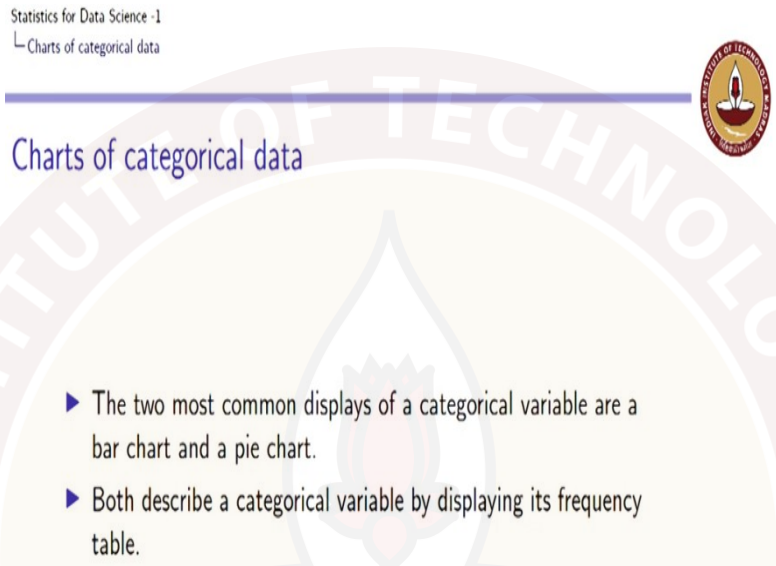
IIT Madras

ONLINE DEGREE

Statistics for Data Science - 1
Prof. Usha Mohan
Department of Management Studies
Indian Institute of Technology, Madras
Lecture No. 2.2
Describing Categorical Data – Charts of categorical data

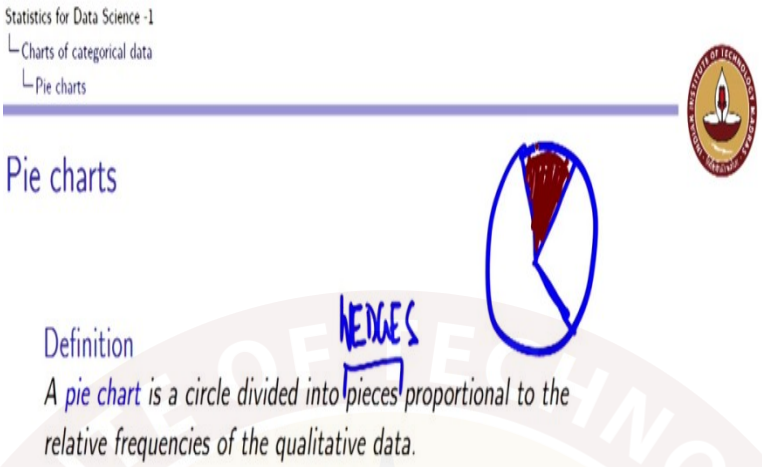
(Refer Slide Time: 00:14)



Now when we come to graphical displays of categorical data what we have learned so far is how to come up with a frequency table and we just also demonstrated how to come up with a relative frequency table. So, once I have the tabular form of summarizing my data and remember we are only working with categorical data at this point of time, and I am looking at only one variable. The next common thing is how do I graphically display this data.

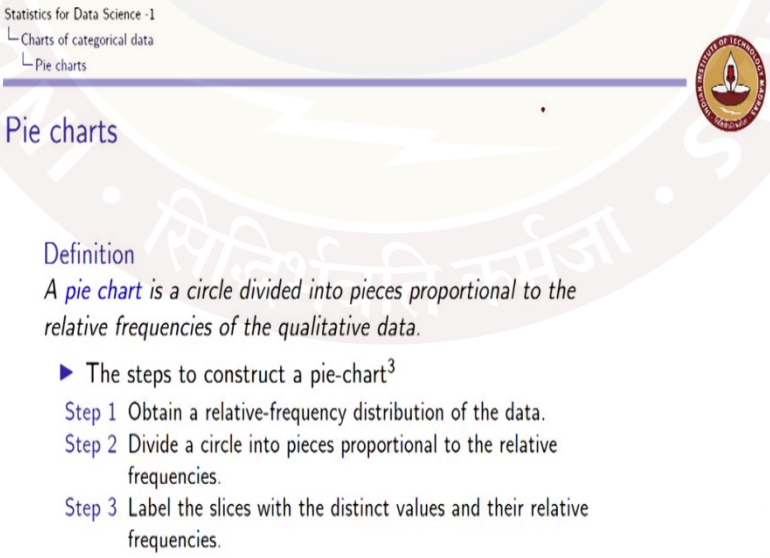
So, when it comes to categorical data, the two most common displays of a categorical data are a bar chart and a pie chart. Since we introduced already a frequency table, we also see that this both the pie chart and the bar chart basically display the data that is given in the frequency table. What do we mean by that?

(Refer Slide Time: 01:14)



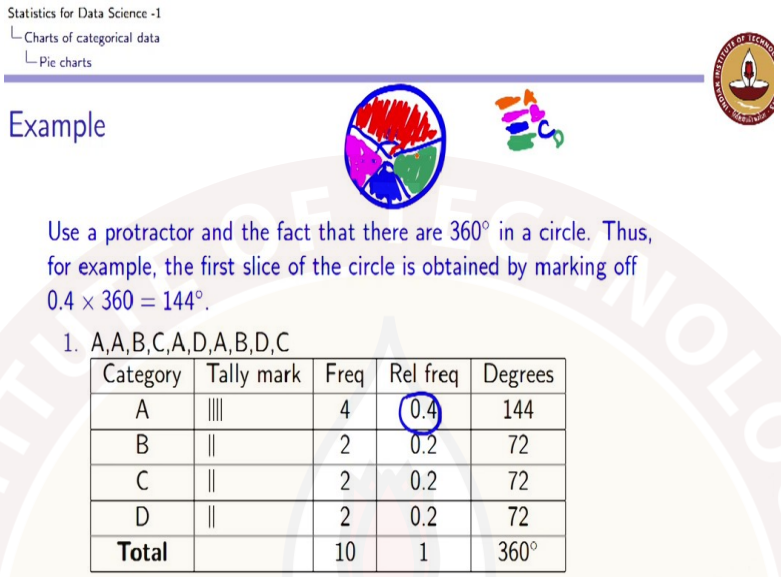
Recall that a pie chart. So, I can define a pie chart as a circle that is divided into pieces or wedges. Some textbooks refer to this as wedges. The reason why it is called a wedge is a pie chart is a circle or a disc which is divided into pieces or wedges. Now this portion I have here is a wedge, this particular portion is a wedge. Now again, another reason why it is called a pie chart is if this is a pie, this shaded portion is basically a share of a pie or a piece of a pie.

(Refer Slide Time: 2:10)



So, a pie chart, how do we construct a pie chart? Again, go back, we know how to construct a relative frequency table.

(Refer Slide Time: 02:19)

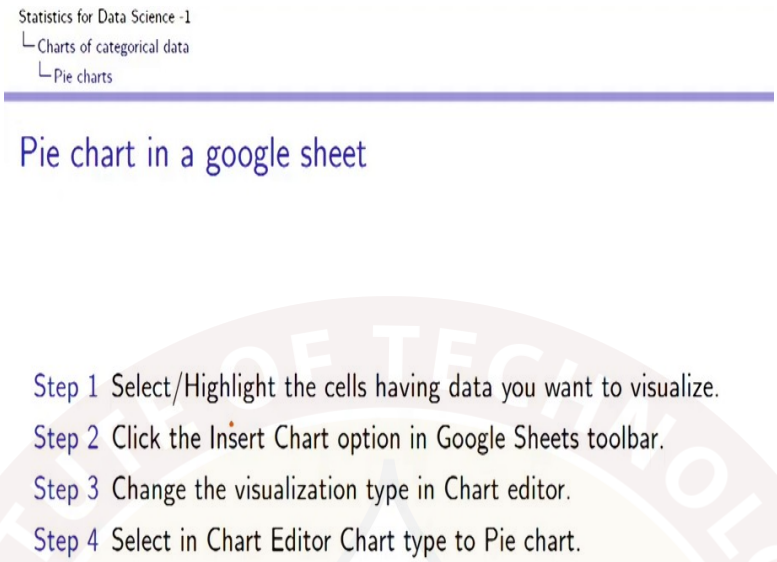


How do we obtain a pie chart? So you can see that I start by drawing a circle. Now, for example, the relative frequency of A is 0.4, I multiply that with the total number of degrees in a circle, which is 360. And I have a 144. This is 72. So, if I start from this, approximately this is somehow this is where I am going to have and this angle. I am not doing it exactly, but I know that this is going to be 144°.

Now the next thing I have is B which is at 72, C it is 72 this is 72°, this is 72°, and this is 72°. So, you can see what you can do is a way good way to have a pie chart is I can shade it with different colors, where color green represents the category D, color blue represents the category C, color purple represents the category B, and color orange represents the category A.

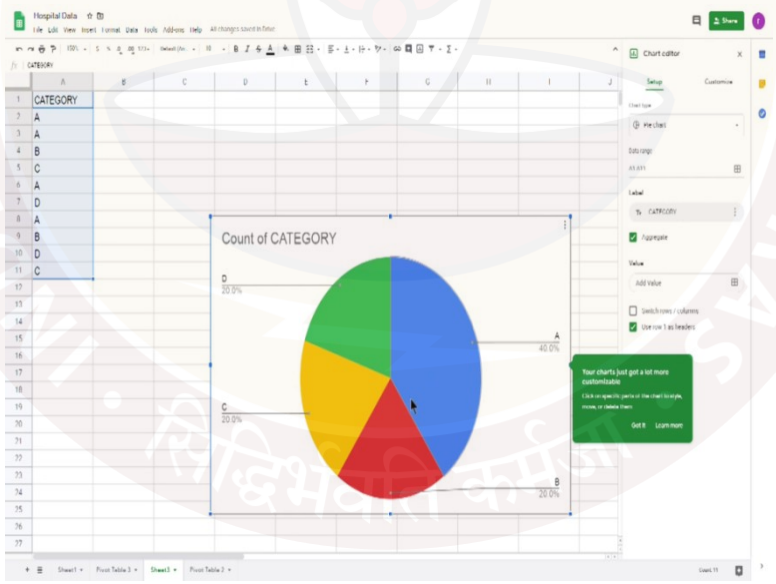
So, in a sense, what a pie chart gives us is, it gives us the share of a pie. In other sense, you can say that 40% of my data, which is the share of this pie is category A, the rest of them you can see the purple, the blue, and the green are almost same. They are same, in fact, they are same share, which is 20% each and that gives a share of a pie. So, whenever I want to actually show to my audience or to I want to display the share of a particular category, then I a pie chart is the most appropriate graphical display.

(Refer Slide Time: 05:11)



So you can see that for this, I can do a pie chart in Google Sheet as well.

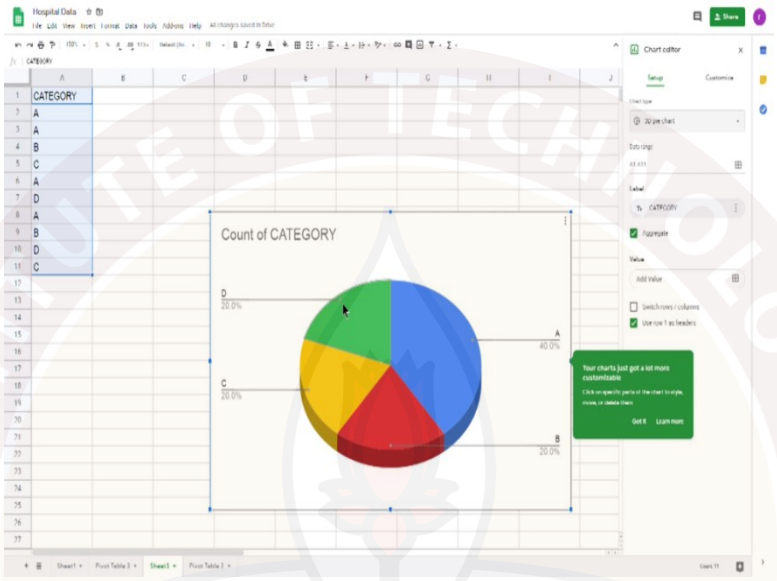
(Refer Slide Time: 05:16)



So, let us go and do a pie chart for the same data. So I have this as my data, I have this as my relative frequency. So, I go back to my data, which I have calculated here. So, you can see that, how do I do a pie chart again, I highlight the data I want to visualize, what is the data I want to visualize, I want to visualize this data. This is the hypothetical data I have created. I go to Insert Chart option.

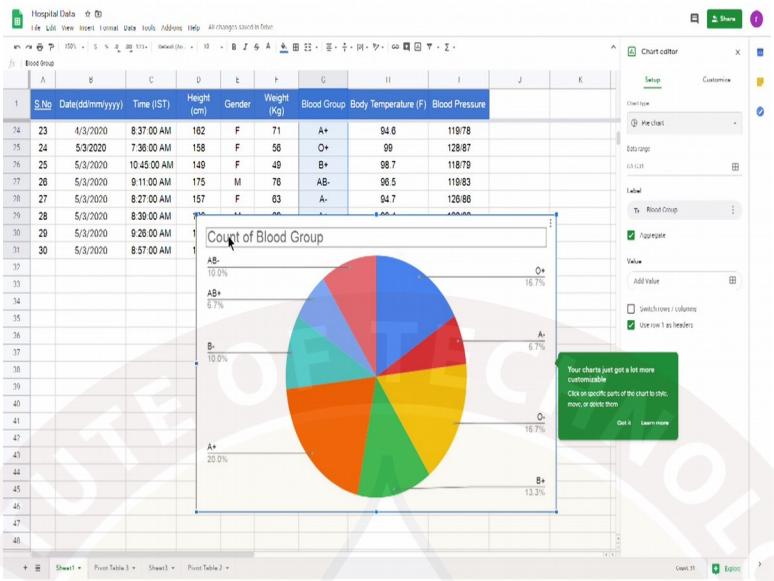
So, you have an insert option here. I have a chart option here. And you can see that this is precisely what we had earlier. A is 40%, B is 20%, C is 20%, and D is 20%. So, you can see that this is how my pie chart, you have the chart editor, within the chart editor, you can actually tell what is the kind of pie chart you want.

(Refer Slide Time: 06:16)



If you want a three dimensional pie chart, you can also click on that three dimensional. You have A again is 40%, B is 20%, C is 20%, and D is 20%. So, in Google Sheet, you change the visualization and chart editor to a pie chart and you get the pie chart. Now let us do the pie-chart for our blood group data.

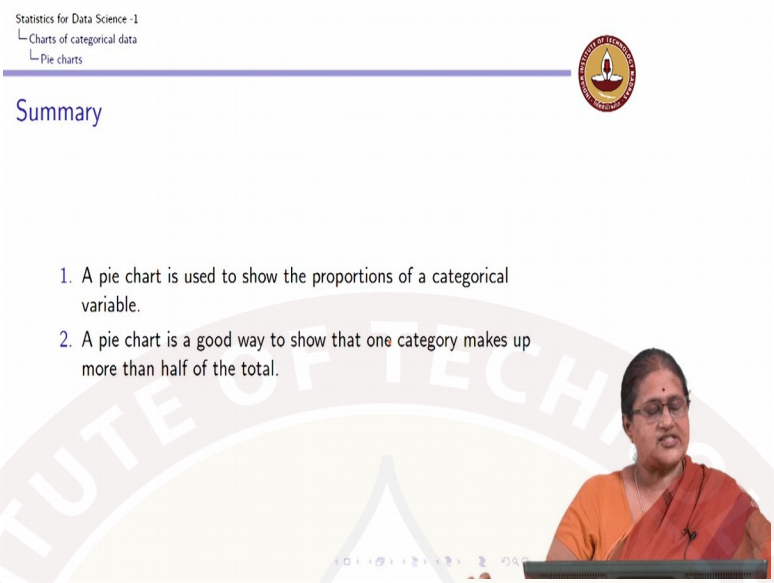
(Refer Slide Time: 06:44)



So, this is my blood group data. Again, what I do, I have to select the datasets, I want to get hold of the distribution of my blood group. So again, I go to Insert, I go to a chart and within the chart, I am going to look at a pie chart, and you can see that this is the pie chart, which says O+ is 16.7%, A- is 6.7%, O- is again a 16.7%, A+ is 20%, B+ is 13.3, B- is a 10%, AB+ is 6.7%, and AB- is 10%.

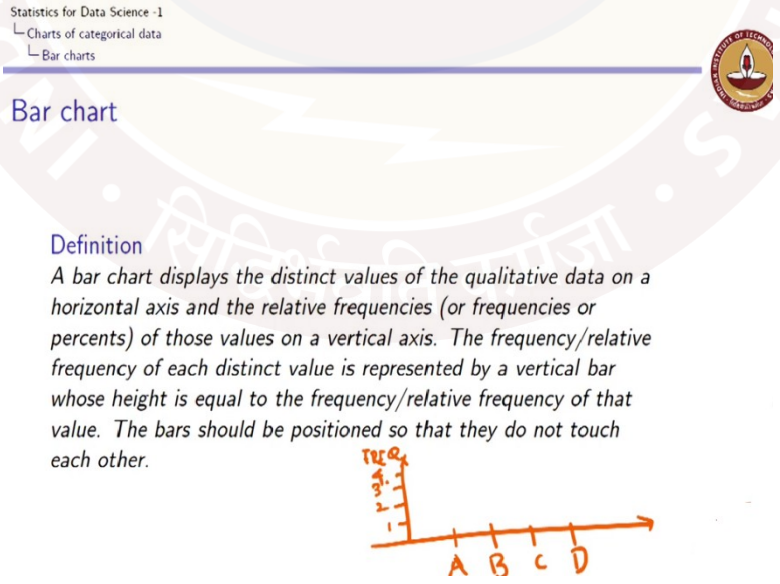
So, you can see that how we can construct and what are we constructing, the variable is nothing but the count of the blood group. This is giving me a pie chart, which is giving me a distribution of the blood group. Of course, within this you can customize it, you can again, do what is the legend and all of that and these things would be taken up in the tutorial sessions on the discussion board.

(Refer Slide Time: 07:56)



So by this time, you should have learned how to plot a pie chart. Remember, whenever the message is to show proportions of a particular categorical variable, a pie chart is a good way as it makes up too. And it will always tell you that the message, if one category, like in our example, where we had A made up for more than actually 144° is made up, you can see that out of, it has more than one way you can always use a pie chart to tell the share of a pie.

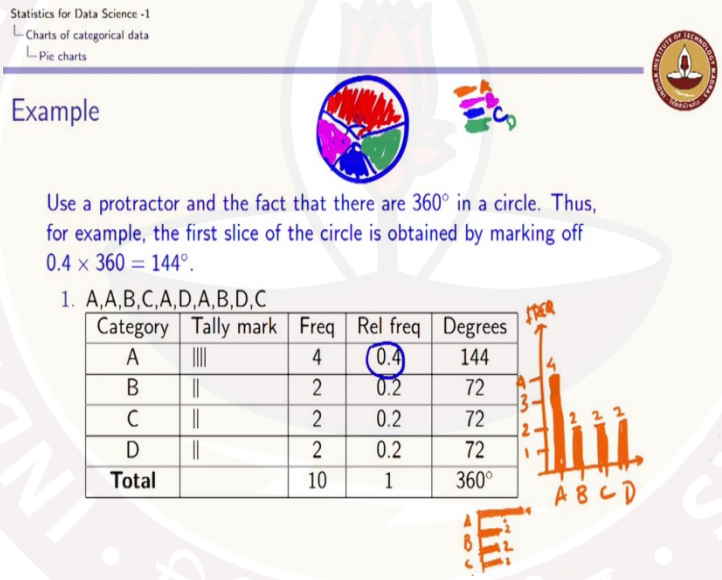
(Refer Slide Time: 08:34)



The next graphical summary which is also very important way, we said that when it comes to categorical data, the two most popular graphical displays are the pie charts and the bar chart. What is a bar chart? Again a bar chart, again, it displays distinct values of qualitative data on horizontal axis. So what do I have if I have a graph on my axis, I give the distinct value. For example, the distinct values were A, B, C, and D. These are the distinct values.

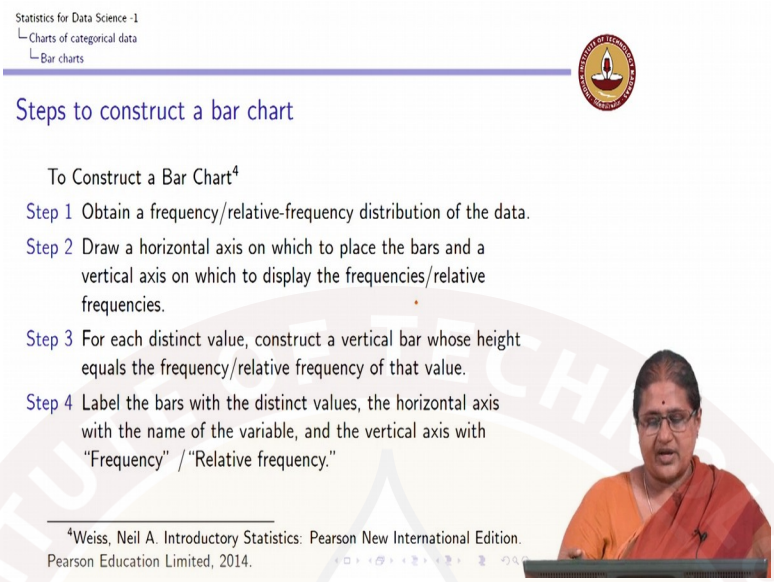
Now this, I really do not care whether it is a B, A, C, D because there is no order in this particular variable. But however, I need to be very clear as to what is the variable and the distinct values are given on the horizontal axis. On the vertical axis, I either can plot the frequencies or the relative frequency depending on what is my interest. We start with just a frequency. For example, if this were a 1, this is a 2, this is a 3, this is a 4.

(Refer Slide Time: 10:14)



Let us go back to our table, draw horizontal axis. Now, if I go back to my example I had here on the horizontal axis, I am just going to plot A, B, C and D. This is what I plot on the horizontal axis. On the vertical axis, I have 1, 2, 3, 4. So, A I draw a bar corresponding to A, the bar should be of equal width for each category, B is a 2, C is a 2, D is a 2. So here, I have the frequency, which is on my Y axis, I could have different colors, but I can have the same color also. But this is what is a typical bar chart.

(Refer Slide Time: 11:12)

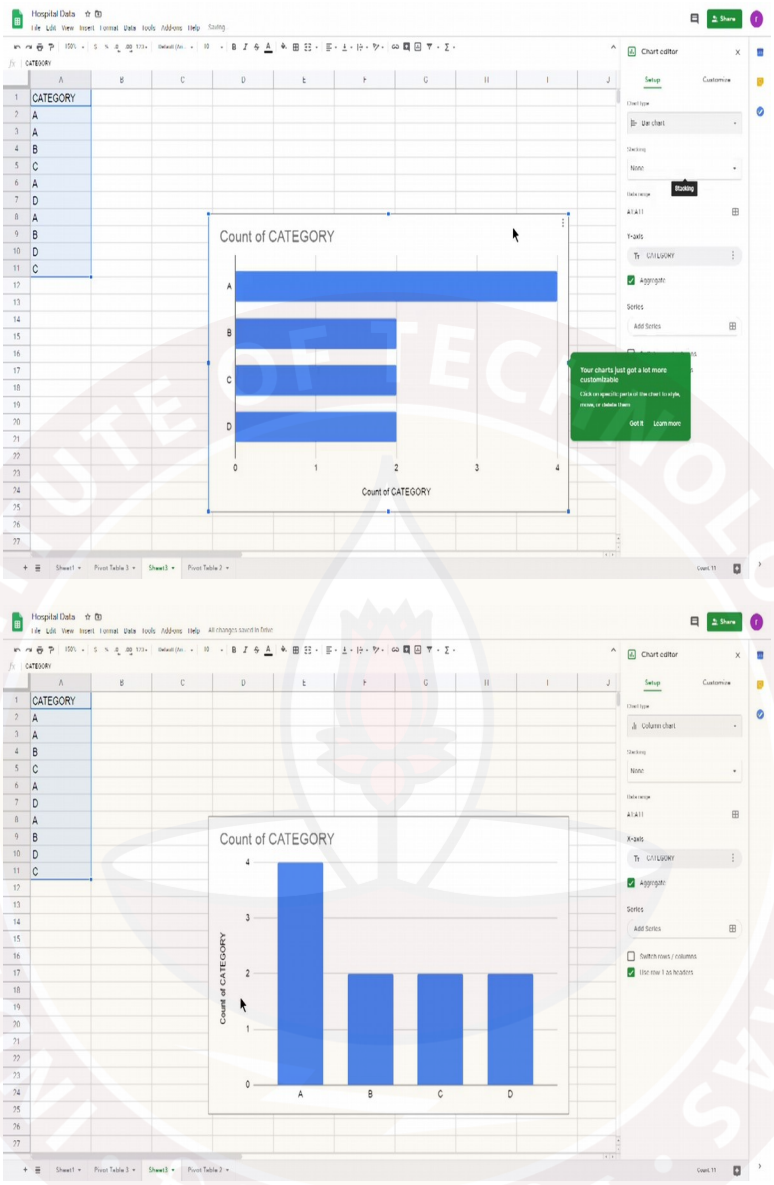


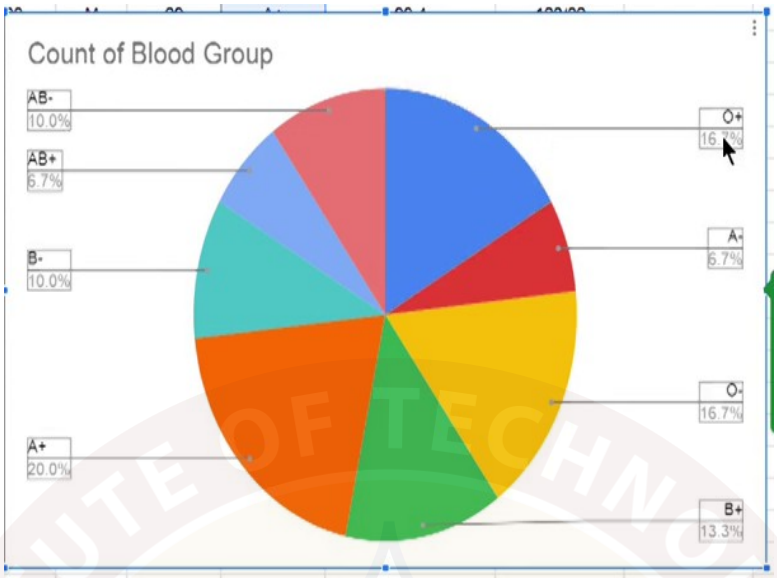
So what is needed, on the horizontal you have the bars, on the vertical the frequency. Label the bars so I can also go and one way I can annotate these bars, so I here I can right since I know the counts, just as I wrote the angles and the percentages for a pie chart, I can write a 4 here, or 2 here, or 2 here and a 2 here, which actually gives me the count of the Category A in this bar. So, this is a typical bar chart.

Now, a bar chart need not be only vertical, you could also have a horizontal bar chart where I have the categories on my axis which is A, B, C, D, and the counts C and D, and the counts which are given, here, I can have a A which is taking the value 4, so I have an axis which will give my value 4, this is B which is 2, C which is 2, and D which is 2. So, bar charts could be either vertical or horizontal. And depending on how you want to display your data, what is the message you want to give with your data, you can choose to have a vertical bar chart or a horizontal bar chart.

So, how do we create a bar chart in a Google Sheet, we go back to the data we have been working on. So I have, I highlight my data, that is my first step, which is given here.

(Refer Slide Time: 13:17)

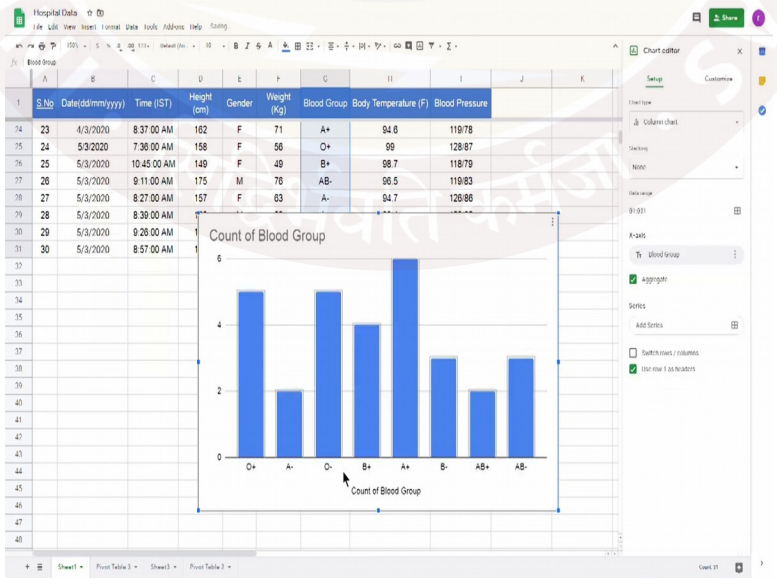




Once I highlight my data, I go to the Insert option with the chart, I choose a column chart here. Now you can see that I can choose either a count of category, which is a chart I said before, or I can also choose what is called a column chart. This is a column chart, just. So you can see that this is a column chart where again you can see on the y axis I have a count, I had 4 of Category A, I had 2 of Category B, I had 2 of category C, and 2 of category D.

Now, let us repeat that with our blood group data. If you go to the blood group data, again, I highlight the blood group data which is already highlighted. I go to insert. I go to chart.

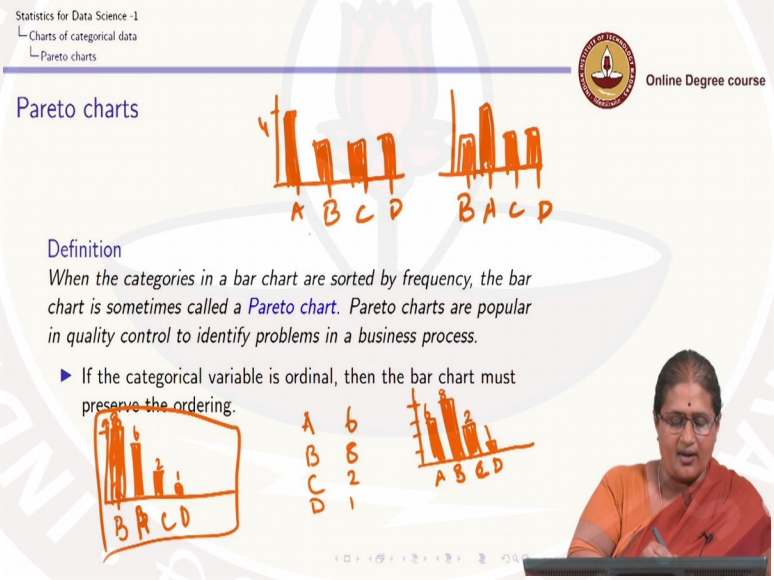
(Refer Slide Time: 14:02)



And you can see that this is exactly the count I had. And this is how this comes again from a relative frequency and my pie chart, you can see that the difference between the pie chart and the column chart is here. I know what is my count and I can annotate each one of the bars with exactly what is the count of people who are having that particular blood group. For example, there are 6 people out of 30 who have blood group A+, there are 2 people who have blood group A-, there are 2 people who have AB+ so you can easily see the count of people, there are 4 people who have B+.

What the pie chart gave us was the share of total, what is the relative frequency and you can see a bar chart is actually giving you the frequencies. You can also plot the relative frequency in a similar way.

(Refer Slide Time: 15:57)



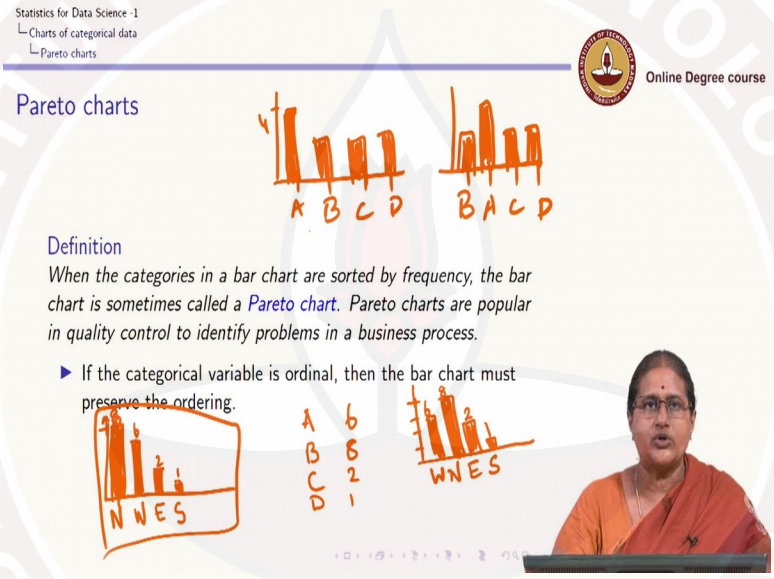
Now, many a time, what you might want to know is, for example, in this suppose I have been given this as a chart, I might want to have the bars which are arranged in a particular order. For example, here I see, A bar has the highest bar or the longest bar if it is a horizontal bar chart, next is both O- and O+, then I have B, B+ then I have AB-, B- and then I have A. So, I might want the bar chart to be arranged in a, as I already said, there is no problem in having a different order.

So in a sense, what I mean is when I have a bar chart, whether A, B, C, D. I have a 4, I have a 2, I have a 2, I have a 2. I could also displays display this as B, A, C, D, where this is a 2, this is a 4, this is a 2, and this is a 2. No problem with either of these displays because they convey the

same information and there is no order between A, B, C, and D. But however, in this I have the highest frequency appearing first and then I have the lower frequencies which are appearing.

For example, if I had a frequency distribution where A was 6, B was 8, C was 2, and D was 1. If I plot a bar chart with 2, 4, 6, 8, my A was a 6, category A was 6, category B was 8, category C is a 2, and category D is a 1. So, I have a 6, 8, 2, 1. This is giving me a count. A Pareto chart is something where I can just look at the category B which has the highest frequency, which is a 8, then comes B, A which is 6, then come C which is 2, and last is D which is 1.

(Refer Slide Time: 18:53)

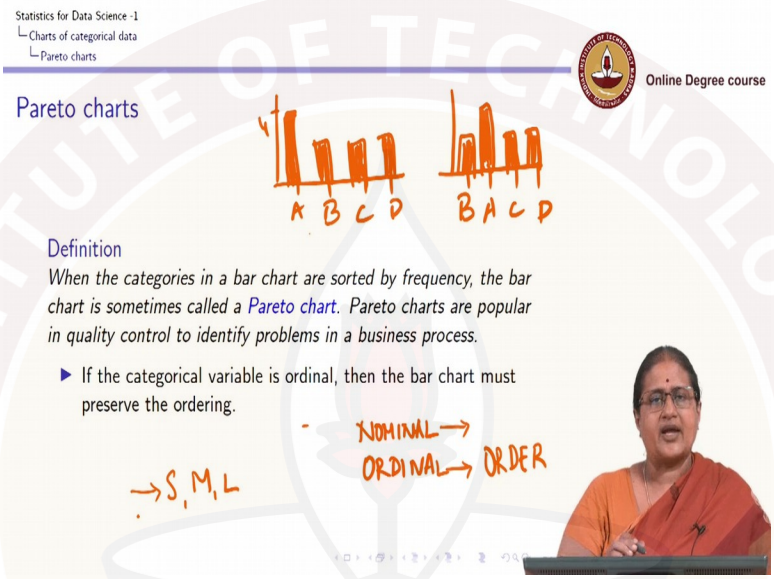


Now when would a chart of this kind be of use to us. Suppose instead of A, B, C, D I had our states, suppose or I had our regions, suppose I had the northern region, the southern region, the eastern region and the western region. Broadly classify any particular geographical area into four particular regions, and I am interested in knowing what is the employment rate or what is the number of students who are passed in each region.

If I present a chart in this way, and for just hypothetically this represents 8000 students, 6000 students, 2000 students, and 1000 students and I have a distribution. Again, just for hypothetical purposes, this is the northern, this is the western, this is the eastern, this is the southern region. So at a snapshot, I can immediately see, now this data would have been the same as this data, where this was the southern region, this was the eastern region, this was the northern region and this was the western region.

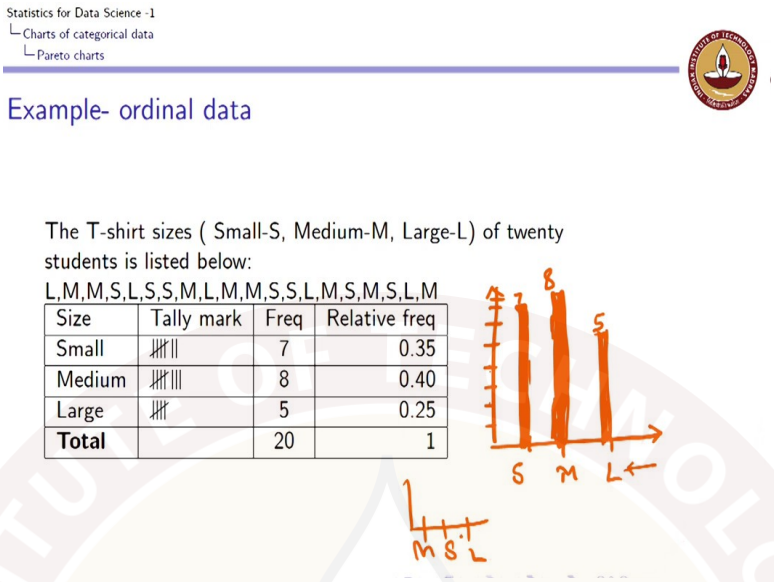
Both the data convey the same message, whereas the Pareto chart in some sense, gives me the idea at a snapshot. I can know that the northern region has the highest number of people followed by the western, then the Eastern and the southern. So, you can see that the bars are in descending order. I can also have the bars in ascending order. Again, it depends on what is the message you want to convey. Hence, in this chart, these type of charts are referred to as Pareto charts.

(Refer Slide Time: 20:03)



Now recall when we talked about categorical variable, we said categorical variable can be measured in two scales and that scales I call them as a nominal scale and I also call them as an ordinal scale. Now the key difference between a nominal scale and an ordinal scale is in this I just have name or values, whereas ordinal Scale I have a natural order. For example, if I am looking at small, medium, large, there is a natural order even though these are categorical and I don't have numerical values assigned to them.

(Refer Slide Time: 20:47)

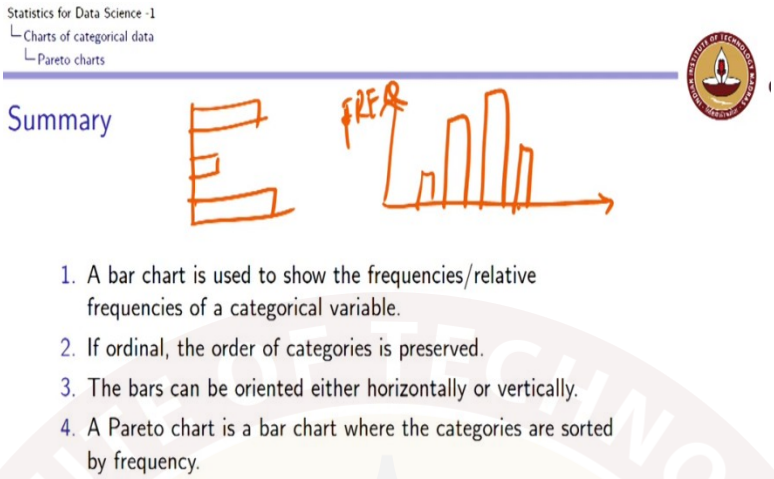


Now when the categorical variable is ordinal, like I have in this data, I have T-shirt sizes small, medium, large of 20 students. So, I have constructed the tally mark. So, I have 7 students who have small T- shirts, medium are 8, large are 5. So, I can construct my frequency or my bar chart for this data. So I have the size of the T shirt, which is small, medium, and large, the sizes, so I have a 1,2,3,4,5,6,7,8. I have a size 7 for my small, a size 8 for my medium.

The bar should be of the equal width since it is free hand. The bars do not appear to be but you can see that your Google Sheets always provide bars of the same width. 3, 4, 1, 2, 3, 4, 5, and large is 5. I further annotate them, this is a 7, this is a 8, and this is a 5. So, whenever there is an ordinal it is good to maintain the order of the categorical data.

For example, I do not want to have a chart which is medium, small, and large, because the order of this categorical data is not maintained. So, even within a bar chart for a categorical data, if you have an order, please maintain that order of the categorical data.

(Refer Slide Time: 22:35)



So in summary, we know, a bar chart can be used for frequencies or relative frequency depending on what you want to convey. On the x axis, I have the categories, and on the y axis I have the frequencies which are the counts if I am plotting the count, or the relative frequency if I am looking at relative frequency, I can either have a horizontal bar chart, where I have, a horizontal bar chart would look something of this kind or a vertical bar chart, which would look something of this kind. A Pareto chart is where the categories are sorted. And if you have ordinal data please try and preserve the order of the categories.