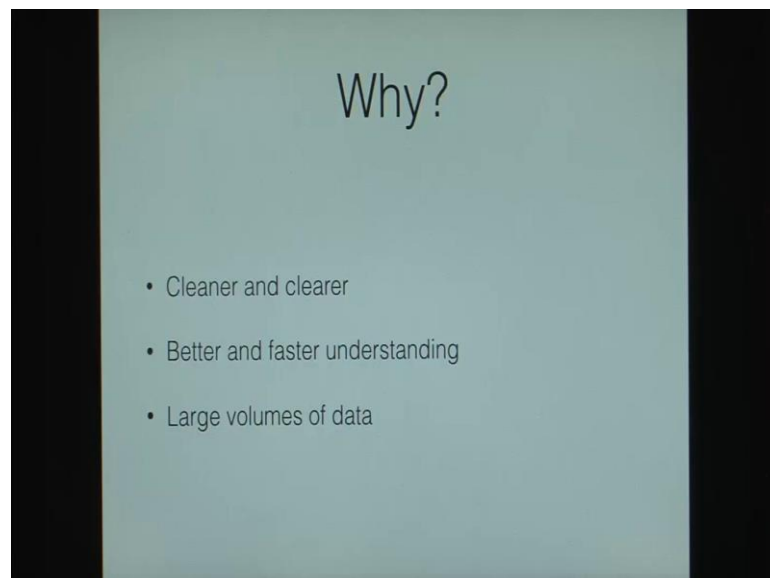


Forest Biometry
Prof. Mainak Das
Dr. Ankur Awadhiya
Department of Biological Sciences & Bioengineering & Design Programme
Indian Institute of Technology, Kanpur

Lecture – 05
Graphical presentation of data

[FL], we can calculate the measures of central tendency and dispersion, but it is also important to be able to graph these values to present our data graphically. Why is that so, well this is because graphical presentation is cleaner and it is clearer. It permits better understanding of data specially so when we are dealing with large volumes of data.

(Refer Slide Time: 00:33)



So, let us now look at some ways of representing data graphically.

(Refer Slide Time: 00:48)

Stem and leaf plots

Data: 11, 12, 9, 8, 15, 25, 21, 19

Suppose these are the heights of different saplings (in cm) in a nursery. We need to find out the most frequent / common heights.


Stem and leaf plots are a quick way of discerning this.

Let us look at this first example stem and leaf plots.

(Refer Slide Time: 00:58)

$x = (11, 12, 9, 8, 15, 25, 21, 19)$ 8, 25

Interval width \rightarrow 1, 10, 100, ... } 10
0.1, 0.01, ...



0		9	8			
1		1	2	5	9	\leftarrow 10 - 19
2		5	1			

36 / 53

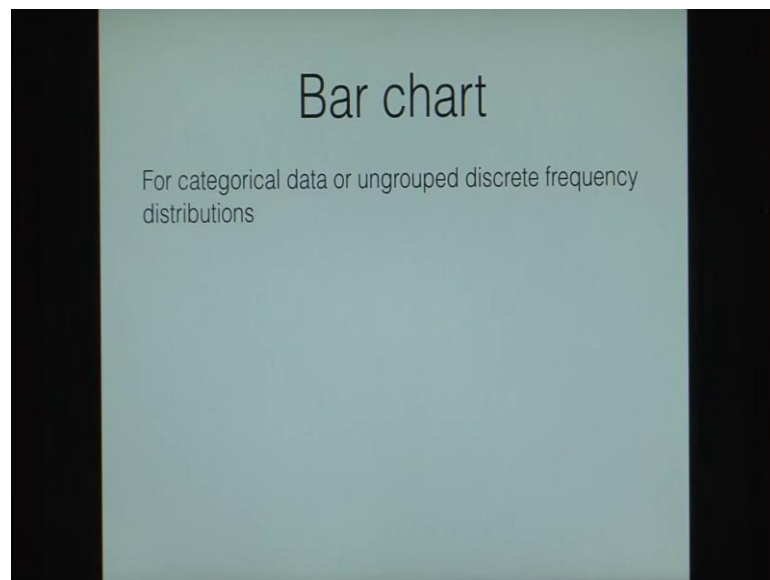
So, we have a set of data that is given by x equals 11, 12, 9, 8, 15, 25, 21 and 19. Suppose these were the heights of different saplings in centimetres in a nursery. We need to find out the most frequent or the common height. Stem and leaf plots are a quick way of discerning this. So, how do we make a stem and leaf plot out of this given data. We will first of all we need to define an interval width. Most commonly, we take an interval width that is a multiple of 10 or a fraction of 10. So, may maybe we could take 1, 10, 100

or so on or we could even take 0.1, 0.01 or so on for this particular distribution an interval width of ten looks as the most appealing one.

Now, that we have decided on an interval width, we divide this data, we take the smallest value which is 8 and we take the largest value that is 25. Now, 8 would come under a stem and leaf plot of 0, so 8 can be written as 08. We will have another stem that is one and another stem that is 2. So, now, let us take these values one by one and start adding leaves to these.

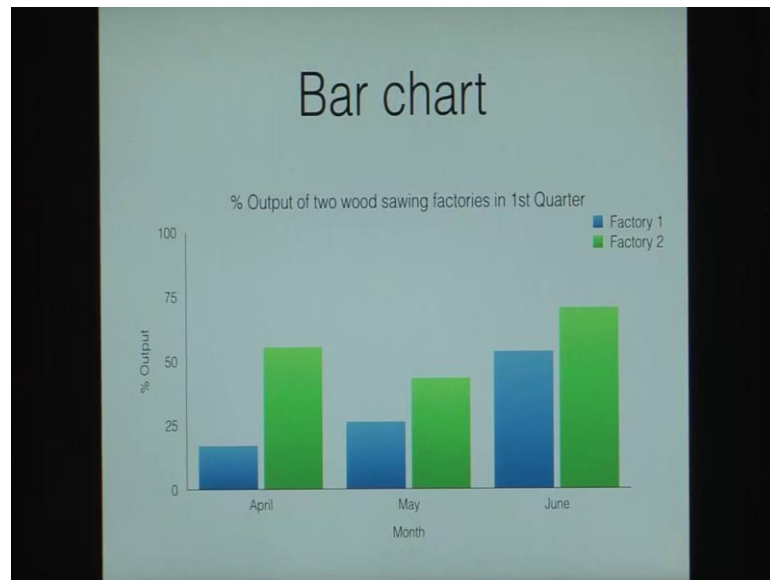
So, essentially what does stem and leaf plot is that you have a stem; on that stem you have different leaves. So, this is how we are going to make this plot. So, consider the first value that is 11. So, 11 come under the stem of one. So, we write 11 like this. Next value is 12, so we write it as 2. Next value is 9 that is 09. Next we have 8 that is 08, next we have 15 that is 1 5, then we have 25 then we have 21 and lastly we have 19. So, this would be a stem and leaf plot. So, this speaks it very easy to understand that most of the values lie in this interval width of 10 to 19, this is how we do it, the most common height is 10 to 19 centimetres.

(Refer Slide Time: 03:44)



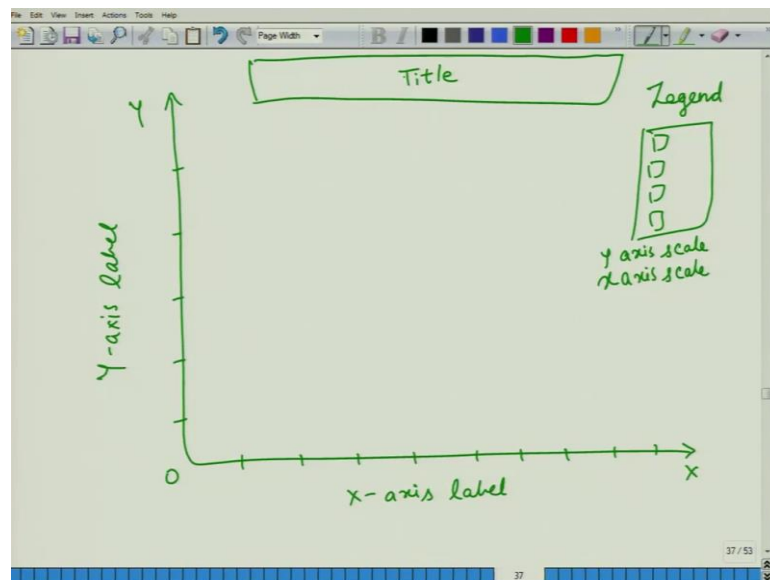
Next have we have a bar chart a bar chart is useful for both categorical data and for ungrouped discrete frequency distributions.

(Refer Slide Time: 04:00)



So, let us have a look at a bar chart. This bar chart on the example depicts the percentage output of two good sawing factories in the first quarter.

(Refer Slide Time: 04:15)

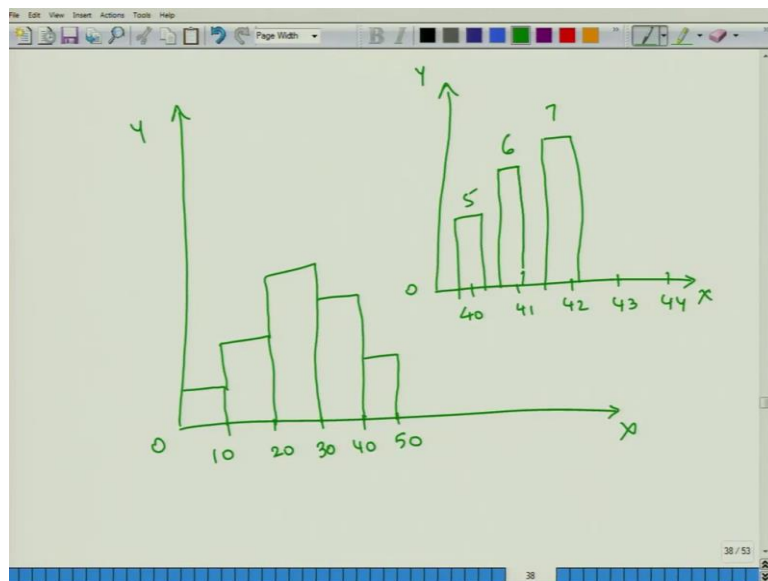


So, every chart needs to have the axis that are labelled. Then it will have some values on the y-axis and the x-axis, then it will have a label a y-axis label and an x-axis label and any then it will have a legend it will show what all values refer to and then it will have a title on top which tells us what this chart is all about. If possible we also add scales, so

for instance here you can add x-axis scale and a y-axis scale; this is especially useful if you are making a plot on a graph sheet.

Now, going back to our bar chart on the bar chart, here we can see that it shows that in the first quarter that is the months of April, May and June we have two factories factory one as seen in the legend on the top right. So, the factory one is represented in blue colour; the factory two is represented in green colour. So, in the case of factory one, we can see that the percentage output goes on increasing with the months. So, the output of May is greater than that of April and the output of June is greater than that of May, but in the case of factory two, the output decreases from April to May and then increases again from May to June. So, a bar chart makes it very easy to interpret these data, it would be extremely difficult if we showed these data in the form of tables only.

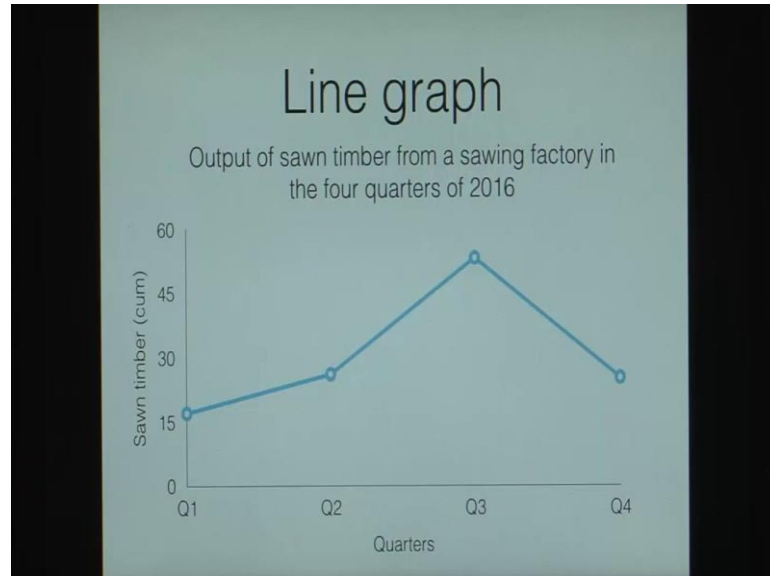
(Refer Slide Time: 06:35)



The next graph is called a histogram it is used for continuous variables, it is similar to a bar chart, but the bars are joined together. So, for instance suppose we had on the x-axis suppose we have distribution that go like this 0 to 10, 10 to 20, 20 to 30, 30 to 40, 40 to 50 and so on. And suppose these values were like this. So, here the values are joined together. On the other hand, if we had a discrete distribution say number of children in different classes number of pupils in different classes suppose we had the values like 40, 41, 42, 43, 44. And suppose each class had a different values, suppose this was grade 5, grade 6, grade 7 and so on. So, in this case, we would not be able to join these bars

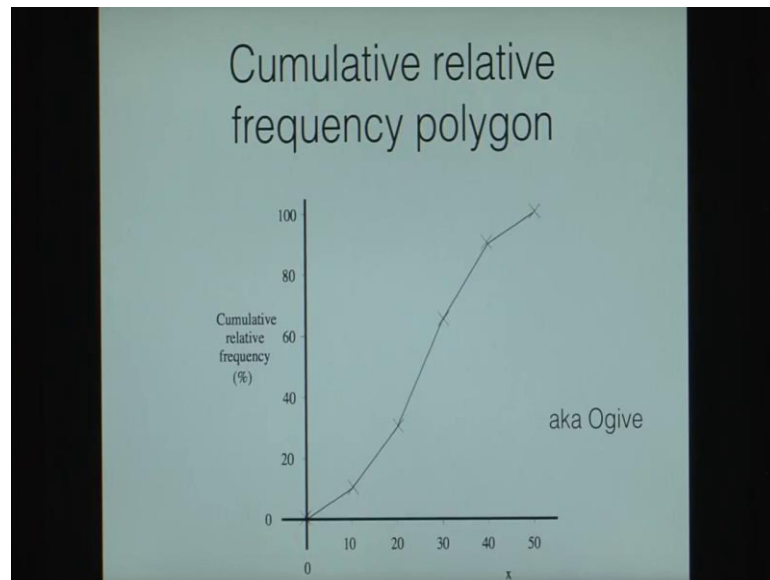
together, but in the case of a histogram, we join these bars together because it is a continuous variable.

(Refer Slide Time: 08:02)



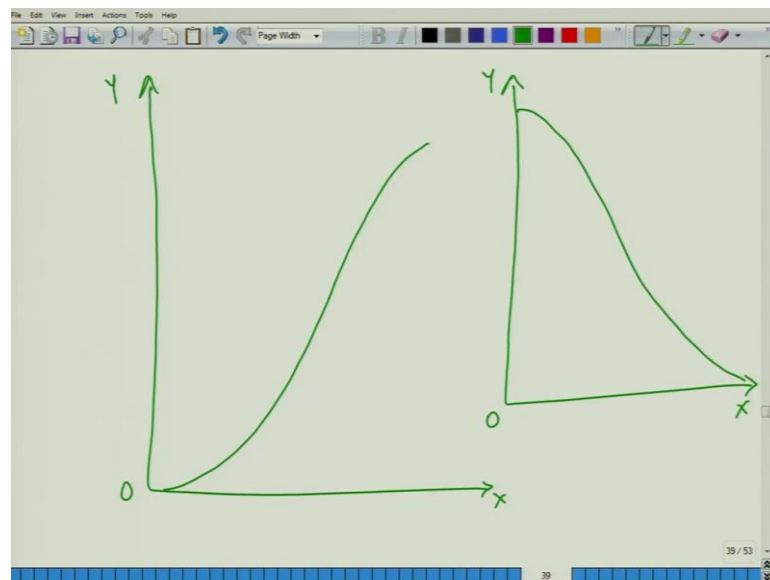
Another graph that we routinely use is a line graph. So, in this example, on the screen, you can see a line graph that represents the output of sawn timber from a sawing factory in the four quarters of 2016. Now, in all the cases, we are not required to make bars on the graphs, but we can always take a central value and then join them together by straight lines to make a line graph. So, as we can see in this graph the output increases from quarter 1 to quarter 2, then increases again from quarter 2 to quarter 3 reaching a maximum and then falls from quarter 3 to quarter 4. So, this is a line graph.

(Refer Slide Time: 08:43)



A similar graph is a cumulative relative frequency polygon also known as an ogive. So, an Ogive can go on increasing or decreasing as you move towards the right.

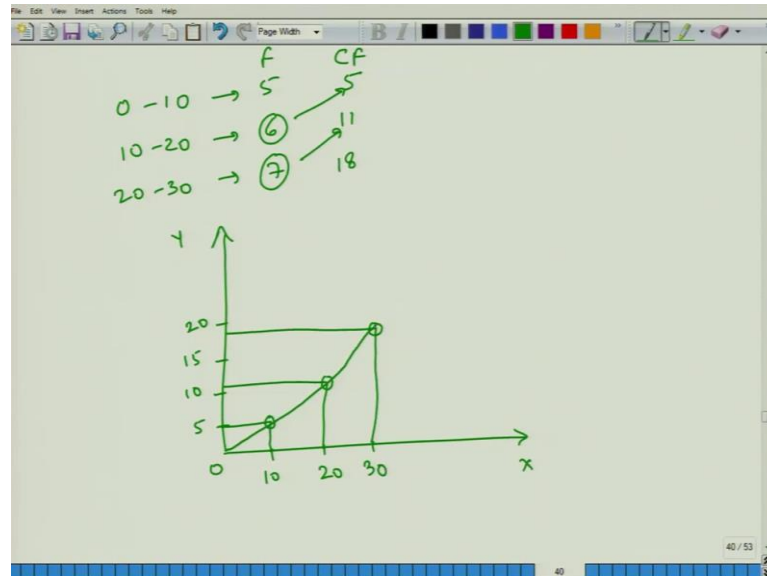
(Refer Slide Time: 09:00)



So, essentially you can have an Ogive that goes like this. or you can have another Ogive that goes like this. So, in one case the values go on increasing as we go towards the right; whereas, in the second case the values go on decreasing as you go towards the right. So, if we looked at the chart on the screen, we will see that if you look at the value the x

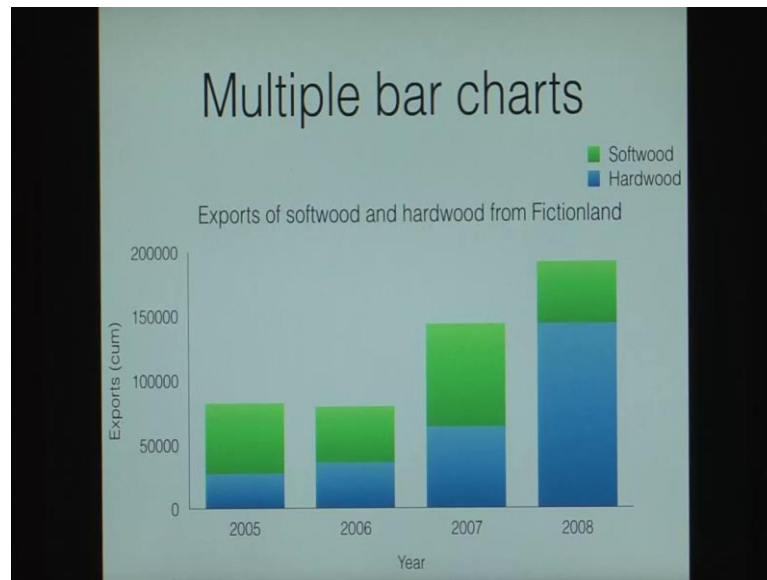
value of 40 we will get a y value of somewhere around 90. So, this means that 90 percent of all the values are less than 40, and only 10 percent values are greater than 40.

(Refer Slide Time: 09:49)



So, this represents a cumulative distribution. So, essentially if we had a distribution of say 0 to 10, 10 to 20, 20 to 30. And suppose these had the values of say 5, 6 and 7. To plot an Ogive, you will make a cumulative frequency. So, this is the frequency and the cumulative frequency this becomes 5, the next value becomes 6 plus 5 that is 11, the next value becomes 7 plus 11 that is 18. So, now, if we plotted, so 10, 20, 30 and suppose this was see 5, 10, 15, 20. So, what are the values that are less than ten we have 5 values that are less than 10. So, we take this point. Then how many values are less than 20; we have 11 values that are less than 20. And how many values are less than 30 we have 18 values that are less than or equal to 30, when we join all these points together we get the Ogive.

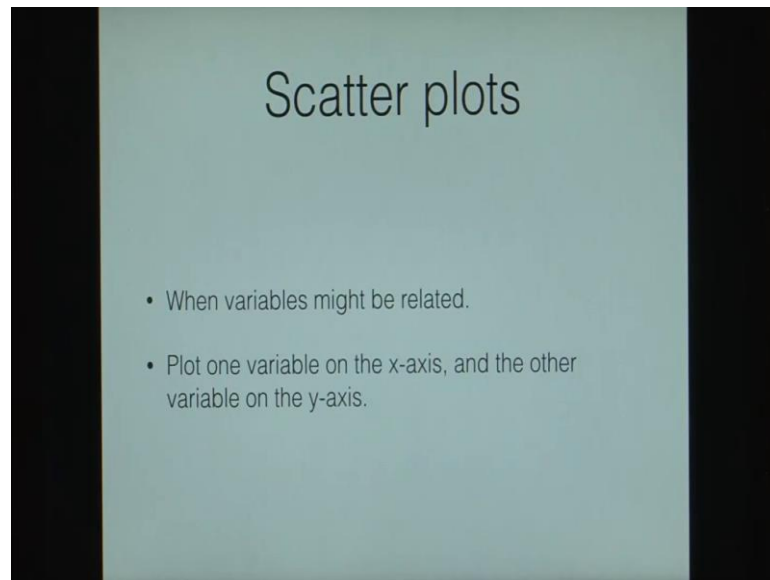
(Refer Slide Time: 11:14)



Next we have multiple bar charts. Now, multiple bar charts are deployed if we wish to show different kinds of data. The multiple bar chart that you see on the screen depicts the exports of softwood and hardwood from a fictional country called fiction land in various years. So, we can see on the x-axis, we have the years 2005, 6, 7 and 8; on the y-axis we have the exports in cubic meters the softwood is represented in green colour and the hardwood is represented in blue colour. We can make out that while the total exports have been increasing.

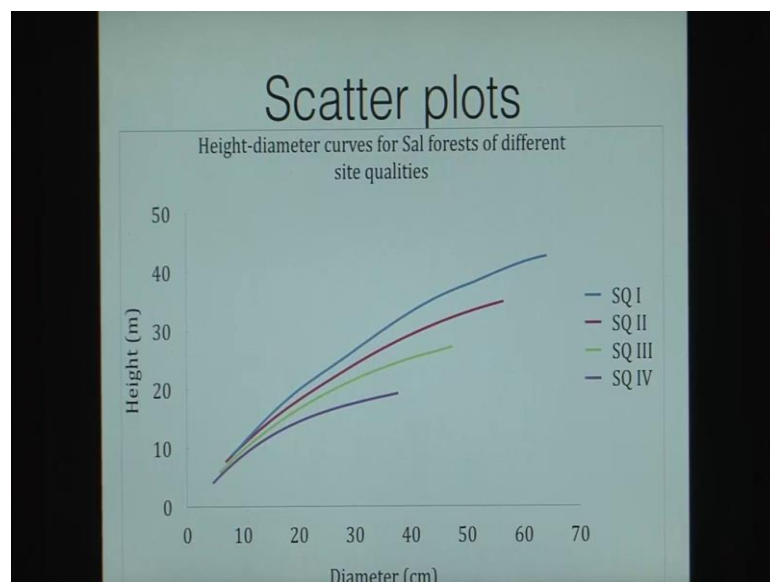
Now how do we make out these total exports by looking at the cumulative length of the bars with the green and the blues combined. So, total exports have been increasing and the exports of hardwood have also been showing an increasing trend if you looked only at the blue portion. However, we can make out from this graph that the exports of softwood are showing a mixed picture because between 2005 and 6 the exports decreased they increase again if we look only at the green portion, the size of the exports increase again between 2006 and 7 and then they decreased again. So, a multiple bar charts help us to show different kinds of data on the same plot.

(Refer Slide Time: 12:39)



Now, when the variables might be related, we use scatter plots. So, in the case of a scatter plot, we plot one variable on the x-axis and the other variable on the y-axis if there is a relation we will observe a trend.

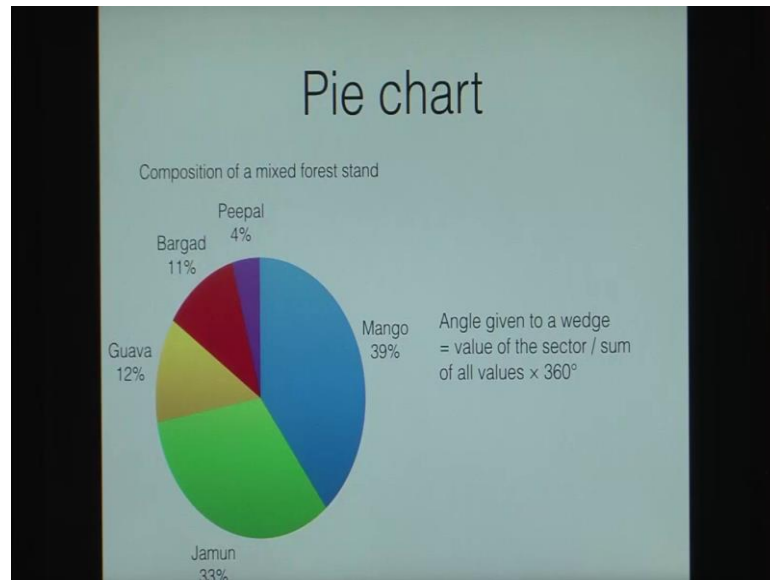
(Refer Slide Time: 12:56)



So, let us look at some scatter plots. Now, this is scatter plot shows height that is represented on the y-axis and the diameter of trees diameter is represented on the x-axis in the case of Sal forests of different site qualities. Here we observe that as diameter increases, the height also increases which tells us that both these variables are related to

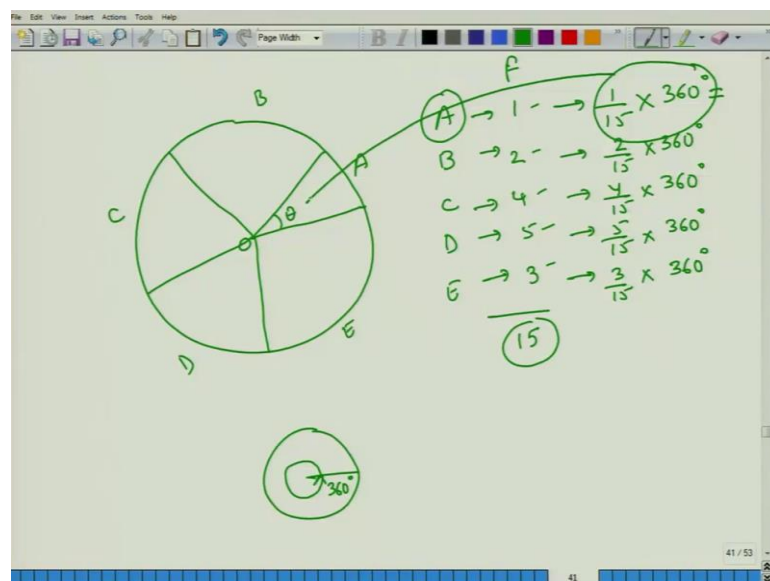
each other and the common variable that governs both of these is the age. So, as the age of a tree increases its diameter increases; at the same time its height also increases. So, if you plot diameter on the x-axis and height on the y-axis, you will see an increasing trend. So, these are related.

(Refer Slide Time: 13:43)



We could also use a pie chart for visual depiction of data. So, for instance this pie chart reveals the composition of a mixed forest stand.

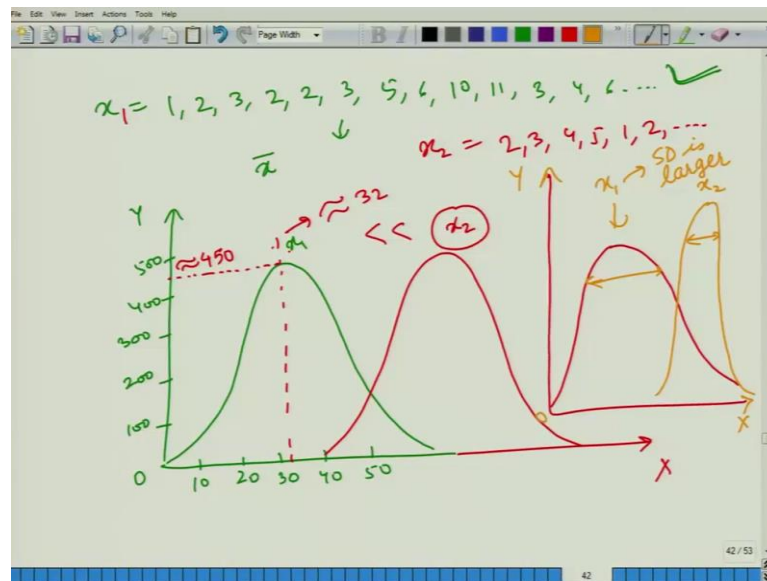
(Refer Slide Time: 13:56)



So, how do we make a pie chart? So, a pie chart consists of a circle and with a centre O, and it is divided into a number of wedges. So, suppose we had five different data values data value a b c d and e and the values were say 1, 2, 4, 5 and say 3. So, what is the sum of all these values? So, these are the frequencies in which these data appear what is the sum of it? So, you have 5 3 of 15. So, to find out the angle that is subtended by each of these a b c d and e on the pie chart we divide the frequency by this total 15. So, you have 1 by 15, 2 by 15, 4 by 15, 5 by 15 and 3 by 15.

Now, remember when you have a circle and if you have the radius that goes all around it makes 360 degrees. So, if we wish to figure out, how much is the angle subtended by any one value say A, we multiply these values by 360 degrees. And the value that we receive out of this would be the angle that is subtended by A in this case. So, here we observe that every tree species has been given its own wedge; and the wedge angle is calculated by the value of the sector divided by the sum of all values, and then multiplied by 360 degrees. So, the total in this case, we will always come out to be 360 degrees.

(Refer Slide Time: 16:19)



Now, representation of data in the form of charts is important because suppose I gave you a distribution of x equal to 1, 2, 3, 2, 2, 3, 5, 6, 10, 11, 3, 4, 6 and so on. Suppose, I gave you a large distribution and if I asked you to find out a measure of central tendency suppose I asked you to find out what is the mean in this case. It would be extremely difficult, because it will be a computationally intensive. On the other hand, suppose I

gave you this distribution and we have these values of 20, 30, 40, 50 and suppose this is how your curve looked like on the y-axis also we have see 100, 200, 300, 400, 500.

Now, if I ask you what the mean was in this particular case in this case it would be computationally intensive to calculate the mean, but in this case you would be very easily saying that the mean is close to this value. So, the mean would be close to around 32 and if I ask you how many values of y corresponded to this value of x you would very easy tell me that it would be somewhere around 450. So, this makes understanding of these data very intuitive. On the other hand, suppose I give you two distributions. So, this one was x_1 , now if I have x_2 that is again to give some random numbers 2, 3, 4, 5, 1, 2 and so on.

And if I asked you which of these two data sets is larger and which one is smaller again it would be difficult. But suppose we plotted x_1 , so this is x_1 and suppose we have another distribution, let us make it in different colour suppose this is x_2 . So, in this case if i asked you which one is a which data set has values that are larger than the other you would be very easy telling me that x_2 is greater than x_1 . In fact, it is being greater than x_1 .

Similarly, if I ask you what is the spread of these data, so if I ask you the range or the range coefficient or say the standard deviation of x_1 or x_2 , it would be difficult to tell intuitively. But suppose if we plotted these and we had this is x_1 and say this was x_2 , you would be very easily telling me that in the case of x_1 the spread is greater. So, say the standard deviation is greater as compared to in the case of x_2 . So, even though we are not getting to the exact values of the measures of central tendency or the measures of dispersion we can very intuitively still something about this data which has a larger mean median or mode and which one has a larger or a smaller dispersion. So, this makes a representation of data in the form of charts very important in our understanding.

Thank you for your attention, [FL].