IIT Madras

ONLINE DEGREE

Statistics for Data Science - 1

Prof. Usha Mohan
Department of Management Studies
Indian Institute of Technology, Madras

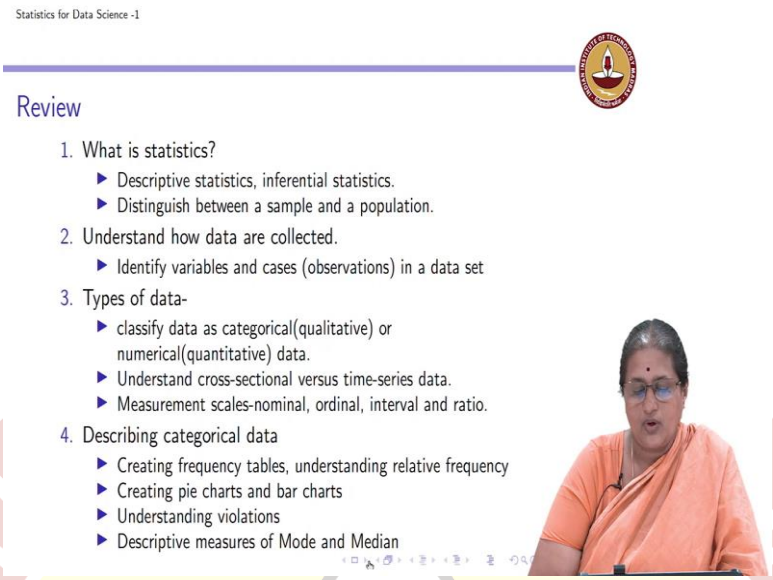
Lecture – 3.1
Describing Numerical Data – Frequency Tables for numerical data

In today's module, so we are going to start with Describing Numerical Data using for a again for a single variable. Just a brief review of what we have done so far. So, if you look at what we have done so far, we started by understanding what is statistics, we broadly introduced to the 2 main branches of statistics which is descriptive statistics and inferential statistics that you will learn as a beginner.

Then, we introduce the concept of a sample and a population and then afterwards we also said; what is the difference between a sample and a population. Then, we went on to understand how data is collected. Now, once you collect the data we did not focus too much on how to collect data, but once you collect the data we said that, you can actually put down this data and identify the variables and the observations and set up what is called a data set.

Once that is done, then we also discussed about the types of data. Broadly, you can classify data as numerical and categorical and then we also understood the difference between cross sectional and time series data. Finally, we looked at the measurement scales; namely, nominal, ordinal, interval and ratio scales of measurements.

(Refer Slide Time: 01:47)

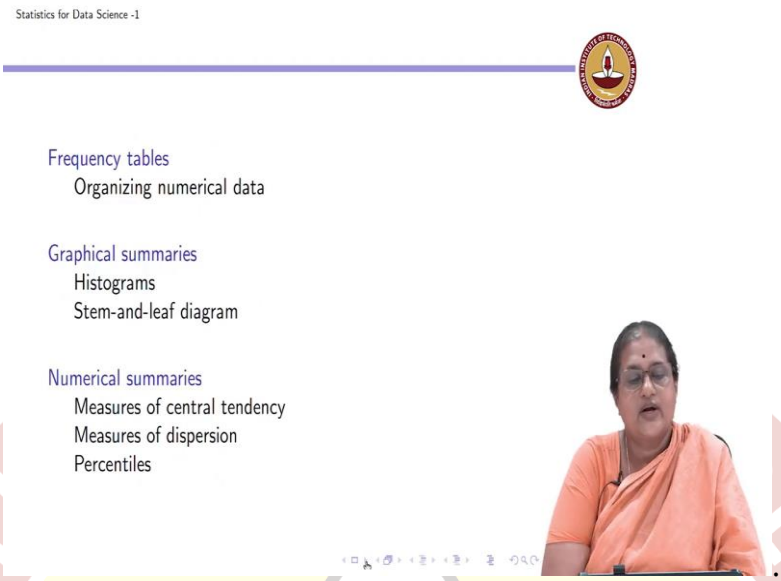


Then, we went on to understand how to describe categorical data. When we talk about describing categorical data, the key point we have noticed are we first looked at what are the how to create frequency tables. Now, when we create a frequency tables we introduced the concept of what we call relative frequency. After constructing frequency tables we looked at the graphical summaries of data.

For this we focused on 2 main graphical summaries; namely, the pie charts and the bar charts. We also looked at what could be the possible violations that could occur in terms of misrepresenting data and how to overcome the same.

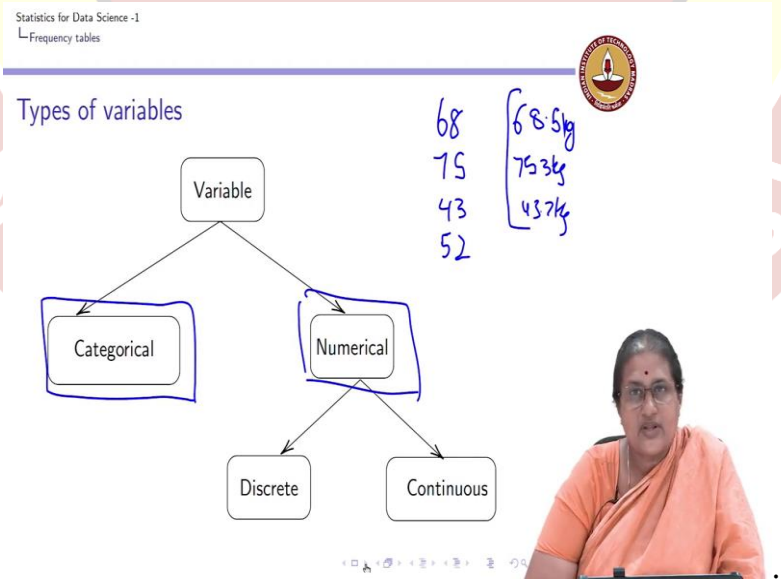
Finally, we looked at 2 descriptive measures which are used for categorical data; the mode and the median, while mode is applicable only when you have while mode is applicable for both the nominal and ordinal data, median can be defined only when you have ordinal data. So, this is where we stand at this point of time.

(Refer Slide Time: 03:07)



What we are going to do next is to understand how to summarize numerical data. The agenda is first we again will go through frequency tables, then we look at graphical summaries and then we will focus on the numerical summaries and measures of description.

(Refer Slide Time: 03:25)

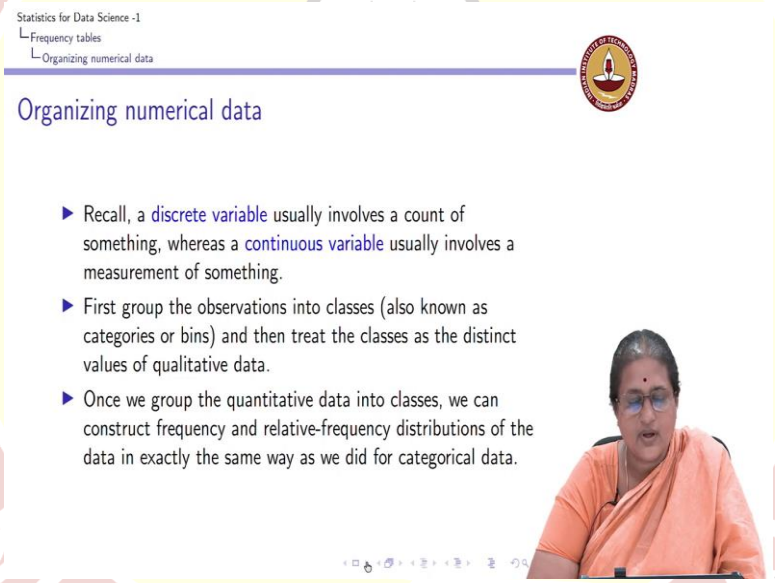


Recall, when we talk about variables we broadly classified variables into the categorical variable and numerical variable. We have understood this portion. In the last module we understood how to describe categorical data. So, now, we are going to focus on how do

we describe or summarize numerical data. Again, when we look at numerical data they can be broadly classified into discrete and continuous. Now, what do we mean by discrete data? Discrete data as the name suggests for example, if I am looking at marks obtained by students and the marks obtained are 68, 75, 43; 52.

In a sense that these are discrete, but now when I am looking at weights of the same say 4 students and I can I am recording it in kilograms, its 68.5 kilograms, 75.3 kilograms, 43.2 kilograms etcetera, I have what is called continuous data. So, we are going to focus on; how do we actually summarize both discrete and continuous data. Again, we will talk about how do we summarize when we have small data that is every discrete data can be treated as a single data value and how do you group data. So, that is the agenda for now.

(Refer Slide Time: 04:57)

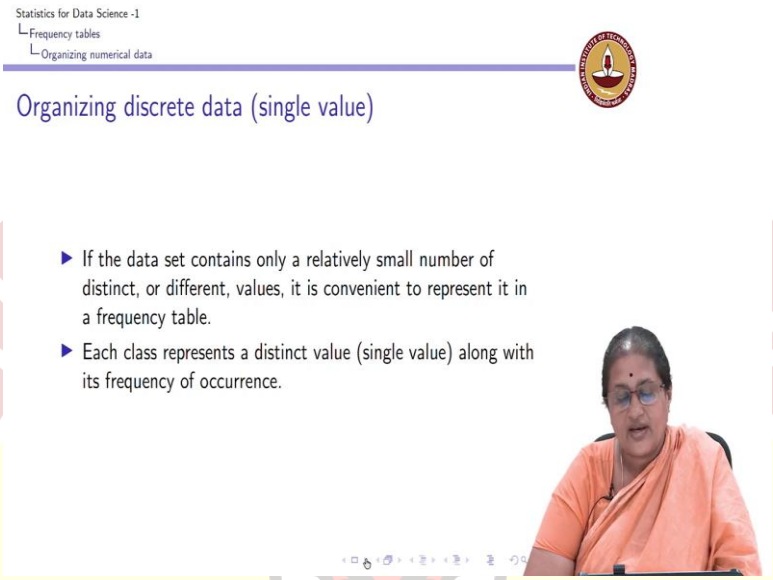


So, now let us go ahead and understand how to organize numerical data. Now, a discrete variable involves a count of something, whereas, a continuous variable involves a measurement of something. We have already given example as to what is a discrete data and what is a continuous data. So, the first idea is we already have seen how to summarize the categorical variables using frequency tables.

So, we now want to see whether we can apply the same technique here or the same idea when we want to summarize numerical data. So, one way to do this is to treat the numerical data as categories. So, that can be done by grouping the observations into categories or bins and then treat the classes as distinct values of qualitative data.

Once I have these discrete categories, then I can I already know how to construct frequency tables for the categorical data. I can use the same thing and construct a frequency table for my discrete or numerical data. So, let us see how that is done.

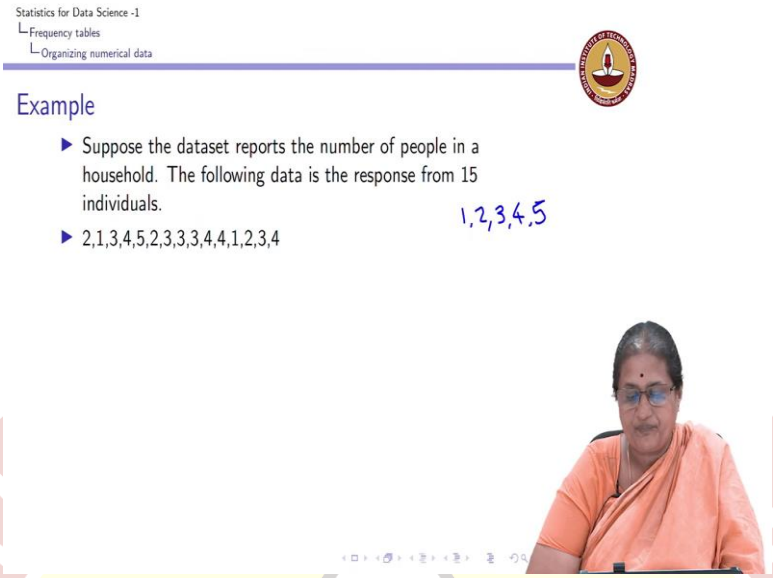
(Refer Slide Time: 06:21)



So, now suppose my discrete data is a single value data, how do I organize it? So, if the data set consists of relatively small number of distinct different values, then I can treat each value as a category. I repeat if my data set contains only a relatively small number of distinct values, I can treat each distinct value as a category.

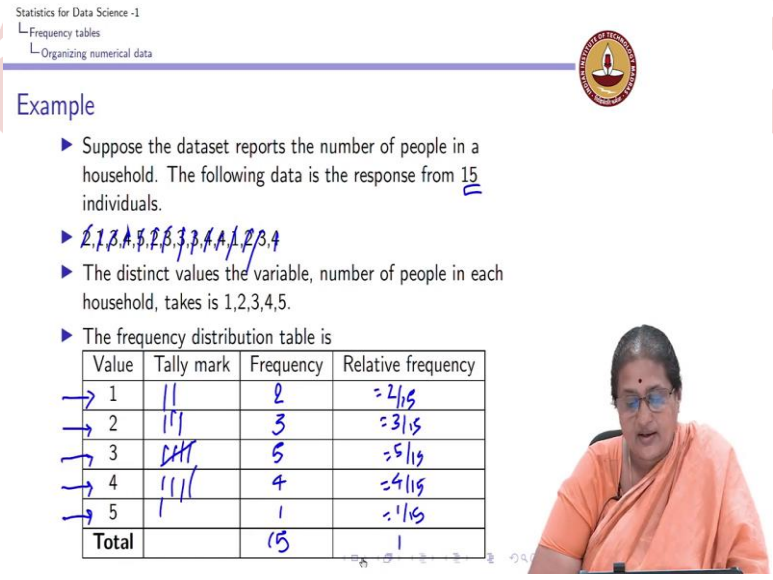
And, then once I have each distinct value as a category, I find out what is the frequency of occurrence of each distinct value and I construct a frequency table the way I constructed a frequency table earlier when I talked about categorical data. Let us see this through an example.

(Refer Slide Time: 07:15)



Suppose, I have a data set that reports a number of people in a particular household. So, I go to 15 people and I have the following response from 15 individuals. The numbers 2, 1, 3, 4 and 5 and so on indicate the number of people in that particular household. There are 15 values and each one of this value has been picked up from these 15 people. So, now, you can see that I have 5 distinct values of my data and the distinct value of my data are 1, 2, 3, 4 and 5. So, this is the distinct value, I can treat each of these values as a category.

(Refer Slide Time: 08:15)

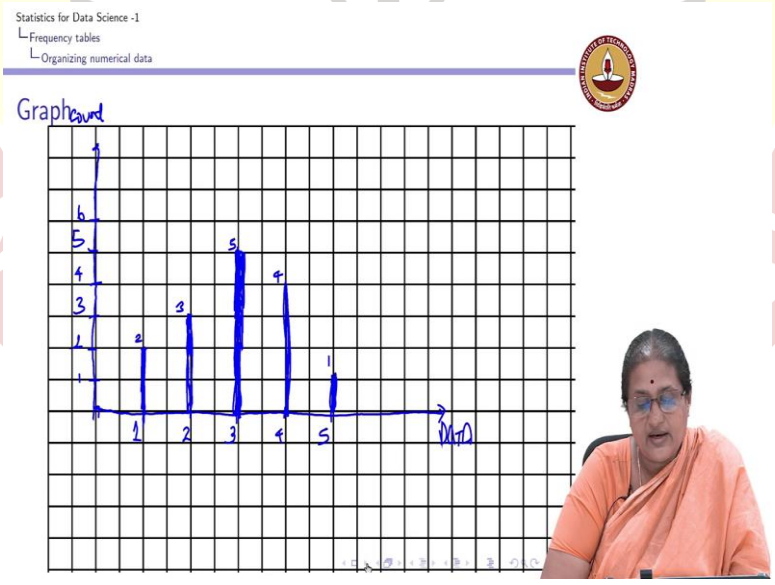


Now, if I treat each one of these values as a category then afterwards I can I put these values. So, you can see in the frequency table I have value 1 here, I have value 2 here, value 3, value 4 and value 5. So, I have these 5 values and for each one of these values I can put a frequency I can construct my frequency distribution table same way I constructed for the categorical variable in my earlier session.

So, now let us construct the frequency table for this data. Now, look at 2. So, I put a tally mark here. 1 is the next value here. So, let us 3, 4, 5, another 2, 3, 3, 3, 4, 4, 1, 2, 3, 4. So, this frequency is 2, this is 3, this is a 5, this is a 4 and this is a 1; this adds up to the 15 which is the number of people we have here. So, we can see that each I can construct the frequency table I have considered each one of these distinct values to be a category.

The relative frequency as we have defined earlier is 2/15, 3/15, 5/15, 4/15 and 1/15; we know the relative frequency adds up to 1. So, this is how we can when my data is relatively small in number of distinct values, I can treat each one of the distinct value as a separate category and construct the frequency table just the way we did it for a categorical variable.

(Refer Slide Time: 10:29)



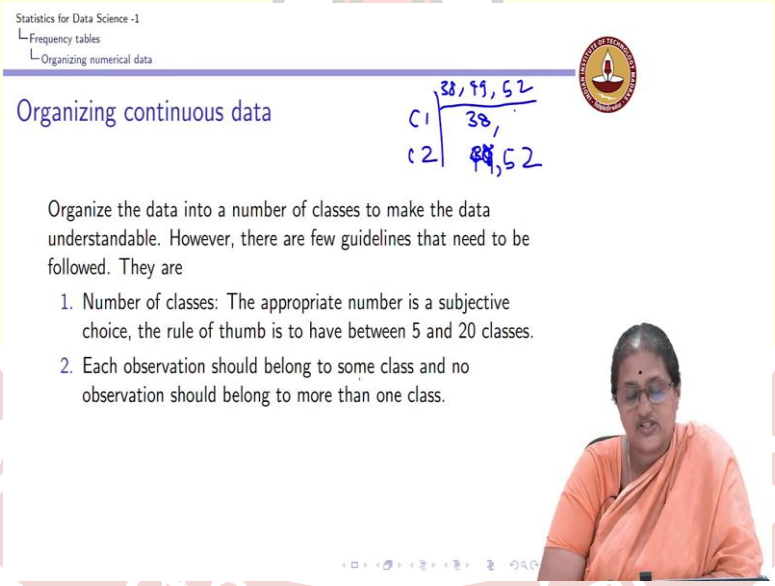
Now, given this I can plot this again, we have seen how we plotted the data. So, I can plot this data in the following way. I have my y-axis, I have my x-axis here. So, let me give the count on the y-axis. The count on the y axis is 1, 2, 3, 4 and 5, 6. I have my data it is 1, 2, 3, 4 and 5 and we can see that, the data frequency the count of value 1 is 2, I

have a bar which is of height 2 here; count of 2 is 3, I have a bar which is of height 3 here; count of 3 is 5, I have a bar which is of height 5 here; count of 4 is 4; count of 5 is 1. So, you can see that, this is very similar to what 2, 3, 5, 4, 1.

So, I can annotate these bars 2, 3, 5, 4 and 1. This is how I can graphically describe the data. What I want you to notice here is, now because this is numerical there is an order in the data. We also described when we talked about plotting bar charts for ordinal data; it is good to preserve the order and data.

So, it is good to have an order here for the data which I have which is number of people in a household. This is the count and this is the data. So, a simple bar chart is what we have shown here. And, since these are distinct values, I am not connecting the bar charts together, I have just listed out what is the height of each bar chart.

(Refer Slide Time: 12:43)



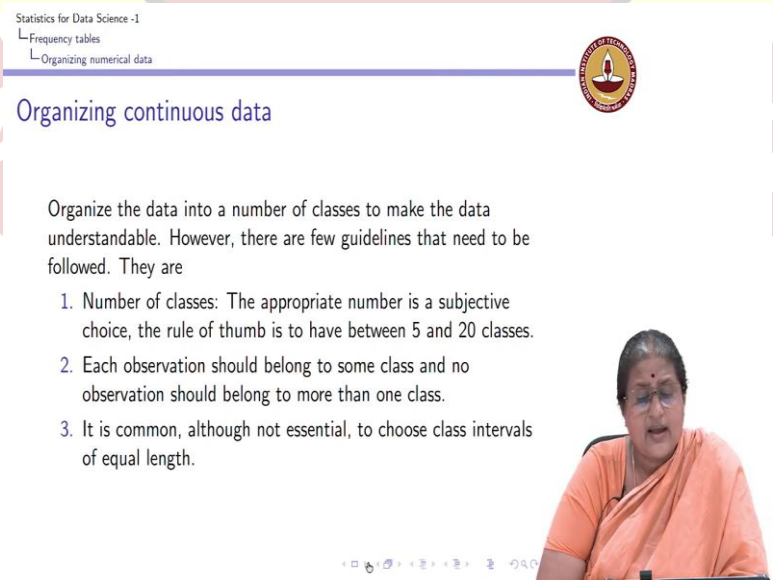
The next thing we want to understand is what would happen when I have continuous data. We said that when I have simple or discrete values which are smaller number then I can treat each discrete value as a category and I can come up with a frequency table. Now, suppose I do not have a small number of discrete values and I have a whole array of values then my; suppose we have a whole array of distinct values then by constructing a frequency table for say I have 65 distinct values, it would be very cluttered.

So, one way to come up with the frequency table for this kind of data is to try and see whether we can group this data into what we refer to as classes. So, now, how do I organize and this can be done even when my data is continuous, ok. So, how do I organize my data into classes?

So, the guidelines we need to follow are the first thing is then I need to choose on how many classes I need. The number of classes is an subjective choice usually, it is good to have between 5 and 20 classes. So, that you do not clutter your table too much. The next thing which we need to understand is each observation and this is very very important; each observation should belong to some class and no observation should belong to more than one class.

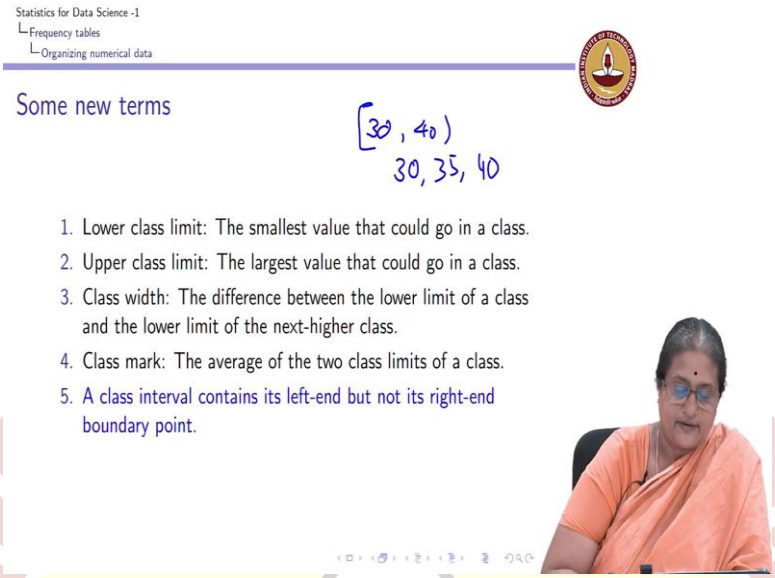
What do I mean by the; suppose I have constructed a class 1 and class 2, we will learn now in very shortly we will learn how to construct these class and I have observations 38, 49, 52. I cannot have 38 which belongs to both class 1 and class 2. So, every observation should belong to either class 1 or class 2 and it cannot belong to more than one class, it should belong to one of these classes. So, I could have 38 here, I could have 49 and 52 in the second class. So, every observation is in a class I have defined that is what this means.

(Refer Slide Time: 15:15)



And, it is also common though it is not essential, but it is a good practice, at least for beginners to start with class intervals which are of equal length.

(Refer Slide Time: 15:29)



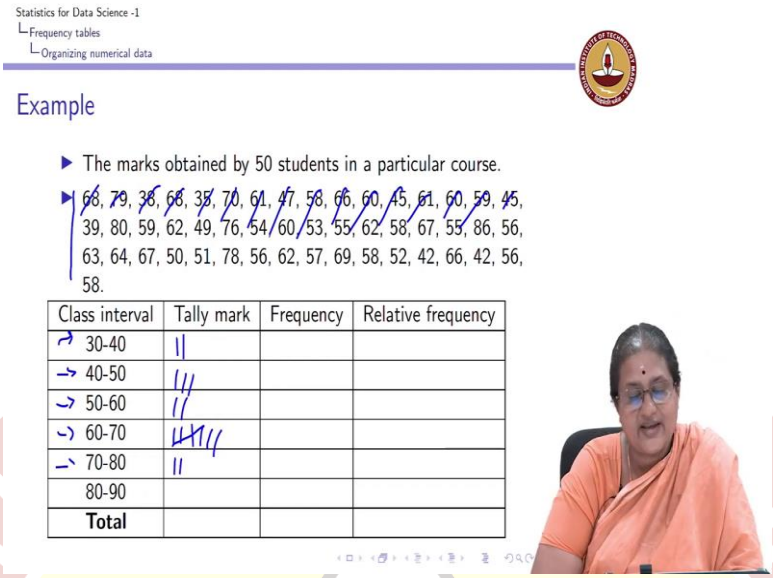
So, now let us look at defining a few terms which would be helpful for us to construct frequency tables for group data. So, now, what do I mean? I said that we are going to construct what are called classes or class intervals because of looking at individual distinct values, now I am going to define an interval of values as a category.

So, the minute I define an interval I have what is called a lower class limit. The lower class limit is the smallest value that will get into a class. I have an upper class limit; an upper class limit is the largest value that will go into a class and the class width is the difference between the upper class limit and the lower class limit.

The class mark is the average of the 2 class limits. And, the convention we are going to use is the class interval contains its left end, but not the right end of the boundary point. Now, what do I mean by? Suppose, I have a class interval I have defined as 30 to 40 and my data has 30, 35; 40.

30 and 35 will belong to this class interval, whereas, 40 will not belong to this class interval. I repeat. If I have a class interval 30 to 40, the points 30 and 35 would belong to the class interval, whereas, 40 will not belong to this class interval because it is a right end boundary point of the class interval.

(Refer Slide Time: 17:07)



So, now let us look at an example and see how to construct a frequency table using this example. So, you can see that this has the marks of 50 students in a particular course. You can see that there are lot of distinct values and there are a few values which appear more than once.

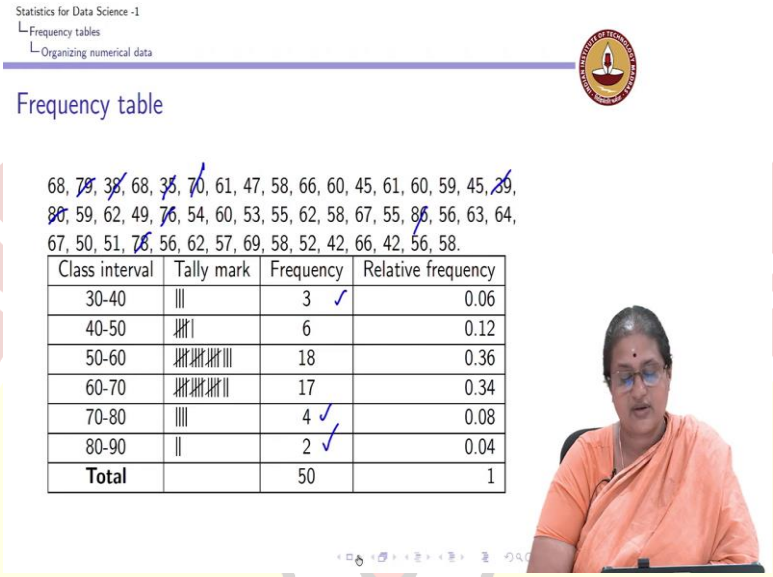
But however, the 50 students so I even if I look at the distinct values and try and adopt what I did earlier for a smaller data set my frequency table is going to get very cluttered. So, what I choose is I am going to choose class intervals of size 10 and start putting in each value after choosing the class interval of size 10.

So, the class interval size once I having defined my class intervals, I see that the minimum occurs in the range of 30. So, I start with a minimum of 30 and go up to 90, I do not have a mark beyond 90 and these are my class intervals. Now, I am assuming each class interval is of the same size. Hence, my class intervals are 30 to 40, 40 to 50 and so on 80 to 90. So, the assumption I have here is each of my class intervals are of the same size.

So, I, now these are my categories and I am going to put in each of the values from here into one of these intervals. Again remember, every value here goes into one of these buckets and no value will go into two of these buckets. I include for example, in this 30 to 40 I include 30, but exclude 40.

So, 68 goes here, 79 goes here, 38 goes here, 68 again goes here, 35 goes here, now come to 70; 70 goes here and because 70 is excluded from this interval, 70 goes here, 61 here, 47 here, 58 goes here, 66 goes here, 60 again comes here, 45 goes here, 61 comes here, 60 again comes here, 59 is here, 45 is here.

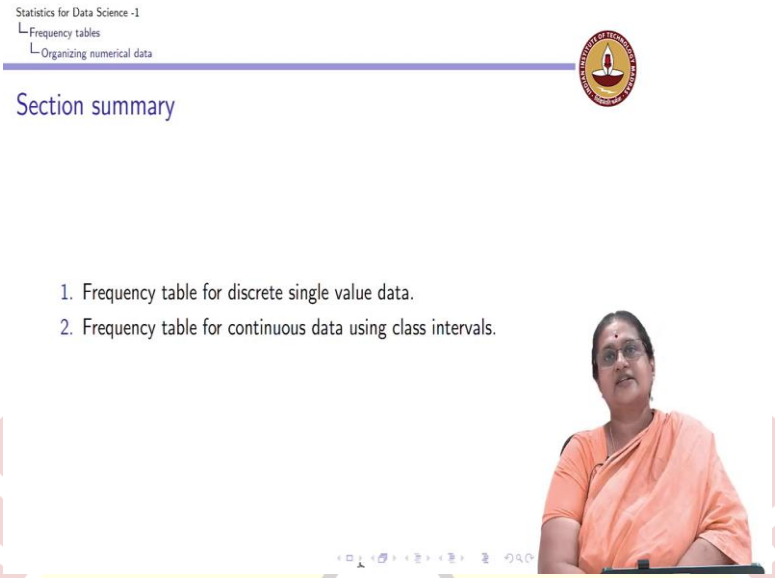
(Refer Slide Time: 19:43)



So, if I continue this I get a frequency table which is of this kind. So, you can see that, 30 to 40 you can easily see its 38, 35 and 39 these are the 3 things. So, it has a frequency of 3, 80 to 90 has 86, 80; because 80 to 90 is in that it has a frequency of 2. 70 to 80 has 79, 70, 76, 78 ok, with the frequency 4 and all these frequencies as always add up to 50, the relative frequencies are also given here.

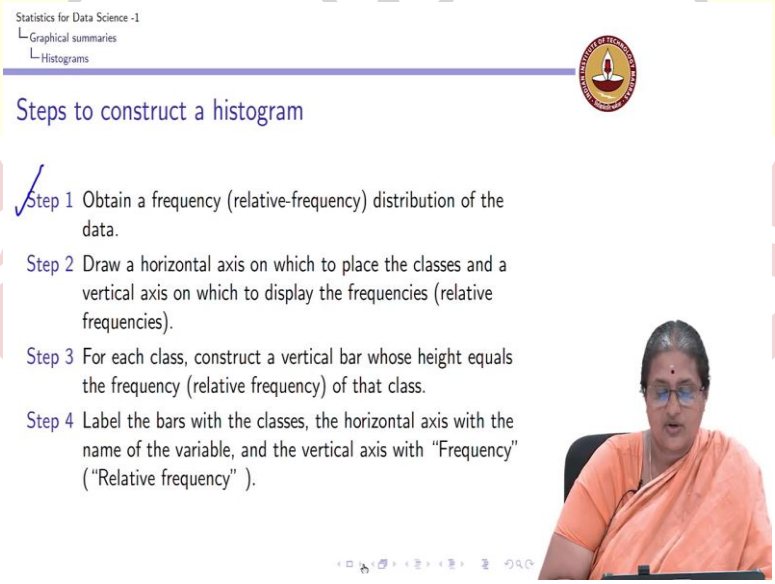
Now, remember here what we have done and what is the difference between what we have done here and the earlier part is we have grouped the data into class intervals. So, this is how we construct a frequency table when I have group data and this could be even for continuous data I can use the same way.

(Refer Slide Time: 20:57)



So, now what we have learned so far is how to construct a frequency table for a single value discrete single value data and how to construct a frequency table for continuous data using class intervals.

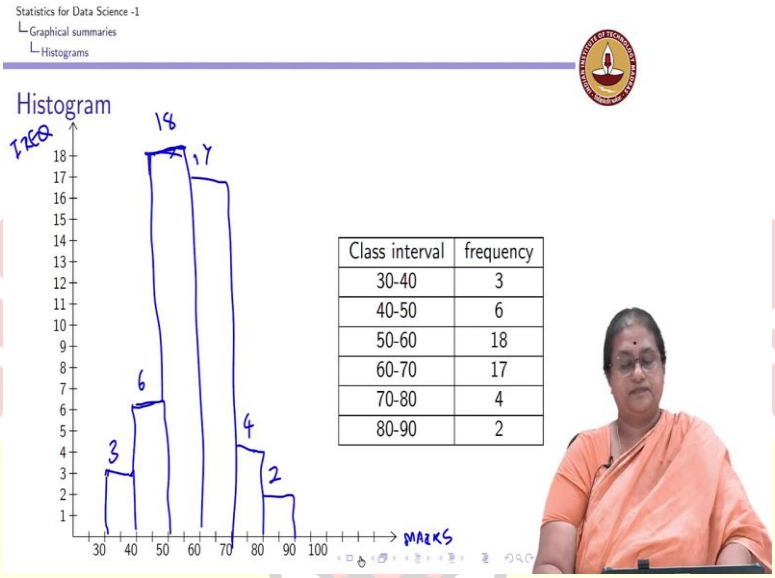
(Refer Slide Time: 21:17)



Now, we will go about seeing now how do I summarize the data for the second case that is where I have constructed a frequency table for continuous data. The way we do it is through what we call histogram. Histogram is one of the most popular graphical summary of a continuous data. So, now, we will understand how to set up a histogram.

Now, when we want to set up a histogram the first thing is we obtain a frequency distribution of data. We have already seen how to obtain this frequency distribution of data. Draw horizontal axis.

(Refer Slide Time: 22:01)

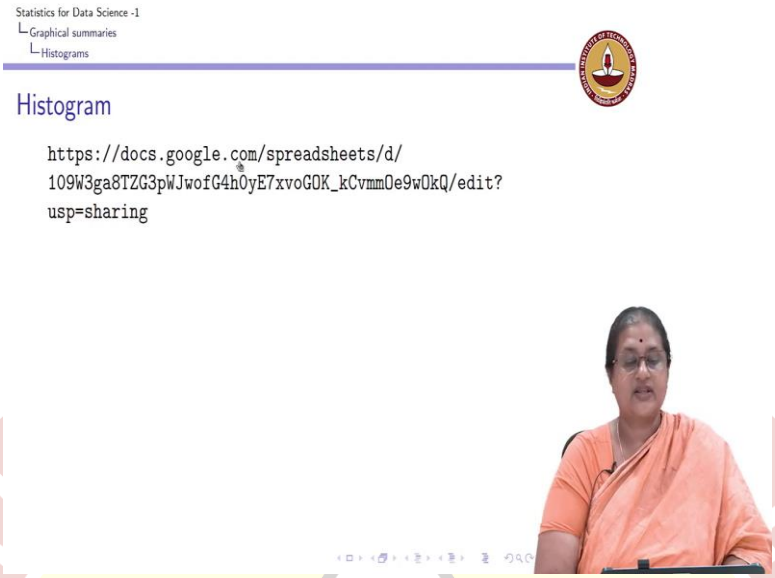


So, on the horizontal axis I have laid my 30 to 40, 40 to 50 and these are my marks that is on my horizontal axis. A vertical axis on which you display frequency. So, on this axis I have the frequency, on this axis I have my marks which are given in the as intervals. The next thing is for each class construct a vertical bar whose height equals the frequency. So, now, for each class, so you can see for 30 to 40 I have a frequency 3; so, I construct a bar with frequency 3.

40 to 50 I have frequency 6; so, I construct a bar with frequency 6. 50 to 60, its 18, I construct a bar with frequency 18 then I come its frequency 17. 70 to 80 is frequency 4 and the last thing is frequency 2. So, this you can annotate it. I have 3, I have 6, I have 18, these are my counts, this is 17, this is 4 and this is 2. So, I have constructed my histogram.

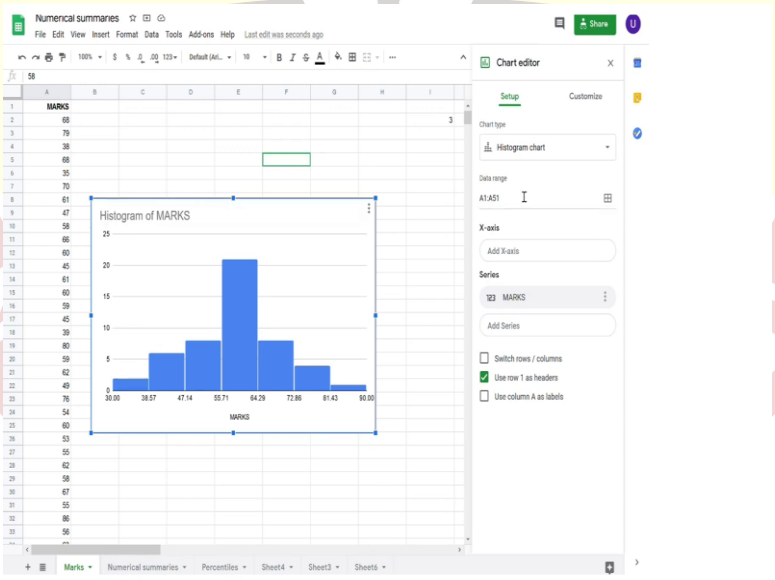
Notice the difference between a histogram and the bar chart, because my class intervals are in a sense continuous. I am not leaving a gap between these bars. So, it is a continuous display of data. The vertical height of this bar represents the count in every class interval. So, this is how we have. So, I label the bars, the vertical axis and this is how we construct a histogram.

(Refer Slide Time: 24:11)



So, I can construct a histogram using my Google Sheets as well and that is what we are going to discuss now.

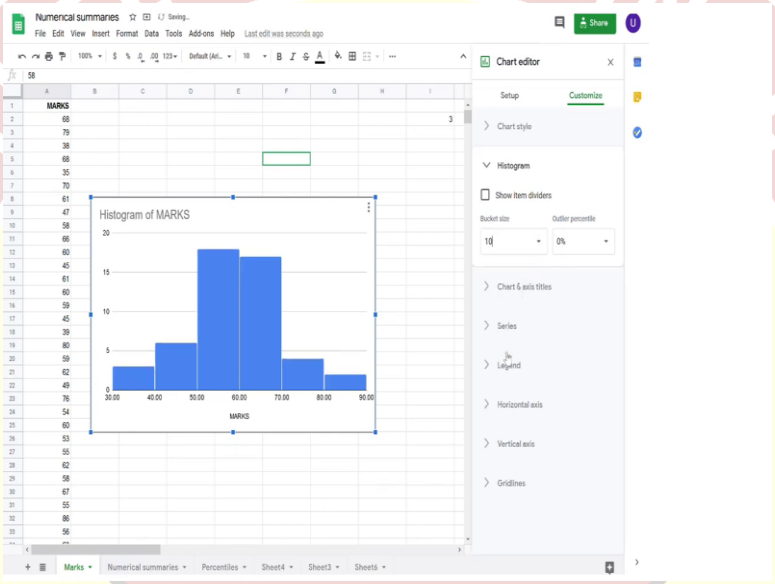
(Refer Slide Time: 24:19)



So, this is the data which we have. So, the same data which was for 50 people is listed here. So, I have a data for 50 students, which is listed here. You can see the data which is here. So, the data is for 50 students. This is the same data which we had here. This is given in column A. The first row is the heading for the row.

So, how do I construct histogram using my Google Sheet? Go to Insert, in Insert you have a Chart. Now, in this chart you can choose the data. So, you can see that the chart type is a Histogram chart. The data range is your A2 to A1 to A51, but I have clicked on using row 1 as headers that is being clicked here. Using row 1 as a header is being clicked, ok. And, the series is the marks. So, you can see that the minute I do not specify the equal buckets here, the histogram randomly chooses the end points of each class interval.

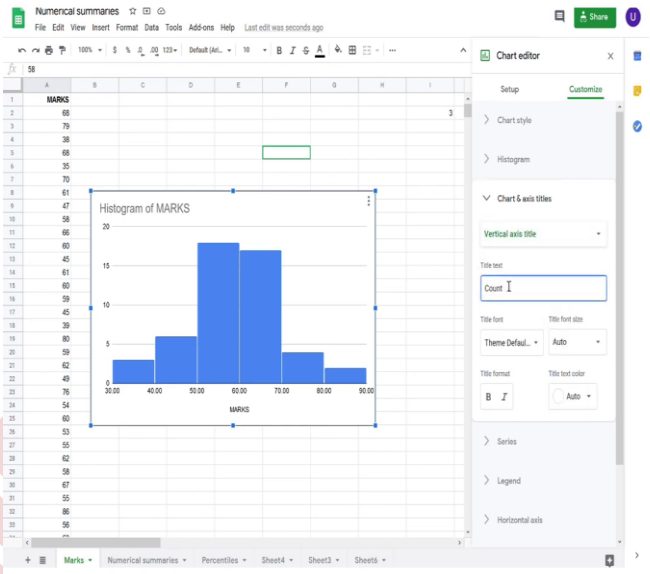
(Refer Slide Time: 25:55)



So, I go to customize, in the customize thing, I go to chart in under my histogram, there is a heading which is called bucket size. In bucket size, I click on 10, that is the size of my class interval instead of auto and you can see that this is precisely what we earlier drew freestyle with 30 to 40 having a frequency of 3 and this is 18, this is 17, this is 4 and this is 2.

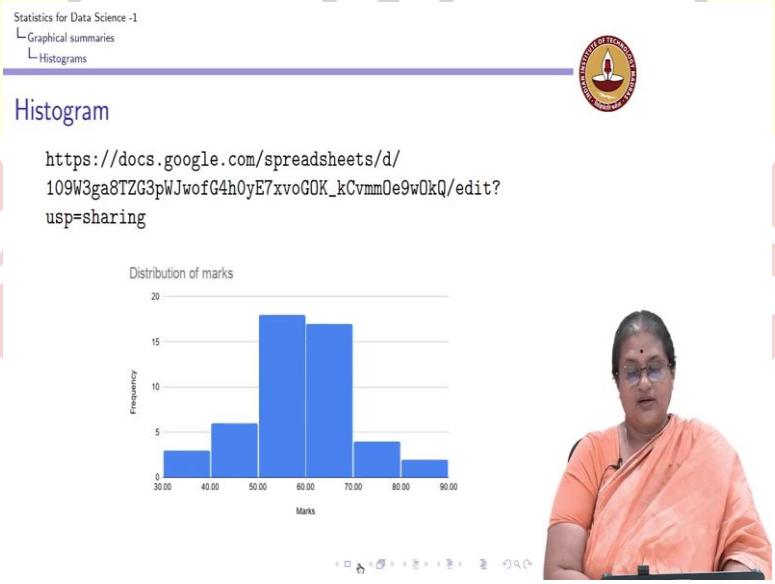
As again as always we can write down what are the legends, we can find out what is the what is my horizontal axis, what is your; you can go to the chart, you can have what is your title text, you can find out what is your subtitle, what is your horizontal access title, I can list it as marks.

(Refer Slide Time: 26:43)



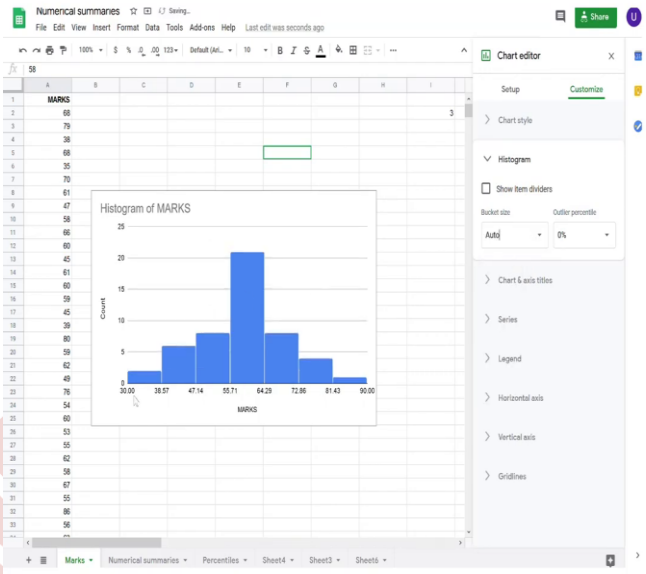
My vertical axis title I can list it as count and that is what I have the count and I can remove grid lines and I can do everything else, but this is how I construct a histogram using my Google Sheets.

(Refer Slide Time: 27:01)



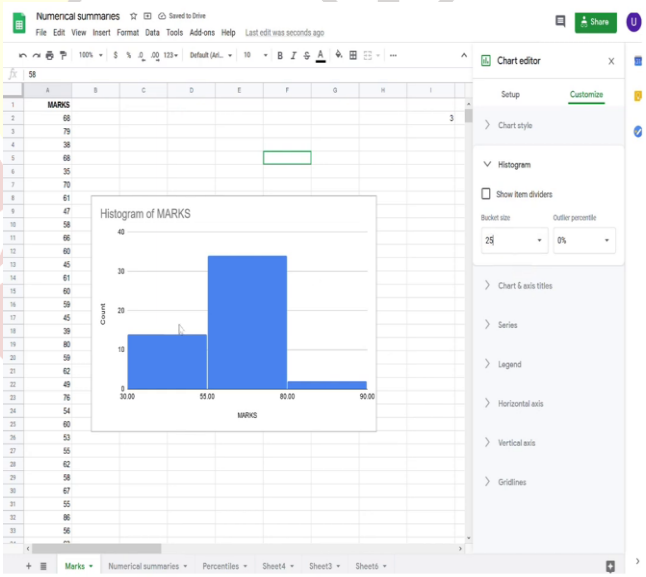
So, this is what we have. And, this is how we showed. We construct a histogram using the Google Sheets.

(Refer Slide Time: 27:17)



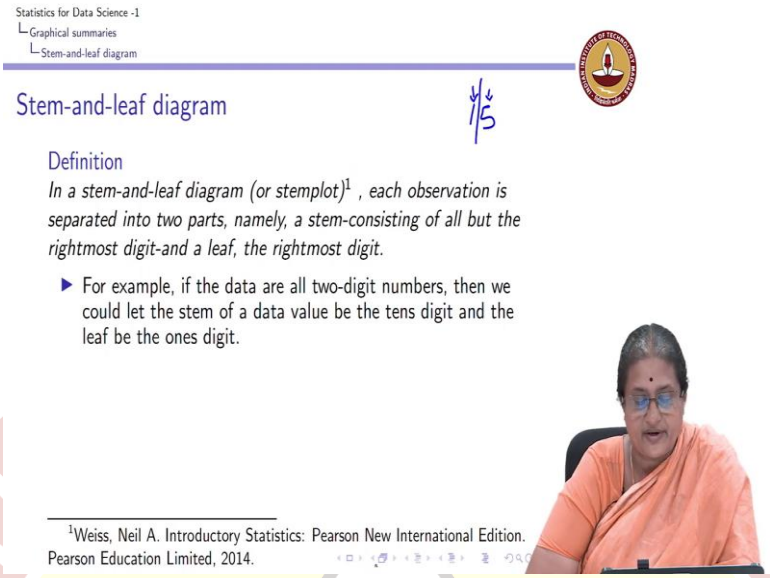
Again, go back and see that if I what I specified here is the bucket size of 10. If you list auto, it will create between 30 and 90 it decides on what is the appropriate class interval. But here I want a bucket size of 10.

(Refer Slide Time: 27:33)



So, you can see that if you increase the bucket size, it is not giving a proper. So, the size of your class interval it actually matters and it is good to have a reasonable size and in this problem I have chosen a bucket size of 10.

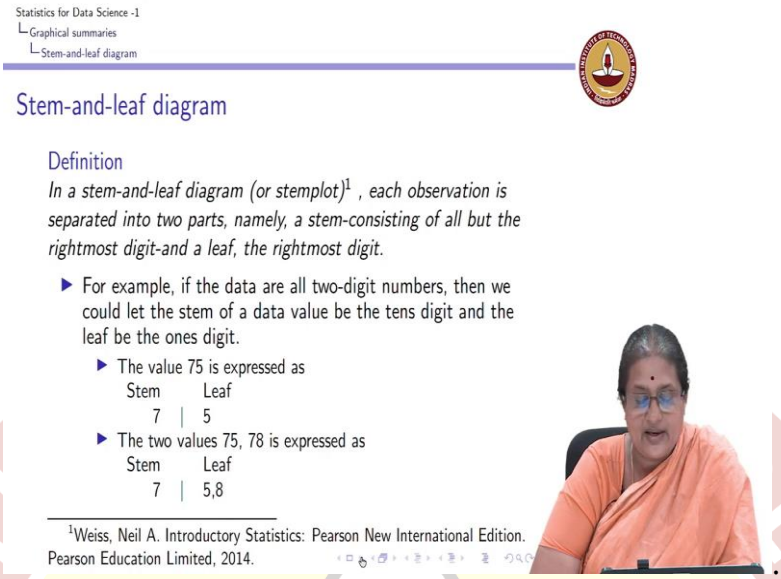
(Refer Slide Time: 27:53)



The next important graphical summary which we are going to discuss is, popularly referred to as a stem and leaf diagram. In a stem and leaf diagram also popularly referred to as a stem plot each observation is separated into 2 parts; namely a stem consisting of all but the right most digit and a leaf which has the right most digit.

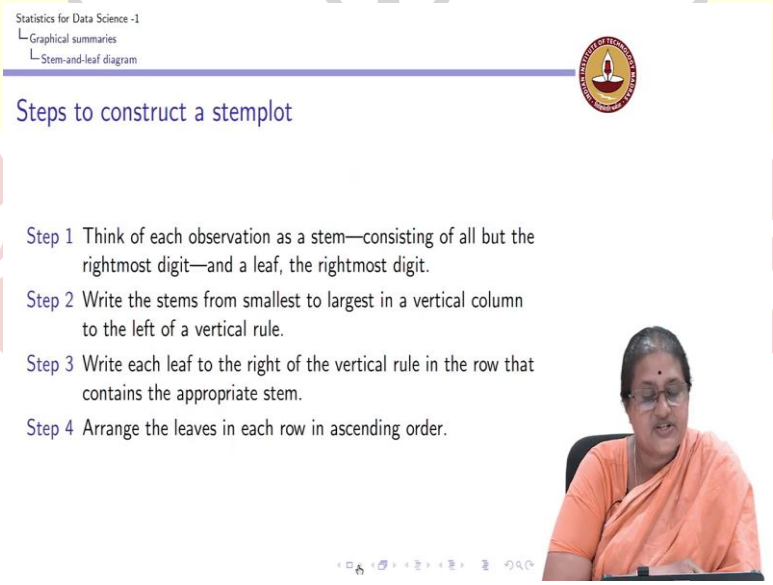
Now, what do we mean by a right most digit and a left most digit? So, in a stem plot for example, if I have 2 digit numbers then we could let the stem of the data to be the value of the tens and the leaf to be the ones. For example, if I have 15 the stem would be 1 and the leaf would be 5 that is what I mean by a stem of a data and leaf of a data.

(Refer Slide Time: 28:59)



The value 75 can be represented as 7 and 5. Now, if I have 2 values then for 75 stem was 7, leaf was 5. For 78 the stem is again 7 and leaf is 8 and I can represent this value by just having 1 stem and 2 leaves. So, this is called a stem and leaf plot.

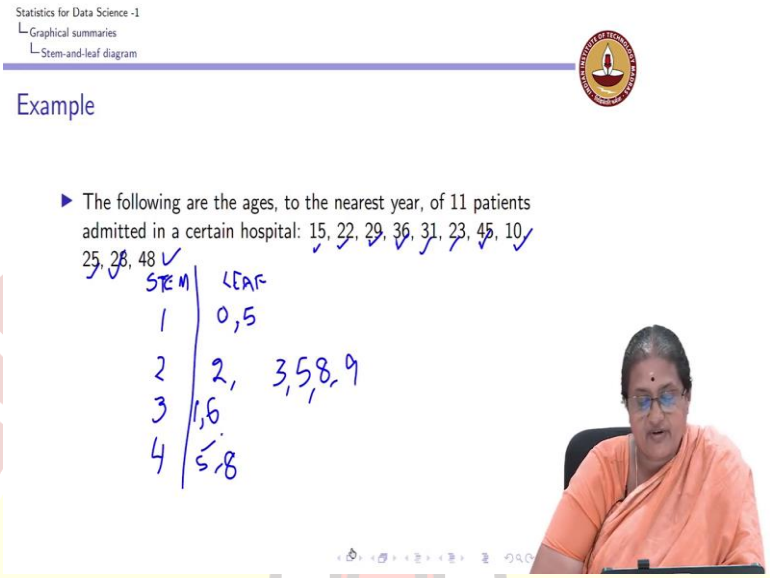
(Refer Slide Time: 29:33)



A stem and leaf plot how do I construct it? So, look at any observation as a stem consisting of all but the right most digit and a leaf has the right most digit. Write the stems from the smallest to the largest in a vertical column to the left of the rule. Write

each leaf to the vertical and arrange the leaves. This is how we construct a stem plot we apply that to an example.

(Refer Slide Time: 29:59)

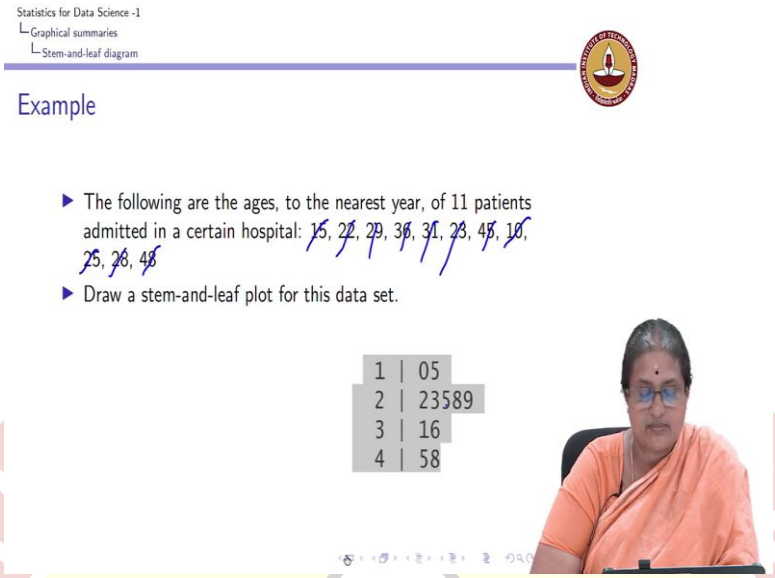


I have the ages to the nearest year of 11 patients admitted in a certain hospital. The ages are 15, 22, 29, 36, 31, 23, 45, 10, 25, 28 and 14. So, let us go back and see; what are these steps. The first is think each observation as a stem and a leaf. Write the stems from smallest to largest. So, if I write the stem the smaller stem is 1, 2, 3 and 4.

These are my stems ok; these are my stems. Now, this is my first step. I have second step where I have written the stems from the smallest to largest in a vertical column. Write each leaf to the right of the vertical rule. So, if I come here these are my leaves. 5 comes here, 2 comes here. So, let me 9 is again here, 6 is here, 1 is here, 23, 3 is here, 45 is here, 10 0 is here, 25 is here, 28 is here and 48 is here.

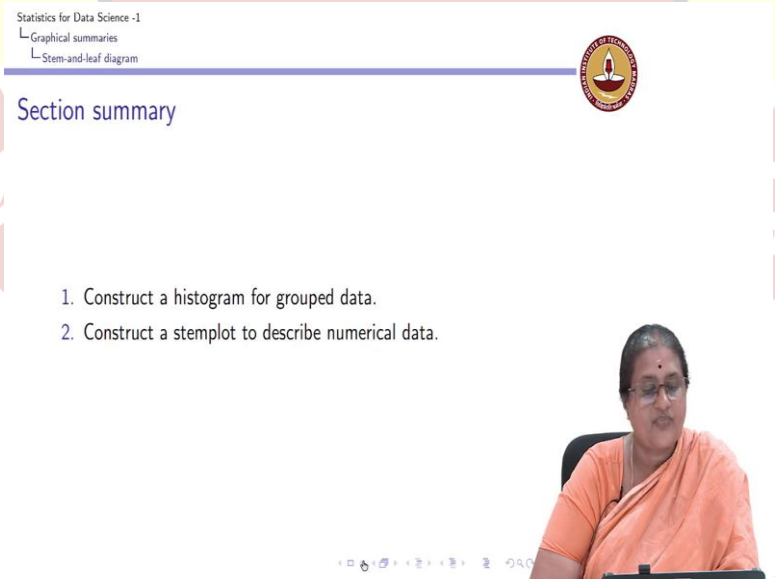
So, this is what I have done in step 3. We have written each leaf to the right of the vertical rule. The last step is arrange the leaves in ascending order. So, the way I can arrange the leaf here is I get a 0 and 5 that is the arrangement. Here this is going to be 2, 3, 5; 8 and 9; this would be 1 and 6 and the last is going to remain as it is to have a final stem plot as the following.

(Refer Slide Time: 32:15)



Which is 1; 0, 5; 2; 2, 3, 5, 8, 9, 3; 1, 6, and 4; 5, 8. 10 corresponds to 10, 15 corresponds to 15 then I have a 22, I have a 23, I have a 25, I have a 28, a 29, 31, 36, 45 and 48. So, this is how we construct a stem plot and we can get a stem plot many of the statistical packages give you a stem plot.

(Refer Slide Time: 32:57)



So, what we have learnt in this section so far is, how do we construct a histogram. First of all we learned as to how do we construct a frequency table both when I have single value discrete values are very small number of data. And, then when I have large number

of values where I group the data. And, once I have the frequency tables for my group data, we saw first how to construct a histogram by constructing class intervals of equal length and constructing a stem leaf plot for my data. This is what we have learnt so far.