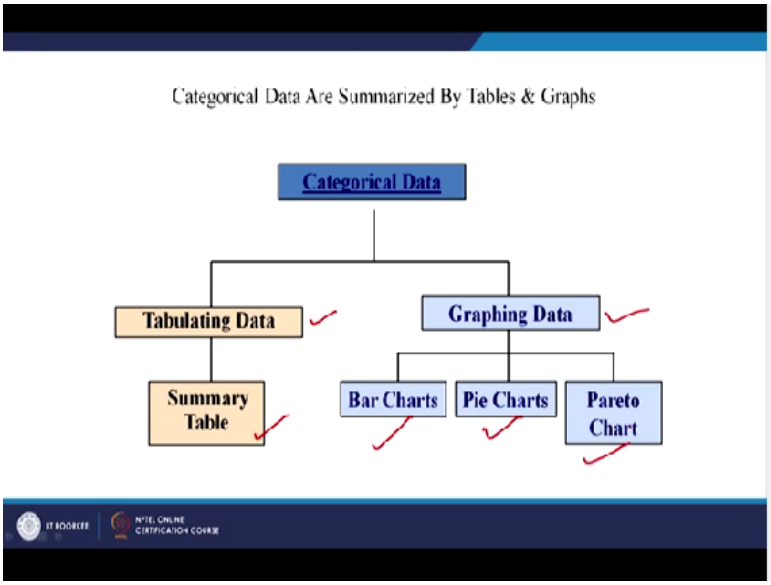
Business Statistics
Prof. M. K. Barua
Department of Management Studies
Indian Institute of Technology - Roorkee

Lecture – 04
Data Representation Techniques - Part I

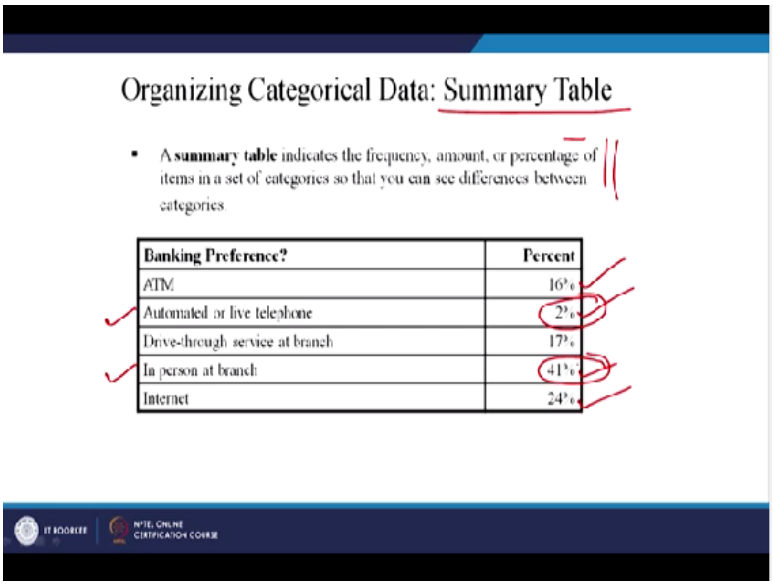
Good afternoon friends. I welcome you all in this session. In today's session, we are going to learn how to present data once you collect data either from primary source or from secondary data. And there are different ways in which you can present data. Let us look at couple of data representation techniques. So as I said basically there are 2 types of data, metric and non-metric, right. So you can call them categorical and non-categorical. Is it not? So let us look at first categorical data.

(Refer Slide Time: 01:01)



We can summarize categorical data by tables and graphs. So we have got tabular data and we have got graphical data. In tabular data, we can represent data in a table form. In graphical data, we have got bar chart, pie chart, pareto chart and there are 100s of other charts. So we would be looking at very few data presentation techniques. So let us look at first of all summary table.

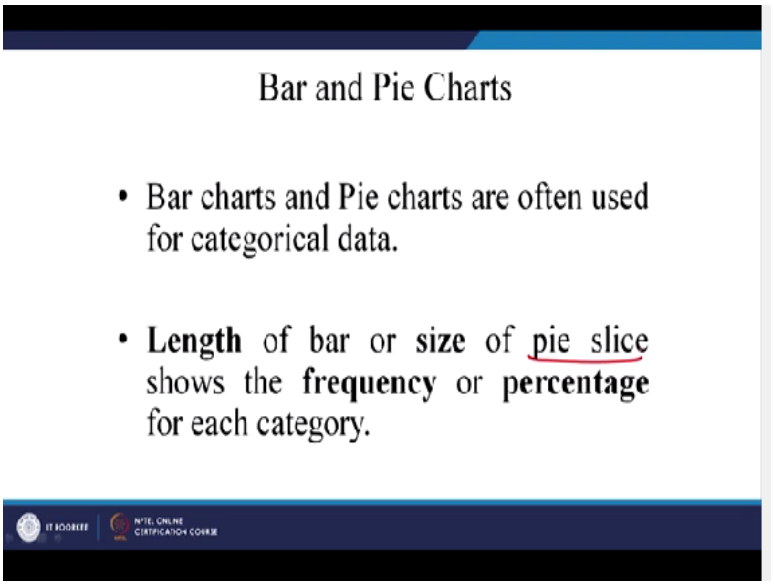
(Refer Slide Time: 01:33)



So there is an example. So a summary table indicates the frequency amount or percentage of frequency in a set of category so that you can see differences between categories. So let us say you have collected data from let us say 1000 people and you have asked them a question about their banking preferences. So out of 1000, 16% said that they would like to use ATM, 2% would like to use live telephone, 17% drive through service at a branch, in person 41% and internet 24%.

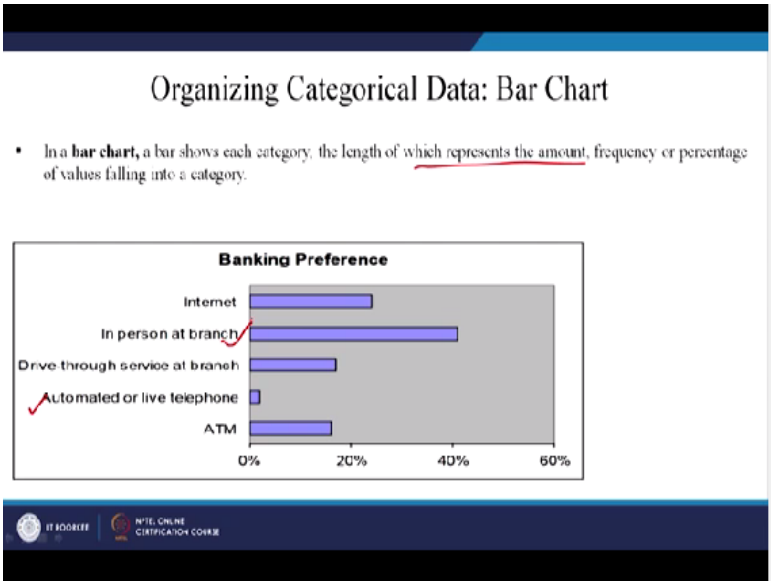
So whenever you talk about any table like this, you need not talk about each and every value in this particular column. You just say that, you should talk about only the least value and the maximum value. For example, in this case we will say that automated or live telephone 2% and in person at branch 41%. So you need to just give the minimum and maximum value. No need to talk about each and every value, right. So this is how you have to present data in tabular form.

(Refer Slide Time: 03:02)



As far as bar and pie charts are concerned, you have got bar charts and pie charts are often used for categorical data or non-metric data. The length of the bar in bar chart and size of the pie slice will tell you the frequency or percentage of each category. And what are those categories? ATM, internet and so on, right. So you can have different categories. Let us look at bar chart.

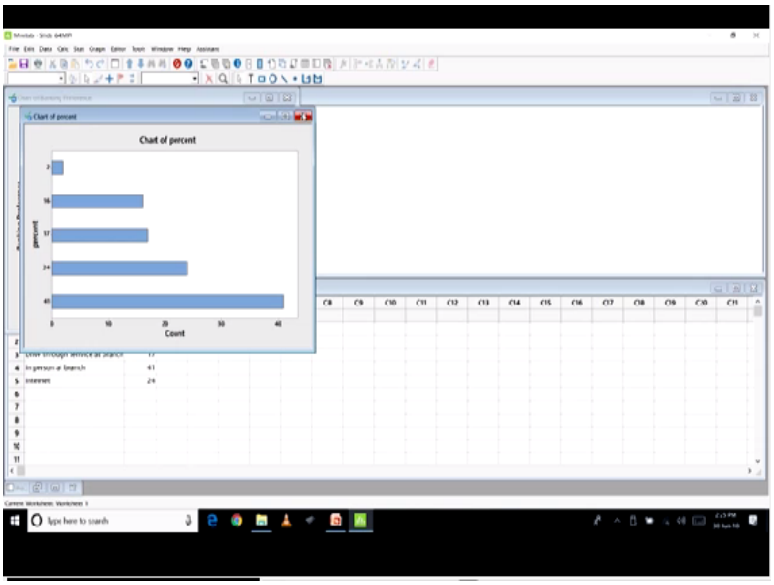
(Refer Slide Time: 03:34)



So in a bar chart, a bar shows each category, the length of which represents the amount, frequency, cumulative frequency, percentage or cumulative percentage, okay. So same information what we presented in tabular form is available here in bar chart. So we will say it is 2% automated or live telephone and this is 41% in person at branch. So you can have, this kind of bar chart, you can prepare using different softwares. So I will show you how to present bar

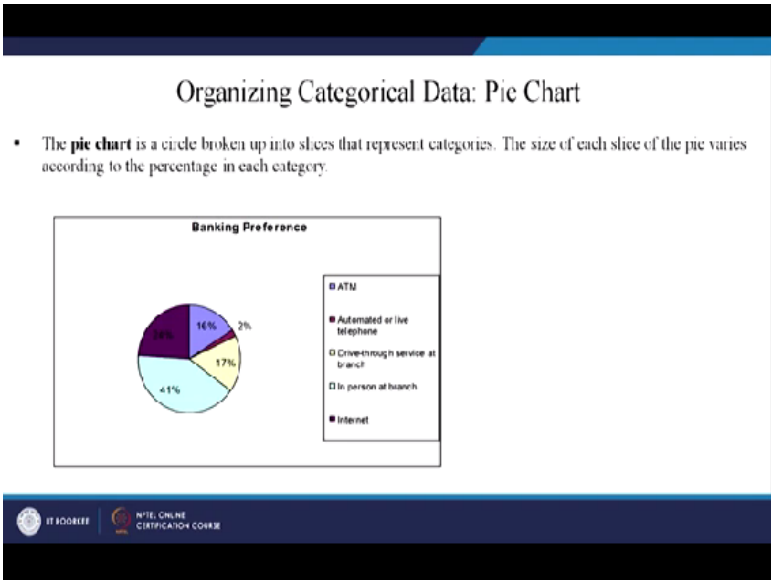
chart using Minitab software. So let us look at Minitab software.

(Refer Slide Time: 04:21)



So let us look at how to prepare a bar chart using Minitab. So you need to enter data first of all in worksheet. Once you are done with data entry part, go to graph. In graph, go to bar chart. Just click over here and okay and you need to select these 2 columns, C1 banking preference, C2 percentage. So select both of them and then press okay. So you will get a bar chart like this for this particular example. So we will move on to slides once again.

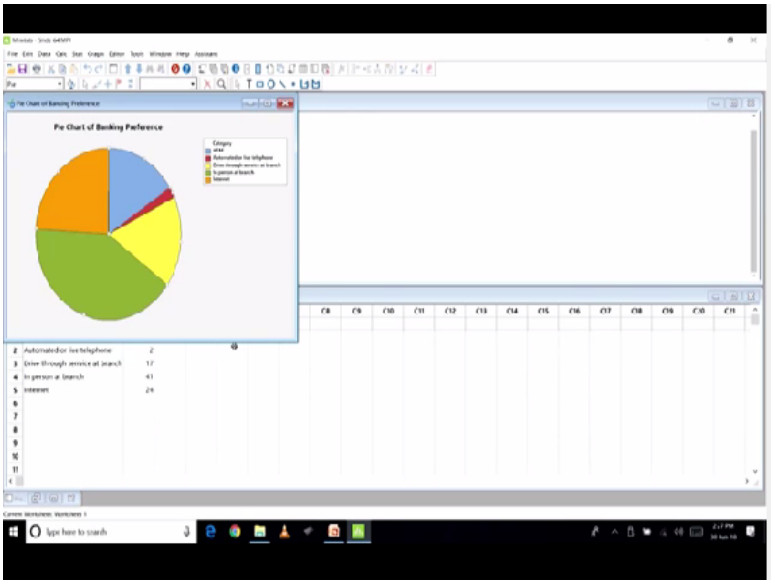
(Refer Slide Time: 05:06)



The next one is pie chart. A pie chart is a circle broken up into slices and those slices would represent what? Different categories. Categories would be represented in terms of their

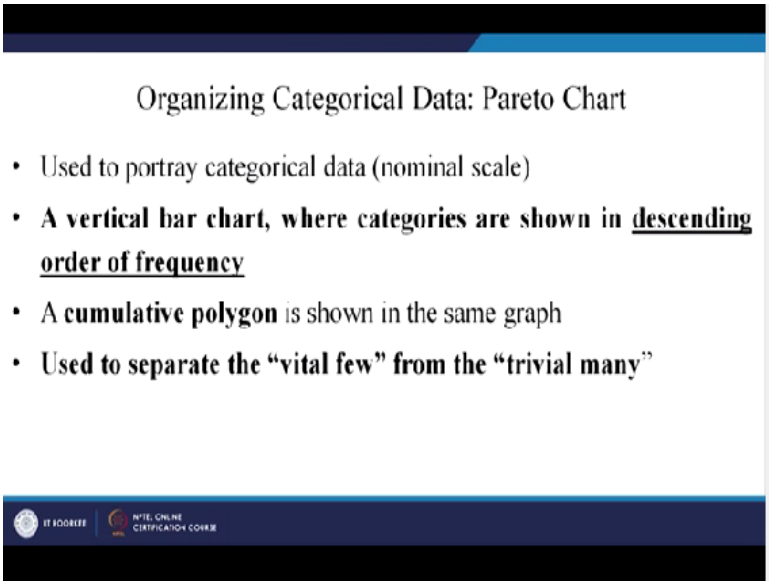
frequencies or percentage, frequency percentage, or cumulative percentage and so on. So let us look at this. So the same information is available here in pie chart as well. So this is 41% in person at branch and this 2% live telephone, right. So we will now draw a pie chart using Minitab software.

(Refer Slide Time: 05:52)



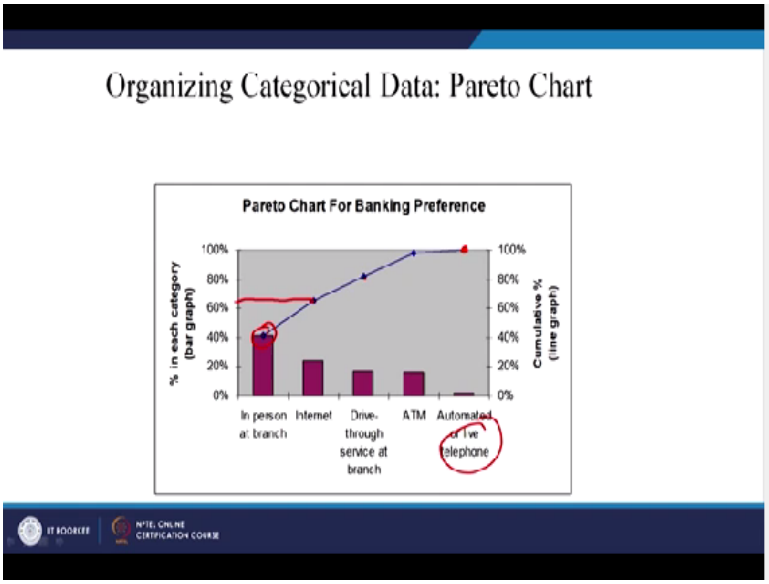
So if you wish, you can enter this data over here. This is, yes, so this is pie chart and of course, you can and there are several options available in this Minitab software where in you can add different values also. So those values would be shown over here. In fact, so if you select this particular column, the second column, then you would get all those values on pie chart. So we will move on to next slide, Pareto chart.

(Refer Slide Time: 06:36)



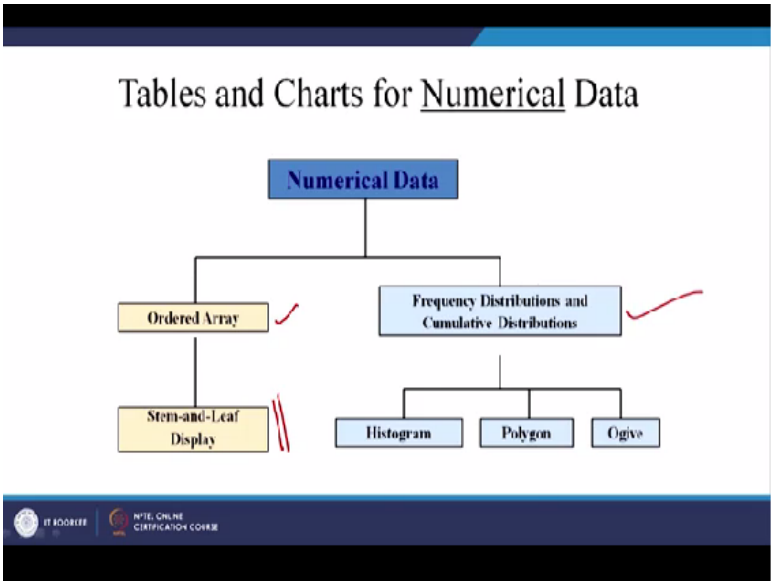
Pareto chart is basically a chart which is a vertical bar chart where categories are shown in decreasing order, right.

(Refer Slide Time: 06:48)



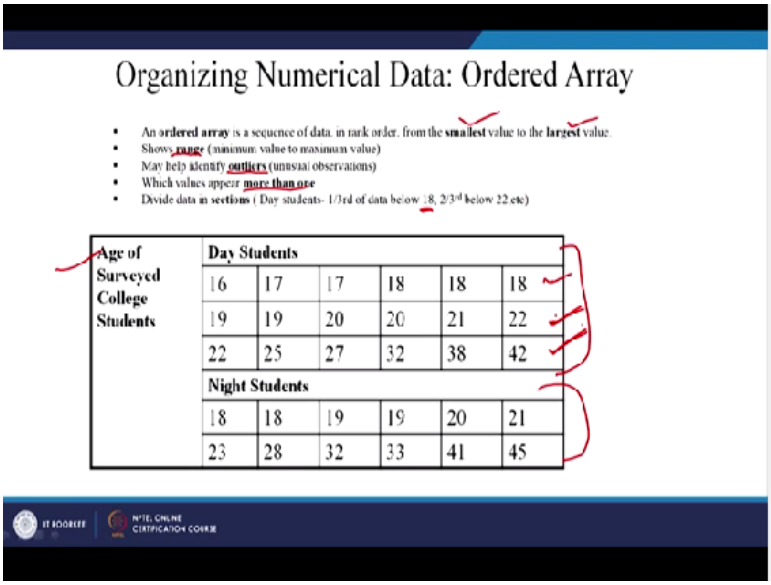
So this is Pareto chart. In this, 41% is in person at branch and 2% live telephone. So here, the point which is to be noted is these are cumulative values, right. So this is, at the end, this is 100%. So initially 41%, then internet was how much? We will go over to previous slide. Internet was? Let us look at here. Internet, we will look at in table itself, 24%. So 41+24, it becomes 65. Is it not? Just see, this is somewhere 65, right and you just add again the value of this particular category and so on, right. So this is Pareto chart.

(Refer Slide Time: 07:45)



Now we will move on to numerical data. So far as we have seen categorical data. Let us look at how to represent numerical data. So we have got 2 broad categories ordered array or you can call simply array, frequency distribution and cumulative frequency distribution. So we will look at Stem and Leaf diagram first of all. In fact, before I go on stem and leaf diagram, let me tell you why you should array, arrange data. There are different benefits of arranging data.

(Refer Slide Time: 08:30)



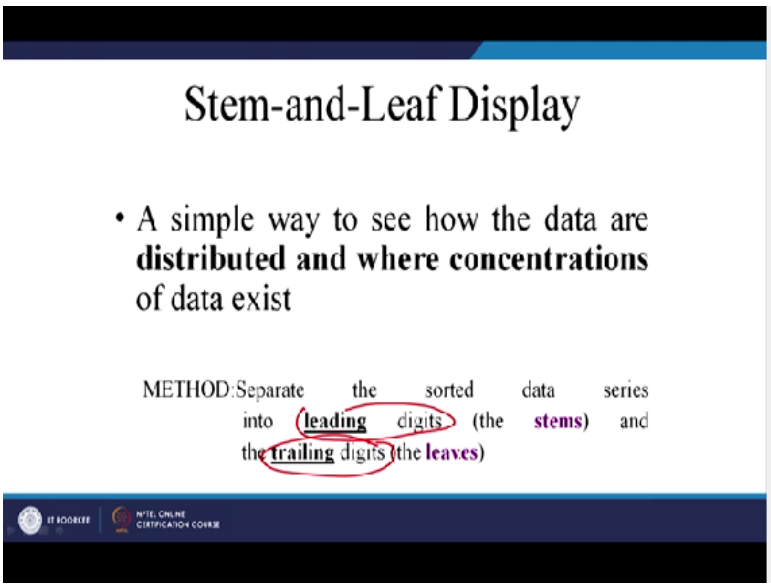
If you arrange data, then you will come to know the smallest value and largest value. Once you know these 2 values, you will also know the range of dataset. An array will help you in identifying outliers because whenever you collect data, there will always be some outliers. And we will discuss in next couple of classes about outliers. What are outliers? How to detect the

outliers and how to remove outliers?

We can also calculate the mode of the data. So we can count which items or which data points are coming more than once. And we can divide data into different sections. For example, I have got table over here in which I have surveyed the age of college students those who are attending classes in daytime and those who are attending night time, right. So day students and night students.

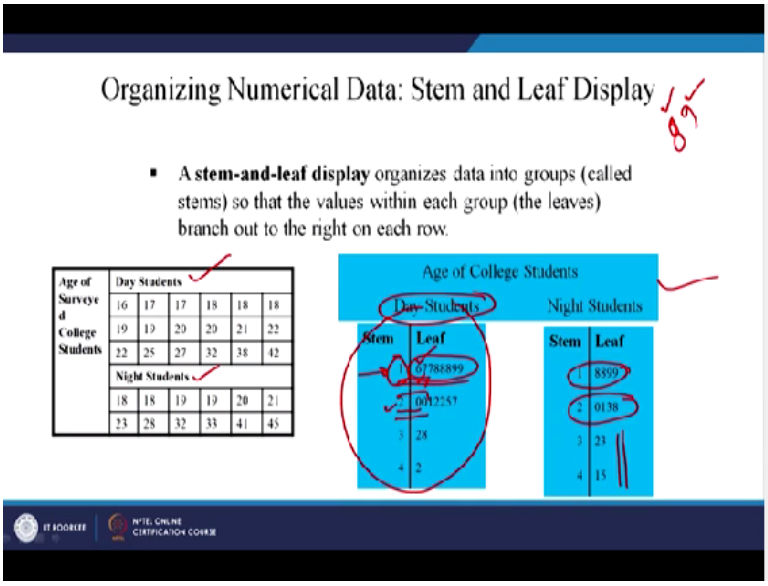
So if you look at, there are total 18 students in this case. So 6+6+6, right. So we can divide this table into different sections. So we will say that one-third of the students are below 18, right. So we have arranged this data in ascending order, right, in increasing order, right. So this is 16 17 17 18 18 and so on. We will say that two-third of students are below age of 22 and all the students are having age either 42 or less than 42. Similarly, you can have for night students as well, right.

(Refer Slide Time: 10:12)



Let us look at stem and leaf diagram or stem and leaf display. It is basically a very simple way which will help you in knowing where data are actually concentrated in a distribution. So it separates leading digits with trailing digits. So we call leading digits as stems and trailing digits as leafs.

(Refer Slide Time: 10:45)

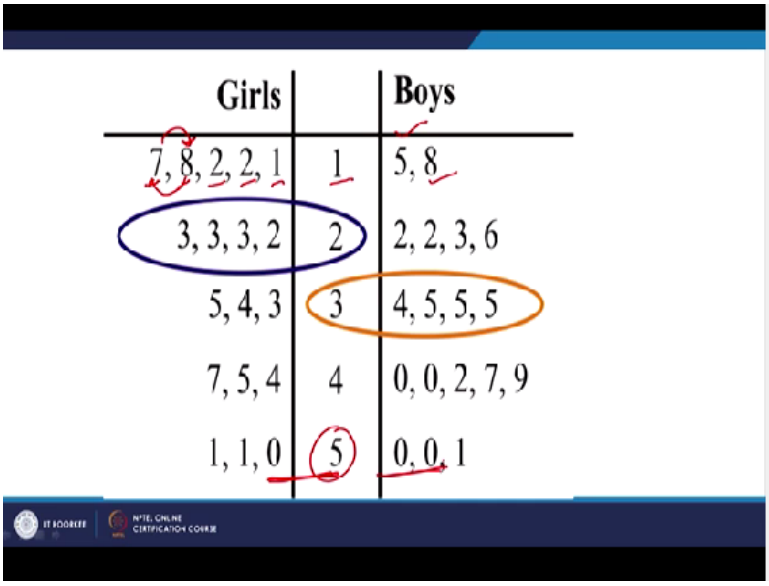


So let us look at the same example wherein we have used students of day time and evening time or night time, right. So if you look at stem and leaf diagram for day students. So in stem side, you have got, let us say 1, right. So the first student has got age as 16 right. So this is 1 and this is 6. The next 2 students have got age 7 and 7, so this is 7 and this is 7. Similarly, next 3, 8 and 8.

And next 2, 9 and 9. So this is your, one is stem and all these are leaves. Similarly, 2. How many students have got age 20? So you need not look at even this table. You just look at this first 20 and this is second 20, right. Is it not? Then 21, 22, 22, 25, 27. So if you look at the stem and leaf diagram, you will see that the data are concentrated at this. It means most of the students have got age 19 or less than 19.

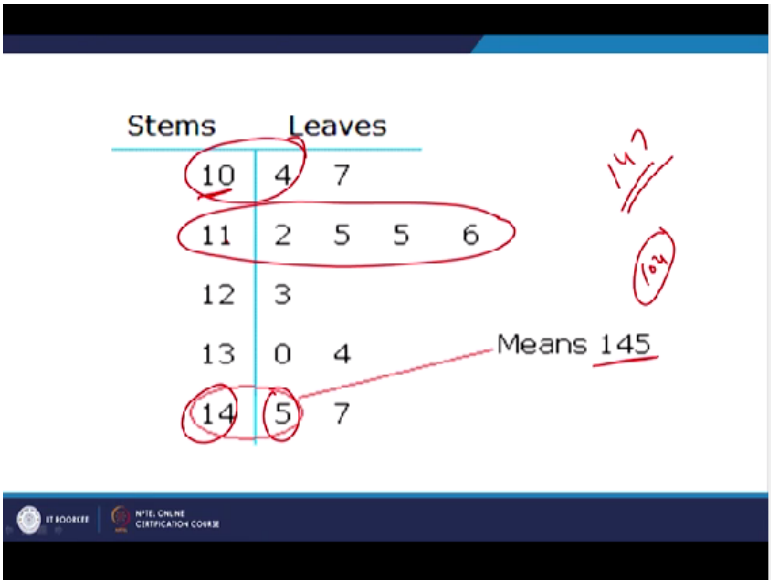
This is how you can easily know about the dataset. Similarly, for night students where there is concentration? You have got concentration at this and at this value, right. So we will say that the 5 students have got age 19 or less than that. And 5 have got age 28 and less than 28, right, but more than 20. So this is again you can draw stem and leaf for remaining students, right. So hope you would have understood what is a stem and leaf diagram, right. So let us say if the number is 89, right. So what is stem? 80 stem and this is leaf, right, okay.

(Refer Slide Time: 12:55)



You can represent, let us say, age of students, girls and boys in this manner as well. So let us say this 1, so the first girl has got age 11. Second 12, third 12, then 17 and then next 18, right. So it should be actually here, right. Is it not? As far as boys are concerned, first boy has got age 15, second 18, right. Similarly, for other values as well, right. So age is 50. So there is only 1 student of, there is only 1 girl who has got age 50 and 2 boys have got age 50, right.

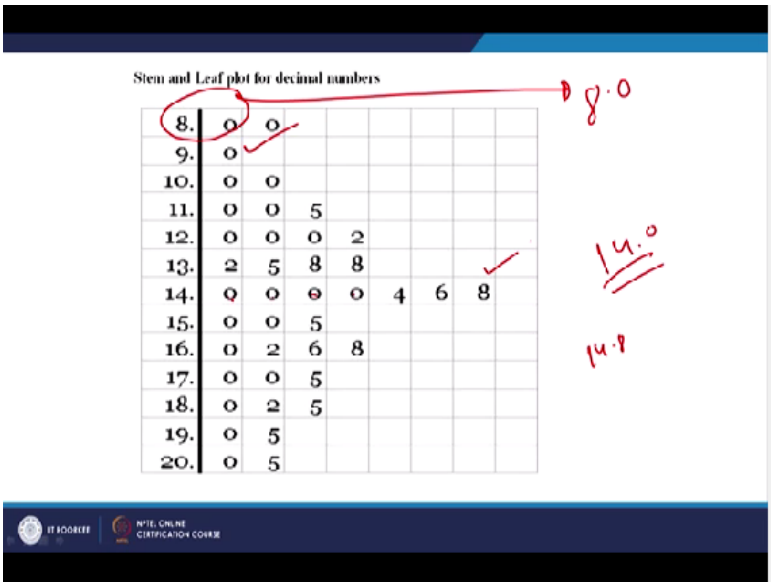
(Refer Slide Time: 13:47)



This is again stem and leaf diagram. Now in previous cases, we have seen stem and leaf diagram only for 2 digits, right. Now here we have got 3 digits. So let us say in stem part, you have got first 2 digits, right. So let us say 145 means 14 here and 5 is leaf, right. So this is, so where data again are concentrated? Here. Is not it? So this is 116, right. If I ask you which is the highest

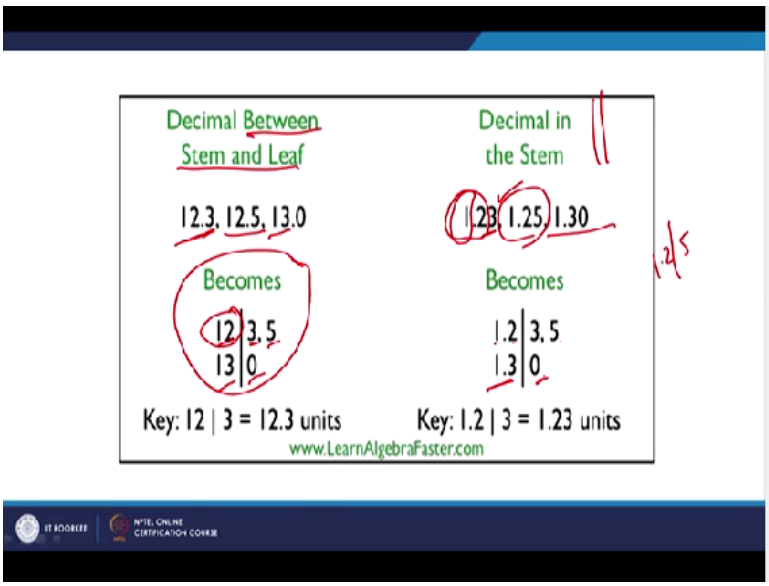
value in this, so that would be 147. Is not it? And the least value is 104. This is here. Is not it?

(Refer Slide Time: 14:34)



Now let us look at stem and leaf plot for decimal numbers if there is a decimal point. So this is 8.0, this is 8.0, right. So there are 2 data points having 8.0 value. So this is 9 value and this is 14.0, so how many 14.0? 1, 2, 3, 4, right. So 14.0 four times, right. And 14.8 is this, okay. So this is how you can represent stem and leaf plot for decimal numbers.

(Refer Slide Time: 15:14)

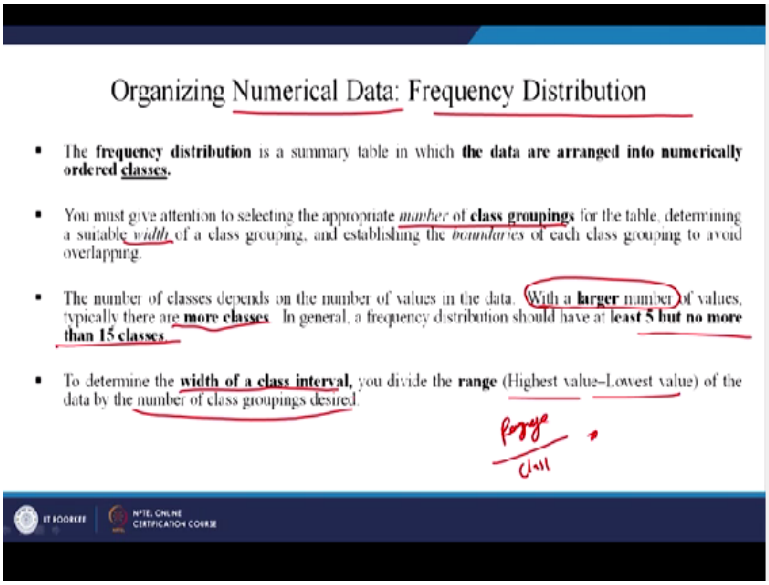


Now if there is a decimal between stem and leaf, then how to prepare stem and leaf diagram? So this is the stem and leaf diagram for decimal between stem and leaf. So let say number is 12.3. So this is stem, 12 and this is 3. This is 12.5. So 12 and 0.5, 13 and 13.0, right. So this is how

you represent stem and leaf diagram when there is decimal between stem and leaf.

If decimal is in stem, then how to represent? So this is your stem and these are your; this is your stem and this is your leaf, right. So 1.23 and this number is, 1.2 and then 5, right. So this is 1.25. 1.30 is 1.3 here and 0 this side. So this is how you can display data, right. Let us look at another method of displaying numerical data which is frequency distribution.

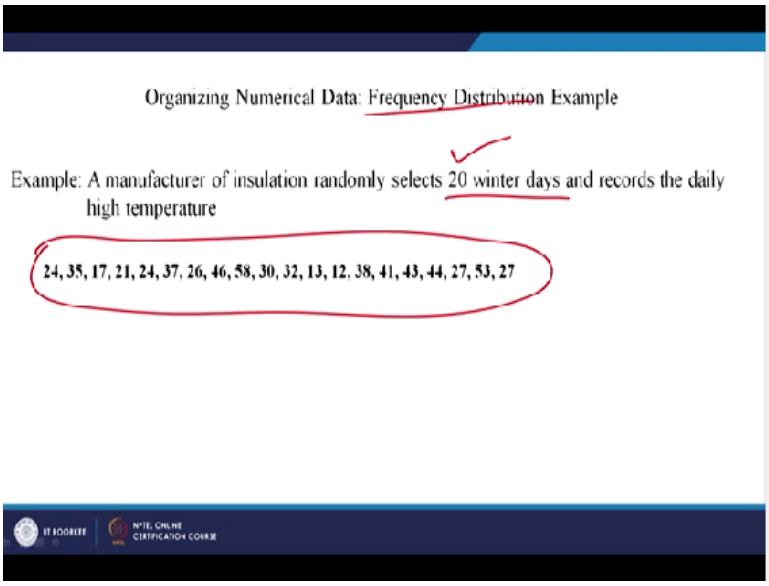
(Refer Slide Time: 16:32)



Now frequency distribution is quite an important type of distribution for data representation. So here we have got data. We arrange data into different classes and in each class, we will have different frequencies, right. So the most important point here is to which one should remember is how many classes will be there and what would be the width of each class. So these 2 important points you should keep in mind. So you can have different classes but the preferred range is 5 to 15 classes.

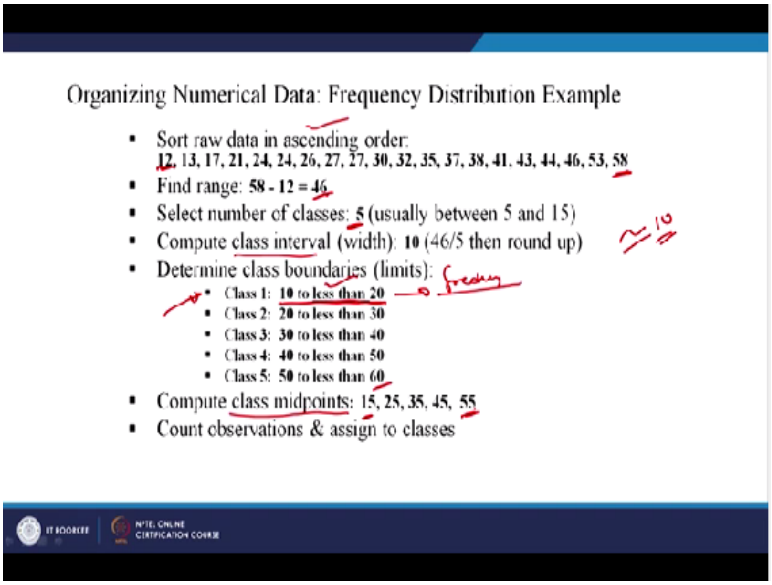
So whenever you have got dataset, either you can have 5 classes or 15 classes. So if the number of items or number of data points are very large, then you can have more classes, okay. So now the point here which you should remember is, how to find out width of the class? So for width of the class, you should have range/number of classes you want, right. So here range is highest to lowest value is range and you just divide by number of classes you are looking for in your frequency distribution. So let us take this example.

(Refer Slide Time: 18:02)



A manufacturer of insulation randomly selects 20 winter days and records the daily high temperature. So there is a dataset available here in which there are 20 data points, right. So these are different temperatures. Now we have to prepare frequency distribution.

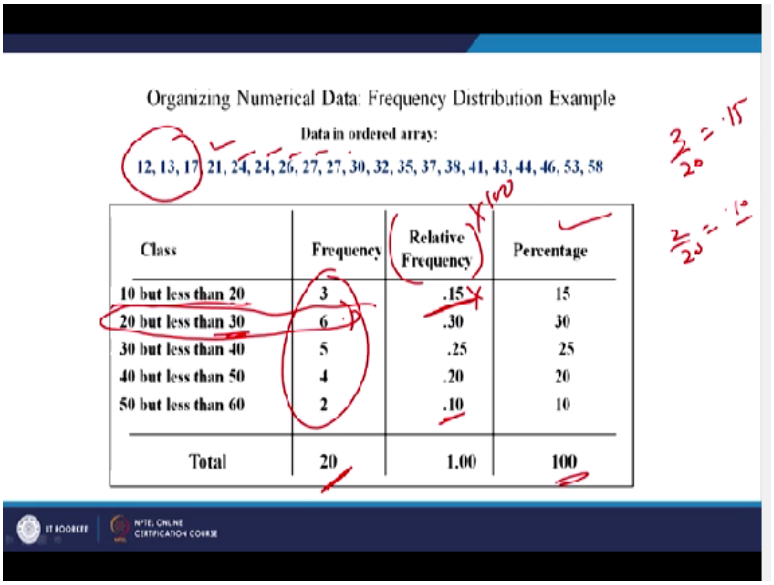
(Refer Slide Time: 18:22)



So first of all range data in ascending order. So the smallest value is 12 and the highest is 58. So this will help you in finding range, right. So the range is 46. Now you want to select number of classes. So let us select 5 classes for this particular example. Compute the class interval, so 46/5, this is approximately 10, right. So the width of the class is 10, right. So the first class is, or the class number 1 is 10 to less than 20.

So all data points which fall in this category, you can write here in frequency column. So that would be the frequency, right. Then 20 to 30, 30 to 40 and so on. So the last value is 60, right. And you can easily calculate the midpoints of all these classes, right. So for this first class, this is the middle point and for last class, this is the middle point. Count observation and assign to classes, right.

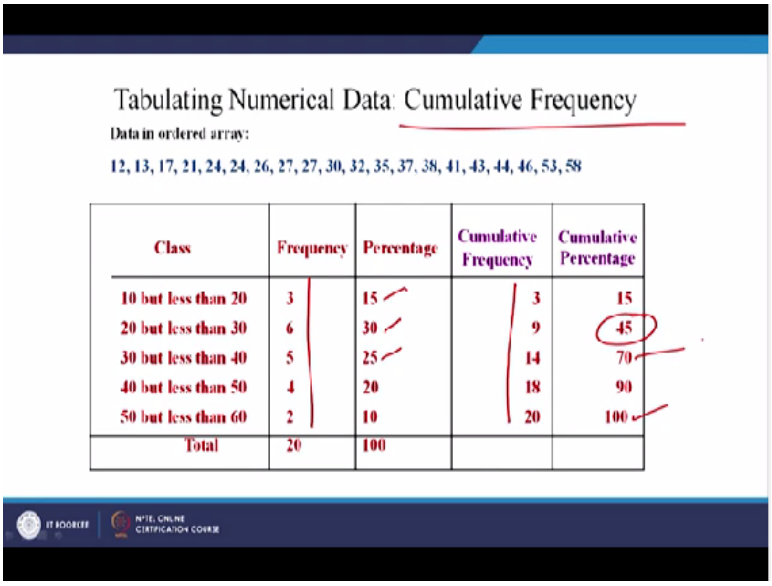
(Refer Slide Time: 19:39)



So this is how you can have different frequencies over here, right. So 10 or more than 10 but less than 20. So what are those values? These 3, right. Then 20 to 30, right, how many? But less than 30, right. So how many? 6 values, right. So 1, 2, 3, 4, 5, 6, okay. So this is how you can do it. Now once you are done with this frequency, this total is 20 because you got 20 data points. Now you can find out relative frequency. This 0.15.

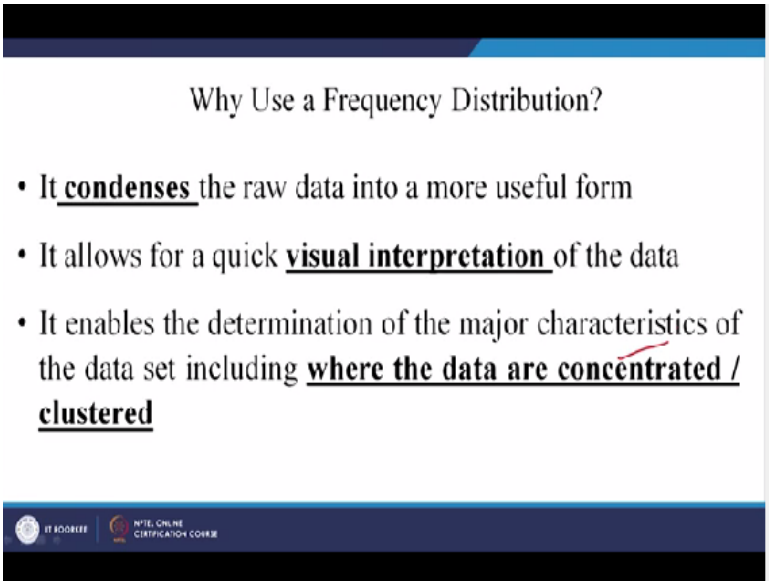
How did you get this 0.15, relative frequency? This 3/20 is 0.15, right. How did you get this 0.10? This is 2/20, right. So 0.10, right. Percentage, you just multiply this particular column by 100, you will get percentage, right. So this should be 100, right. So this is how you can draw a frequency distribution wherein you can have relative frequency percentage, you can also have cumulative percentage as well.

(Refer Slide Time: 21:03)



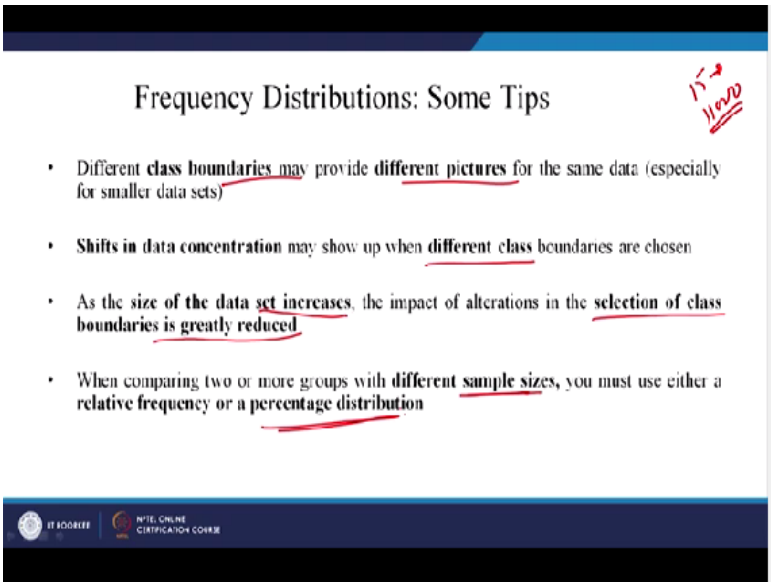
So let us look at cumulative frequency. So this is frequency, these are frequencies, these are cumulative frequencies, right, 3, 9, 14, 18 and 20, right. Similarly, cumulative percentage is 100%. So this is 15+30, 45, right. Then 45+35, 70, right and so on. So why to use frequency distribution? Is it better than earlier methods of data representation? Yes, it is surely better than earlier methods.

(Refer Slide Time: 21:39)



Because it condenses raw data into some more useful information. It is a kind of visual representation of data. So you can easily know where data are concentrated. It enables the determination of the major characteristics of dataset including where the data are concentrated as I have already talked about, right.

(Refer Slide Time: 22:04)

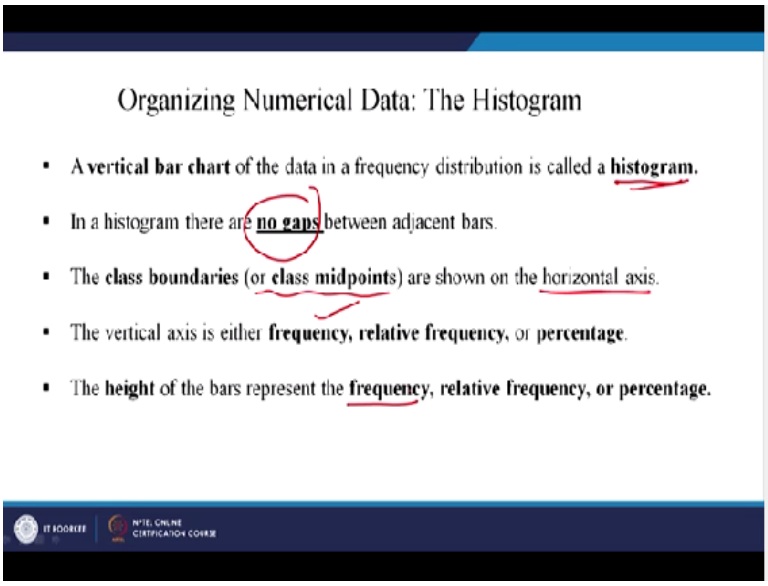


Some of the important points. So different class boundaries may provide different pictures. This is very important. So if you change, let us say in previous example, if you change number of classes from 5 to, let us say, 10. Then you will get some different frequency distribution. Now if you look at this, here the data are concentrated here, right, in this case. But if you change number of classes from 5 to 10, then data concentration would be somewhere else. Is not it?

So that is the point, you should remember. Shifting data concentration may show up with different class, yes data concentration would be shifted if you change number of classes. The size of the dataset increases, the impact of alteration in selection of classes, boundaries is greatly reduced. Now what happens if you have got, let us say, 10,000 data points, right.

Then and if you have got 15 classes, then there would be a class wherein some, wherein data would be concentrated. But let us say if there are, let us say, 11,000 data points and again there are 15 classes, then there will not be much difference in shift of the frequency. When comparing 2 or more groups with different sample size, it is good to use percentage distribution, right, or relative frequency distribution.

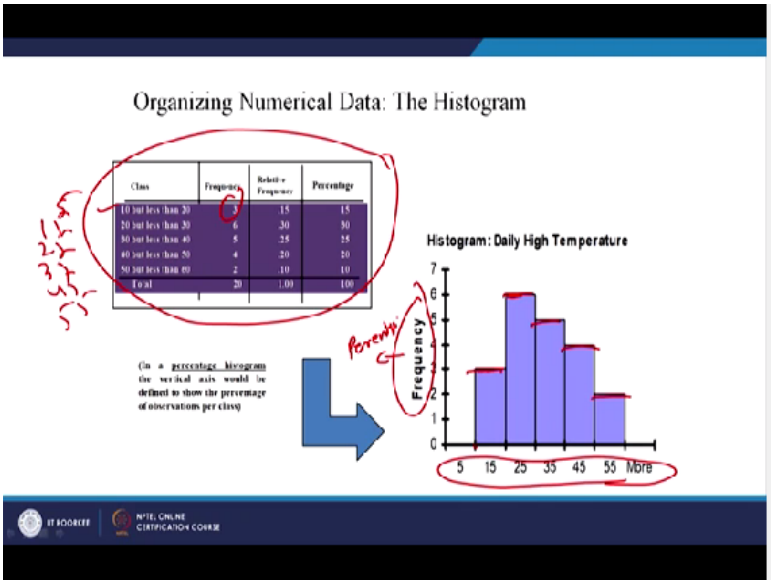
(Refer Slide Time: 23:43)



Let us look at another method, it is called histogram for numerical data presentation. A vertical bar chart of the data in frequency distribution is called histogram. So what is the difference between bar chart and histogram? Here there is no gaps between adjacent bars, right. So of course, you can have class boundaries and midpoints. They are shown on horizontal axis and on y axis, we always show frequency, right.

The height of the bar represents what? Frequency or cumulative frequency or percentage or cumulative percentage.

(Refer Slide Time: 24:27)

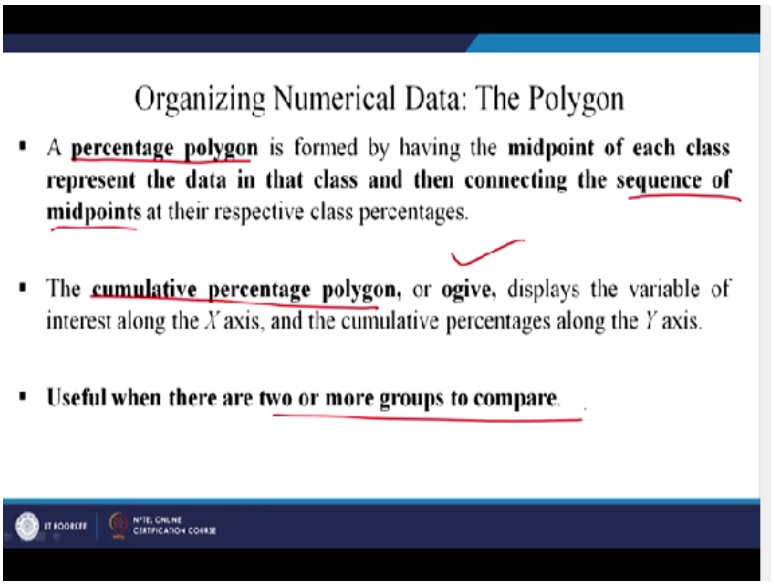


Let us look at how to draw histogram. So we have got this table already available. Now the

midpoints are 5, 5, 5, 15, 25, 35, 45 and 55. So all these midpoints are here, right. Is not it? 5, 15, all these are midpoints. And the first one is 3, right. So this frequency is 3, right. Next is 6, then 5, 4 and 2, right.

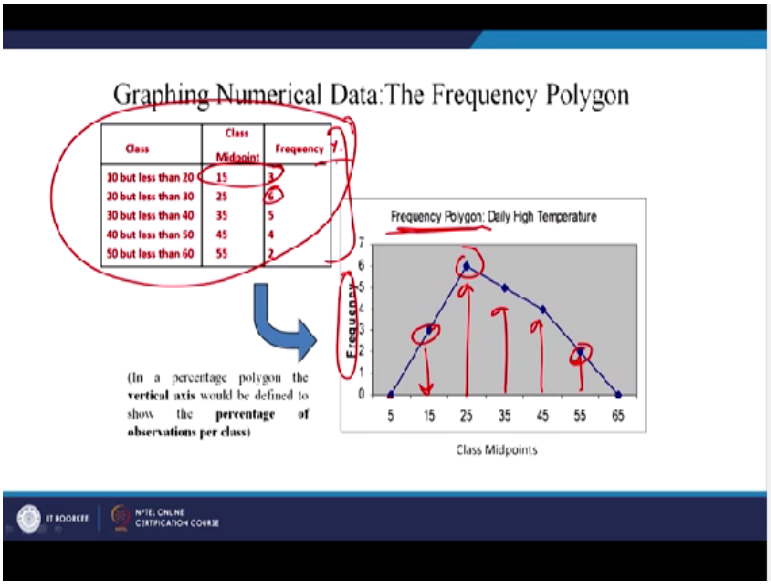
So this is nothing but a histogram, right in which we have shown frequencies. Now if you want, you can show percentage histogram also. So in case of percentage histogram, what will happen? Instead of frequency, you would be writing percentage, right, percentage. Is it not? So this is histogram.

(Refer Slide Time: 25:30)



Let us look at polygon, the last method for today's session. A percentage polygon is formed by having the midpoints of each class represent the data in that class and then connecting the sequence of midpoints at their respective class percentages. So you can also have percentage polygon or cumulative percentage polygon, also known as ogive, right. So we will see how to draw cumulative percentage polygon or percentage polygon. It is useful when there are 2 or more groups of data to be compared. So let us look at this.

(Refer Slide Time: 26:19)



So this is what we have already available with us. And this is your, these are your midpoints, right. So the first 5, 15, 25 and so on. So the first frequency at 15 is 3, right. So this is 3 at 15. Then next is 6. This is 6 at 25. Is it not? Then 5, then 4, then 4 and then 2, right. This is 2, okay. So this is how you can draw frequency polygon. This is called frequency polygon. How would you draw percentage polygon?

You just write percentage over here and calculate percentage in this and you know and in next column, you can calculate percentage, okay. So with this, let me summarize what we have done today. We have seen different methods of data presentation as far as categorical and numeric data are concerned. So we will look at some more methods of data presentation in next class. Thank you very much.