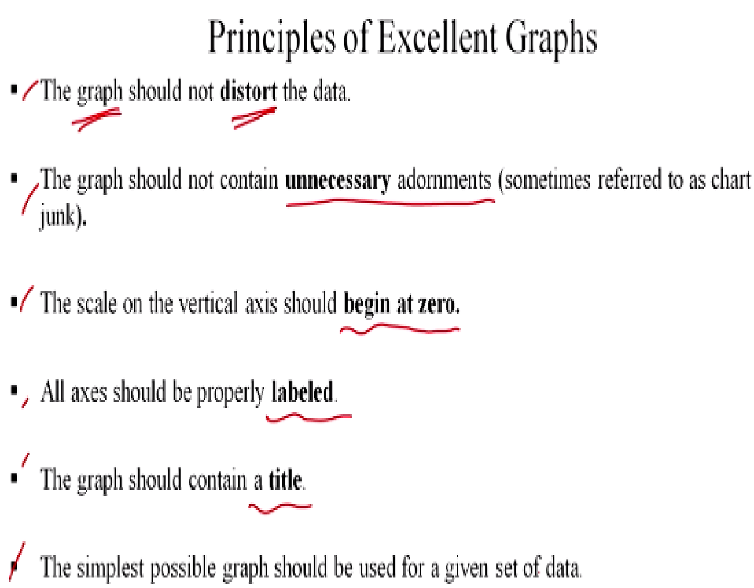
Let us look at another plot, it is called time series plot, now times series plot and scatter plot, these two are almost similar only there is one difference that in times series plot on x axis, you will always have time, right and on y axis you will have a particular variable, right, so this is again for 2 variables; variable 1 is on x axis, let us call it time and variable y is some any other some other variable.

(Refer Slide Time: 11:23)



So, let us say the number of franchises of a product or of a form in 1996 were 43 and in 2004, 95 and other values in between, so how would you plot times series plot, so you just take on x axis this is your time and on y axis is your number of franchise, so the difference is; difference between the scatterplot and times series plot is here you will always have time on x axis.

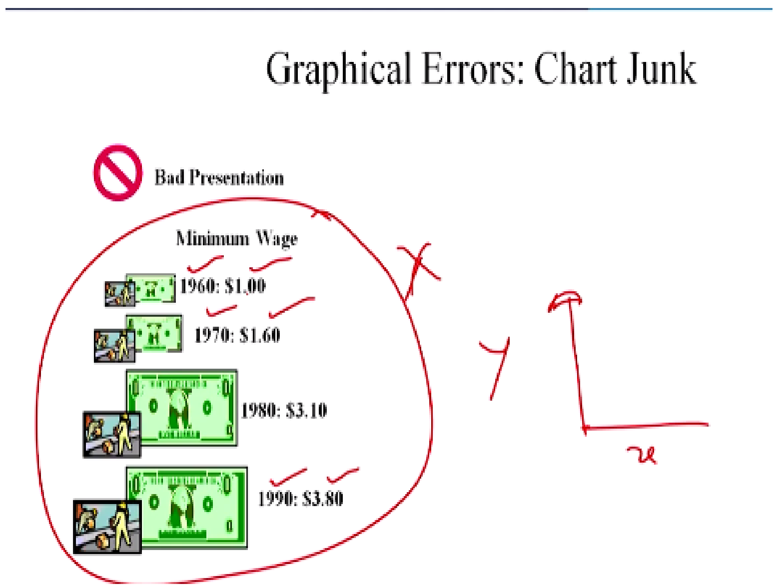
(Refer Slide Time: 11:57)



Now, whenever you go for drawing graphs, there are certain principles, some guidelines which you should keep in mind, so the graph should not distort the data because whenever you collect data, you should represent in a graph properly, you should not distort the data while preparing a graph, graph should not contain unnecessary adornments, so you just; you should not try to give each and every information in the graph, you just give whatever is required.

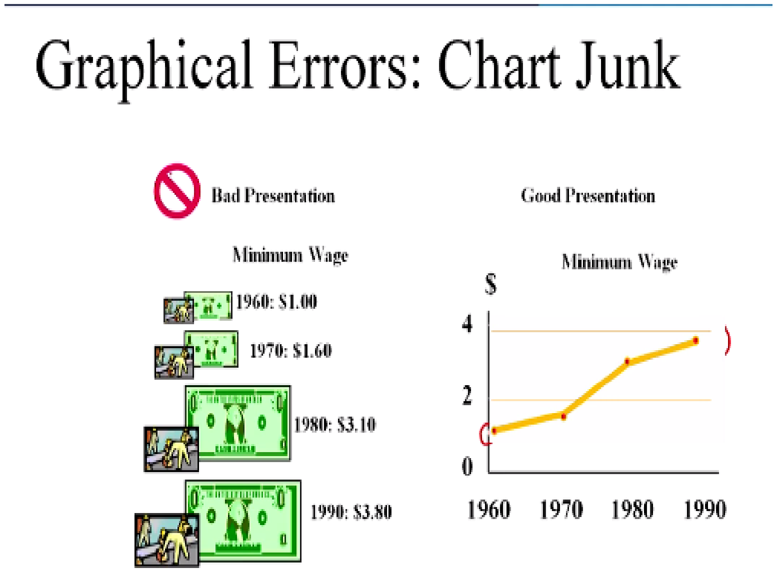
The scale on the vertical axis should always begin at 0, right, we will see example of each one of these points, okay, all access should be properly labelled; labelled means whether x axis is time or let us say volume or cost or whatever it is, right, the graph should contain a title of course, it is good to have title in each and every graph, so the simplest possible graph should be used for a given set of data, right.

(Refer Slide Time: 13:15)



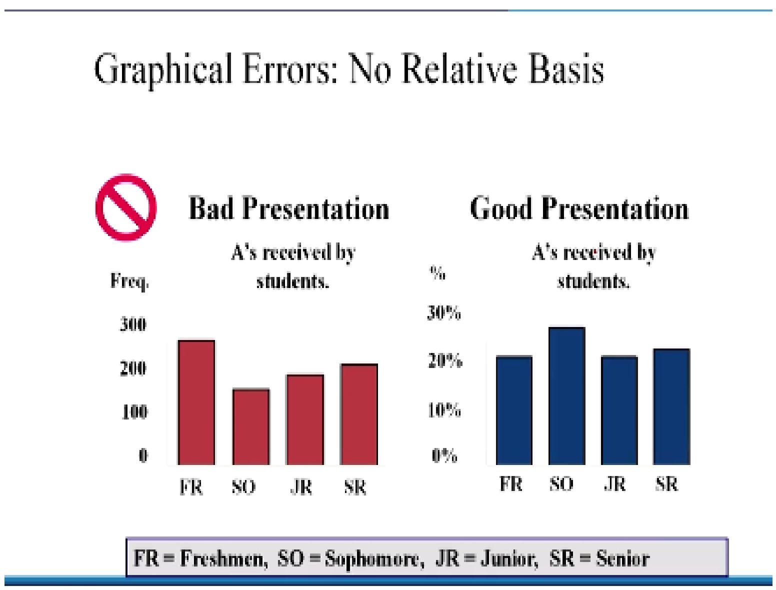
So, let us look at some of the examples of graphs, so this is; you have got a data, so the minimum wage in 1960's 1 dollar, 1970, 1.6 dollars, 1990, 3.8 dollars, right, so is this a good representation, no this is not a good representation, this is a bad presentation of data, so what is wrong here, can you think for a while, what is wrong in this particular slide? The point here is that you can have a better graph by having time on x axis and on y axis wages, is it not.

(Refer Slide Time: 14:10)



So, let us look at this, this is how you should represent, so in 19; 1960, this is the wage 1 and in 1990 wage is 3.8 dollars, right, so try to avoid making such mistakes.

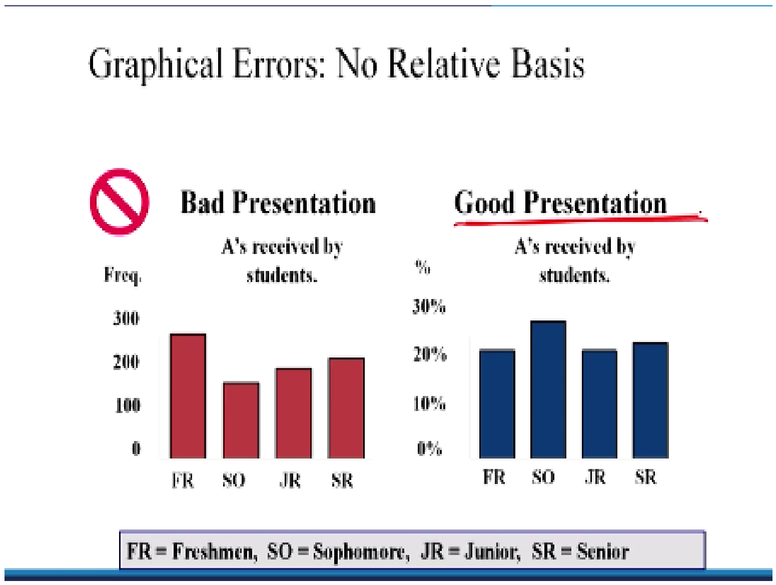
(Refer Slide Time: 14:28)



Let us look at another example, now if you look at this, this is again a bad presentation of data, so let us say these are different let us say grade A received by students, so freshman these many students received grade A, juniors these many, seniors these many, right, so there is a better way of presenting same information and how you can do it; just think for a while, just for 30 seconds how would you improve presentation of this information?

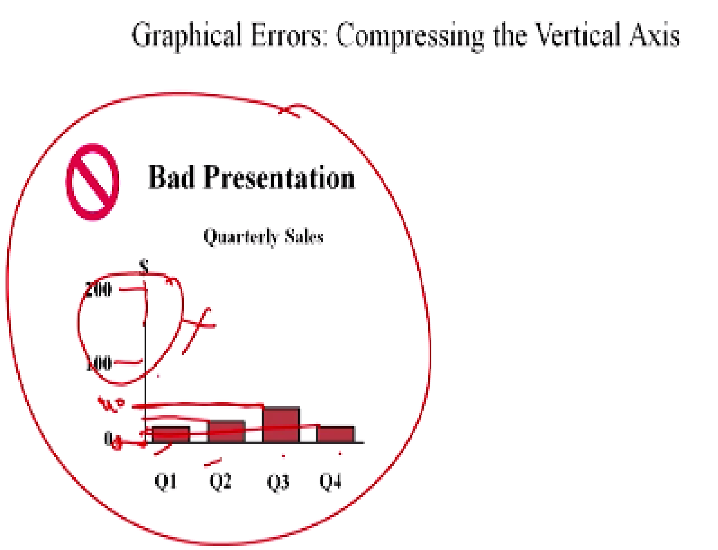
So, you can do it by having all these students on x axis just like here we have given but on y axis you can have percentage frequency, right percentage frequency, okay.

(Refer Slide Time: 15:33)



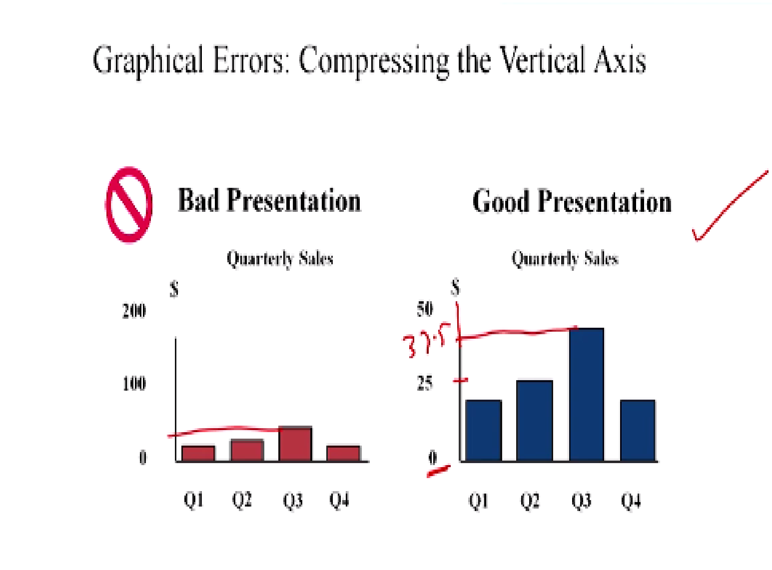
So, let us look at this, just see this, 0%, 10%, 20%, 30%, right so this is a good presentation of data, okay, so let us look at one more example.

(Refer Slide Time: 15:47)



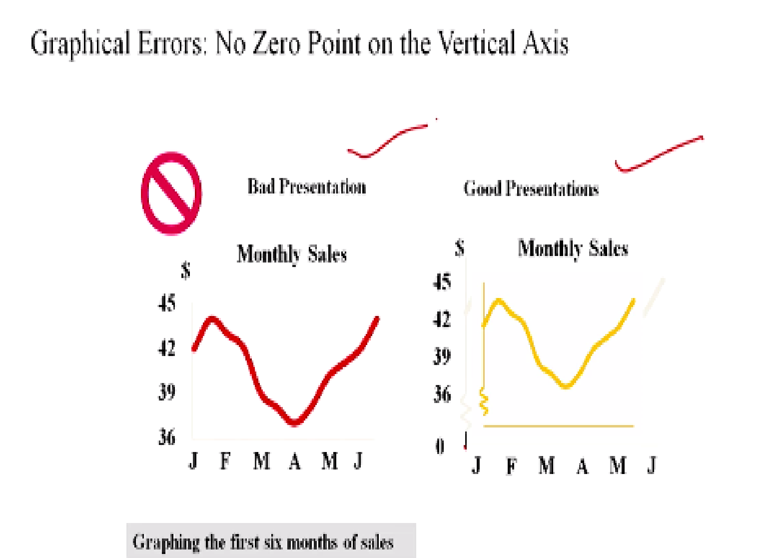
This data, right, so quarterly sales, first quarter, second quarter, third and fourth quarter, right so quarterly sales hour here is this, for second quarter this, third and fourth is this, right, so what is wrong here, just 10 seconds, what is wrong here? Now, if you look at this particular graph then this point is 0 absolutely right but all these the maximum value is let us say this is 40 or 50, right but this value and this value do not have any meaning, is not it.

(Refer Slide Time: 16:44)



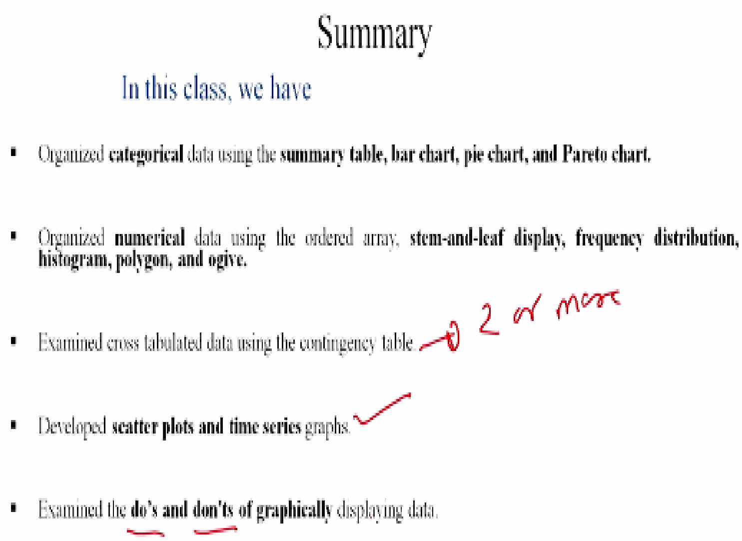
So, you can prepare a better graph, so start with 0, then 25, next value is let us say 50 and so on, right, so here you can easily identify what is; what was these, you know quarterly sales, let us say in Q3, so this comes around let us say 37.5, right but from here you do not know how much it is, is it not, so this is a better graph.

(Refer Slide Time: 17:22)



Let us look at this one, what is wrong here, it is very simple, written here, no zero point on vertical axis, so monthly sales from January to June, so in month of January, it is let us say 42 but you should always begin y axis with 0 point, right, so this would be 0, then 36 and so on, right, so this is a good presentation, this is a bad presentation, okay.

(Refer Slide Time: 18:03)

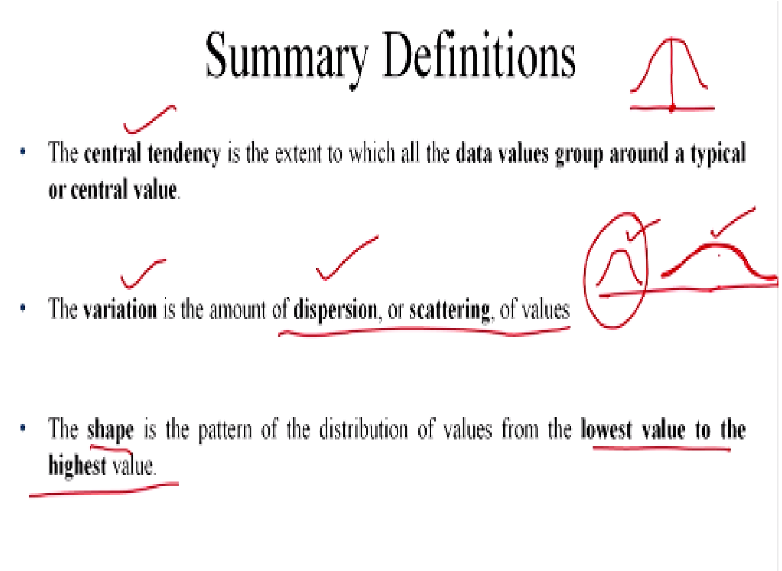


So, let us summarise what we did in this session so far and what we did in previous session, so we have seen how to organise categorical data in terms of table, pie chart, bar chart or Pareto chart, how to organise numerical data in terms of stem and leaf diagram, frequency distribution, polygon, ogive and histogram, we have examined, we have also seen what is cross tabulation, as I said when there are 2 or more variables, right then use cross tabs. We have also seen scatterplot and times series plot and we have seen do’s and don'ts of graphing, okay.

(Refer Slide Time: 19:02)

So, let us look at next topic which is numerical descriptive measures now, what we will see in this particular session is we will talk about different measures of central tendency, different measures of dispersions. So, first of all you should know what is central tendency whenever you have got data, there would be one point on which all data would be resting our lying up on or dependent upon, so that point is known as central tendency.

(Refer Slide Time: 19:49)

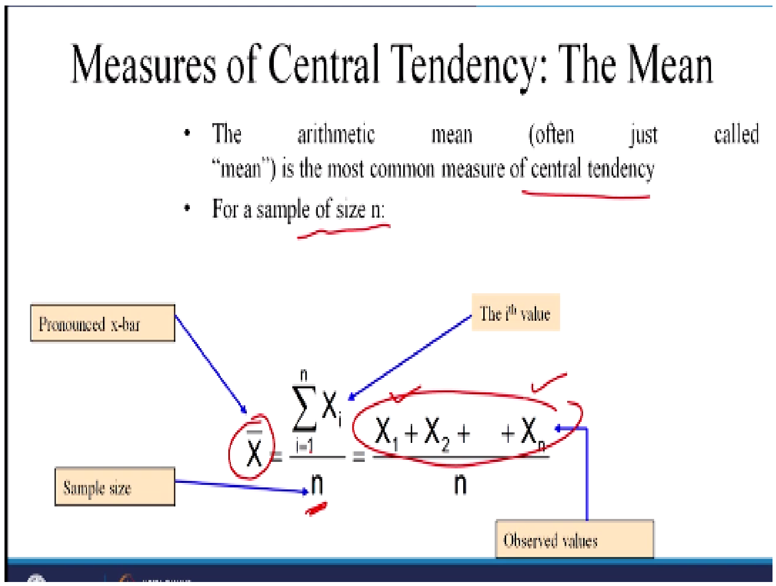


So, let us say you have got a distribution of data for example, let us say this data, right, this is your distribution right, so you can say that this is the point where this particular distribution rests, right, so this would be called central tendency now, since your data would be different, right, it would not always be normally distributed data, you will have skewed data, in most of the times you will see skewed data, right, so this is central tendency.

The variation is the amount of dispersion or scattering of values, so what is dispersion; so, let us say; let us take an example, so this is your distribution, right and this is your another distribution, right, so which one is more dispersed, this one is more dispersed than this, right, here you have data concentrated, right, here they are dispersed, right, so central tendency variation, there is something called shape; is the pattern of the distribution of the value from lowest value to the highest value.

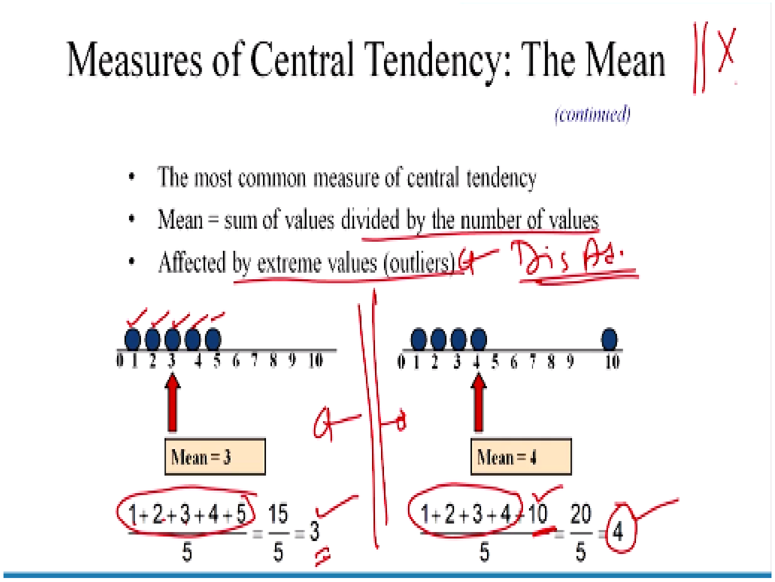
So, how this shape is changing in the distribution, now there is something called kurtosis; kurtosis is the how the height of distribution, right, so for a normally distributed data, the skewness is 0 and kurtosis is 3, so this is one of the ways of defining normal distribution that for a normally distributed data, kurtosis is 3 and dispersion is 0, right and a skewness is 0, right. So, let us look at different measures of central tendency.

(Refer Slide Time: 22:05)



The very first and most important is mean or you can simply call it arithmetic mean or average is the most common measure of central tendency for a sample of size and so, let us say there are different data points, so X1 to let us say Xn, so you just take summation of all these data and divided by n, right, so this is your X bar, right so i is ranging from 1 to n, right and n is the sample size, so this is and all these are your observed values.

(Refer Slide Time: 22:33)

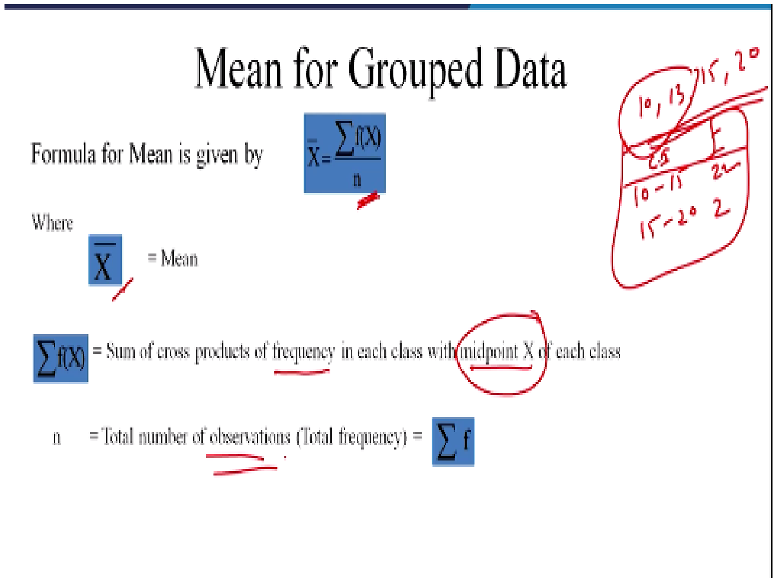


So, this is how you can calculate mean, let us take this example, so it is the most common measure, mean = sum of values divided by number of values affected by extreme values, so this is basically a disadvantage of mean, when you take this as a measure of central tendency, so let us take this example. So, you have got let us say data point; 1, 2, 3, 4 and 5, so what is the mean; you just take summation of 1, 2, 3, 4, 5 and divided by 5, so this is mean, right.

Let us look at this example, so this is; there are 2 examples, this first example this second example, so data points are 1, 2, 3, 4 and 10, this is mean is 4, so as I said if you look at these 5 data points; 1, 2, 3, 4, 5 mean is 3 but if you take 5 data points here, 1 2, 3, 4 and this is 10, right, so you can say that this is an outlier okay, though there are different methods of finding outliers but here just by looking at this pattern, you can see 10 is an outlier.

That is why the mean has gone up from 3 to 4, so not a very good method or measure of central tendency okay.

(Refer Slide Time: 24:09)



Let us look at another method is called mean for grouped data, so you can have 2 types of data; grouped data and ungrouped data, what is the difference? Grouped data is a data wherein you are grouping data by writing frequencies in different classes, so frequency distribution is nothing but the kind of grouped data, right, so let us say you have got data points, let us say 10, 13, 15 and 20, right, okay, so this is ungrouped data.

But let us say if I want to have 2 classes, let us say 10 to 15 and 15 to 20, these 2 are different classes, so 10 to 15, there are 2 data points, so this is frequency, this is; this your class interval, right frequency 2 and frequency 2 here, right, so this is; this becomes a grouped data right, so how would you calculate mean in this case, X bar = summation of fX divided by n, right where of course, this is mean and this is summation of cross products of frequency in each class with midpoint of each class.

This is important, right, so X is nothing but midpoint of each class and n is total number of observations or total frequency, right, so we will take an example and we will calculate mean of grouped data.

(Refer Slide Time: 25:49)

Mean for Grouped Data Example

Find the arithmetic mean for the following continuous frequency distribution:

Class	0-1	1-2	2-3	3-4	4-5	5-6
Frequency	1	4	8	7	3	2

So, let us take this example, find the arithmetic mean of the following continuous frequency distribution, so there are 6 classes, first second third fourth fifth and sixth, right and you have been given frequencies as well, so frequency in first in class one is 0 to 1 is the interval is 1, frequency here is in class interval 2 to 3 is 8 and when frequency is 2, the class interval is 5 to 6, right.

So, how would you find mean of grouped data? So, first of all you need to calculate midpoint of each class, is not it, so what would be the midpoint of class 1 or this class; 0 and 1, midpoint is 0.5, is not it, for this 1.5, right and so on, for this 5.5, is not it, so all these are X values, right.

(Refer Slide Time: 26:50)

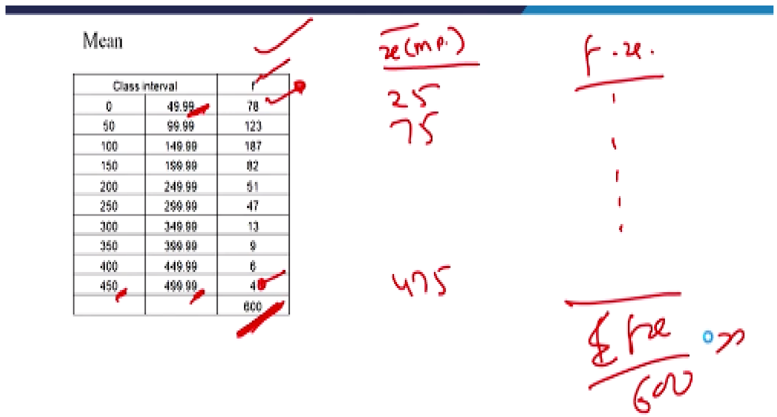
Solution for the Example

	A	B	C	D
1	Class	X (mid pt)	f	fX
2	0-1	0.5	1	0.5
3	1-2	1.5	4	6.0
4	2-3	2.5	8	20.0
5	3-4	3.5	7	24.5
6	4-5	4.5	3	13.5
7	5-6	5.5	2	11.0
8	Totals		25	75.5
9	Mean			3.02

Applying the formula $\bar{X} = \frac{\sum f(X)}{n}$ $= 75.5/25 = 3.02$

Just see this, so all these are your classes, this is midpoint, 0.5 to 5.5 and of course, you have got frequencies, so 1 to 2, right, so total frequency is 25, so fX is this, so column A, B, C and D, right, so this fX, so just multiply column B and C, you will get these values and this is the mean, right and this is the total; total of this is the total height, total is here, 75.5, so just 75.5 divided by 25, you will get 3.02 as mean, right.

(Refer Slide Time: 27:45)



Class interval		f
0	49.99	78
50	99.99	123
100	149.99	187
150	199.99	82
200	249.99	51
250	299.99	47
300	349.99	13
350	399.99	9
400	449.99	6
450	499.99	4
		600

So, this is very simple method of calculating mean for grouped data, so let us move on to one more example wherein there is grouped data available and the example is like this, so you have collected data of 600 account holders in a bank and you have, you want to know what is the main balance in those 600 accounts, so when you collected data of 600 customers, you found that the there were 78 such account holders who are having their balance between 0 to these many rupees this as good as 50, right.

Similarly, there were only four accounts in which the balance was between 450 to 499.9 or 450 to 500, so how would you solve this example, it is very simple, you have been given frequencies, you have been class intervals, you just write down mid points, right midpoint, what would be the midpoint here; let us take this as 50, right so, first midpoint would be 25 then 75 and so on and finally, it would be 475, right.

So, just multiply a f and x, right, this is f and this is x, take the total of f and x, so this is summation of fx, you have got total frequency 600, we just divided by 600, you will get mean, right.

(Refer Slide Time: 29:35)

Class interval		f
0	49.99	78
50	99.99	123
100	149.99	187
150	199.99	82
200	249.99	51
250	299.99	47
300	349.99	13
350	399.99	9
400	449.99	6
450	499.99	4
		600

By taking mid values as 25, 75,... 475.

$$\bar{X} = \frac{\sum f(X)}{n}$$

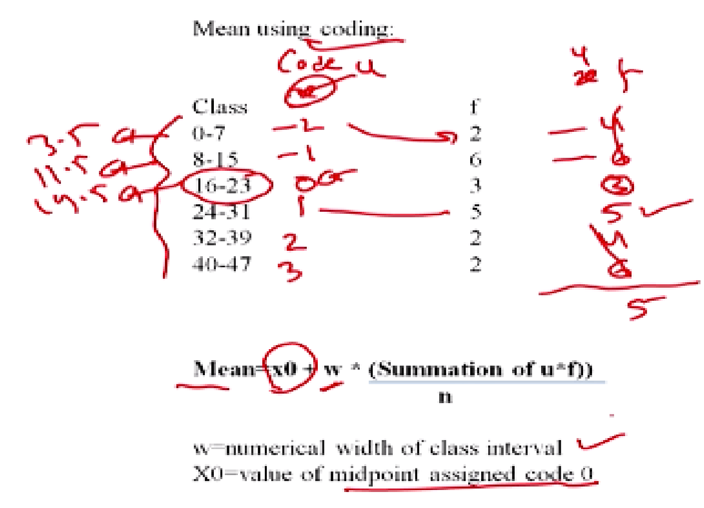
Mean: 142.25

f(X) = 85350

n=600

So, this is how you can solve this example, okay, so fX in this case is 85350, is not it, this is nothing but 85350, right 85350, so divide this value by 600, and you will get mean as 142.25, so this is how you can calculate mean of grouped data and ungrouped data, right.

(Refer Slide Time: 30:11)

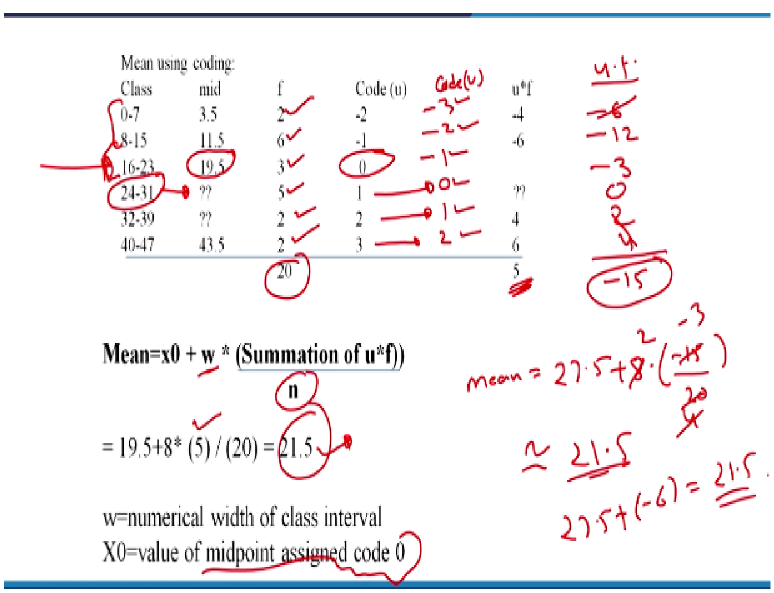


So, let us look at another method of calculating mean using coding, now this is a simpler method by which you can get answer easily, now coding means what; we generally give different codes to these classes for example, let us since here you have got how many; 6 classes right, so you can you coding let us say, 0 to class number 3, then -1 to this class and – 2 to class number 1, right similarly, 1, 2 and 3.

Now, the next step is just multiply, so let us call this as X value, right, this is nothing but code, right, now multiply x and f right, x and f, so this becomes let us say this is -4, -6, 3 sorry, this is 0, right, 3 * 0 is 0, this is 5, this is 4 and this is 6, right, so this total is now what; you just this 4 and 6, you can cancel with this 4 and 6, your remaining is 5, so this 5, right. Now, how to proceed further?

So, mean = x0 + w * summation of u * f, so let us call it u rather than calling it x, right, so this is u, this is also u, right now, x0 is the value of the midpoint assigned codes 0, value of the midpoint assigned codes 0, so code 0 is assigned to this particular class and what is the midpoint of this, so let us say, so the midpoint of this class would be 3.5, then this would be 11.5, this would be 19.5, is not it, so that is the mid value assigned to code 0, right.

(Refer Slide Time: 32:32)



The numerical width of the class interval is 8, right so this how you can proceed further, so this is your summation of u f, now x0 is as I have already mentioned is this, 19.58, right, plus w, what is w, numerical width, so this is width, right, 0, 8, 16 -8 is 8 and so on, right, this is summation of u and f which is 5 divided by n; n is this 20, so this becomes 21.5 as the mean. Now, in this case what we have done?

We have assigned zero code to class number 3, right, so there is no fixed rule that you should assign zero rule; zero code 2, let us say third or fourth or fifth whatever it is, you can assign zero code to any of these class intervals, there is no change in answer, so let us move forward and as change these code and assigned these code 0 to class number this which is fourth class, right, so 21 to 31 is now we have assigned zero code.

Now, this is -1, this -2 and this is -3, this is 1 and this is 2, right, now what would be the u and f multiplication, so this is u and this is f, right, so this is -6, then 6 and -2, -12, the next one would be 3 and -1, -3, then 5 *0 is 0, 2 * = 2; 2 * 2 = 4, now you can calculate what this summation would be, so this would be -15, so this 6 and this 6 gone, right, so this is -12, -3, -15 now, this -15 you need to put it in formula.

So, what would be the mean now, mean = x0 is the midpoint of the class where we have assigned 0, so this is what would be the midpoint here for this, it would be 27.5, right, so 27.5 + 8 * -15 divided by 20 okay, so this summation of u and f is -15 and we are dividing it by n which is 20, so this value is approximately anyway, you will get this answer 21.5, so let us calculate -3, 4, this is 4, you just divided by; so 8 divided by 4, this is 2.

So, 2 * -3 is -6, so 27.5 + -6 which is = 21.5, so this another method of calculating mean using coding, so with this let me complete this session, we will see some more methods in next session thank you.