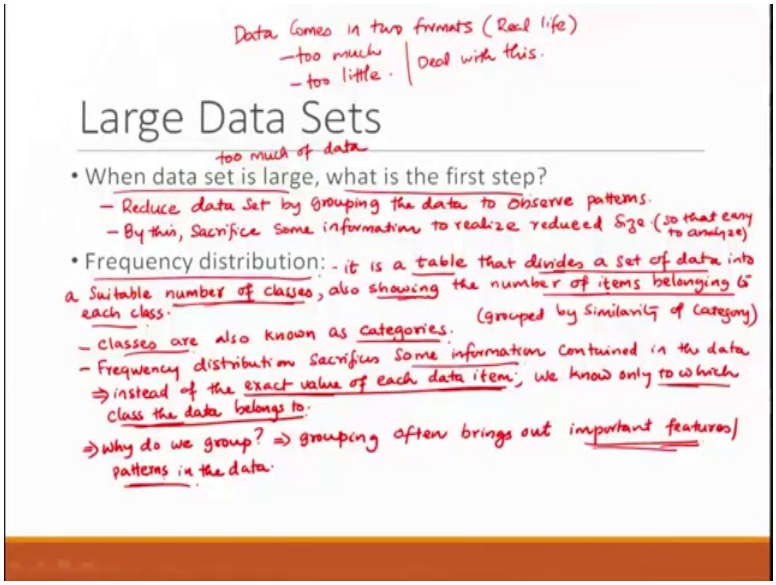
Practitioners Course in Descriptive, Predictive and Prescriptive Analytics

Prof. Deepu Philip
Dr. Amandeep Singh Oberoi
Department of Industrial and Management Engineering
Indian Institute of Technology, Kanpur
National Institute of Technology Jalandhar

Lecture – 10
Frequency Distribution & Histogram

Good evening ladies and gentlemen. Again, welcome to the course practitioners approach in business analytics descriptive, prescriptive and predictive analytics. And today we are trying to get into the new topics of descriptive analytics specifically and how do we describe the data and the importance in the in studying the decision making process or doing the decision making. And the tool that we are going to do today is the frequency distribution and histogram and it is also an important tool in the descriptive statistics.

(Refer Slide Time: 00:52)



So, as we said earlier most of the time data comes into 2 format, unless it is a text book that is in real life, the format is too much or too little ok.

So, then we have to methods or mechanism to deal with this ok. So, today what we are going to do with this? We are going to look into, when the data set is large, if you have a large data set or too much of data too much of data then what is the first step or what we

do? Ok one of the way a simple way to do it is reduce data set by grouping the data to observe patterns ok. So, the idea here is that you group the data and then the grouping process you will end up; you know reducing the data set. The second thing is by doing this by this sacrifice some information to realize reduce the size ok. So, what one you have reduce the data then you are able to so, that they can. So, that easy to analyze ok.

So, one search tool what we talked about is a frequency distribution ok. And what is frequency distribution? Frequency distribution it is a table it is a table with a tabular form that divides that divides, a set of data into a suitable number of classes ah. Also showing the number of items belonging to each class to each class, the idea here is that it is a table or it is a tabular that divides the set of data it divides a set of data into suitable number of classes.

We will talk what are the classes, but it divides into classes or groups. So, all of you are in the class of this prescriptive analytics what does; that means, because all of your taking this course you have the similarity in that. So, classes means a group of similar items and it also shows the number of items that belongs that each class that is also what is being shown ok.

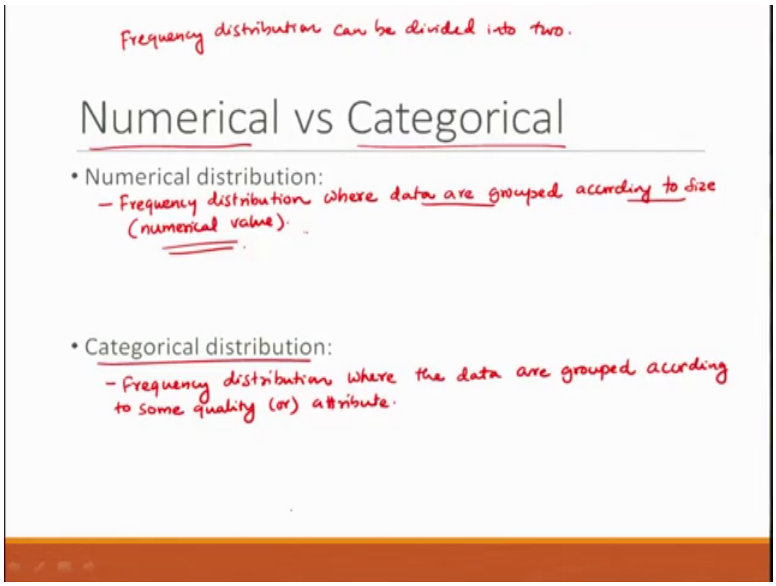
So, in a way classes are also called as also known as categories ok. And the importance of that frequency distribution frequency distribution sacrifices some information some information contained in the data the data ah. Which is what is that one which is implies instead of the exact value of the data exact value of each data item each data item we know only to which class the data belong to belongs to get ok.

So, what happens here is it sacrifices some information what is sacrifice it sacrifice the exact value of each data item and instead it is providing you which class the data belong to. Each individual values you do not know, but you know where it actually belongs to which class it belongs to, which group it belongs to, or which category it belongs to. . So, why do we group into categories? Because grouping often important features in the data, important features slash patterns in the data ok.

So, the main reason why we do this why do we group because to bring out important features or patterns of the data is one of the reasons why we group. So, this is a tool for reducing the data set and as I said earlier, you device data set into suitable number of classes. And showing the number of items that belongs to which class and classes are

also known as categories. So, they are grouped by similarity, similarity of category ok. And it is sacrifices some information, it sacrifices exact value of each item and. Instead it tells you which class the data belongs to. The reason why we do this because by doing this we can identify the important features or patterns in the data.

(Refer Slide Time: 08:04)

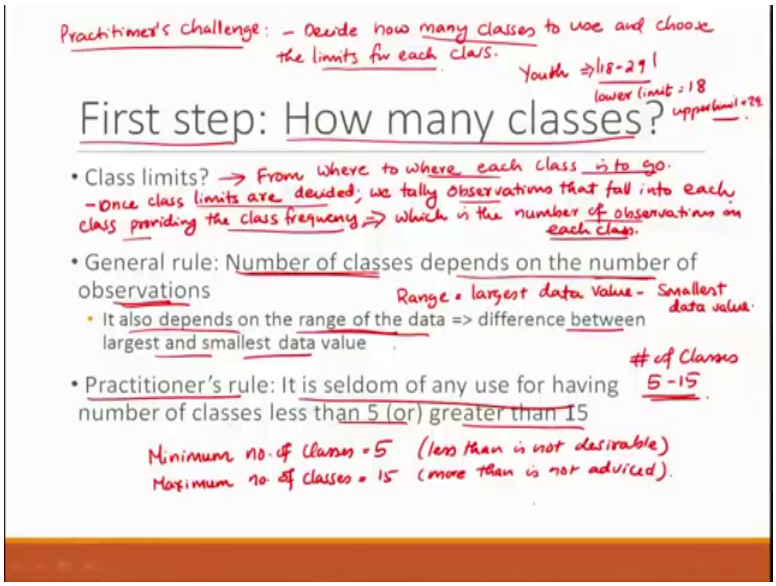


So, if the frequency distribution can be divided into 2 ok. Number 1 is numerical other one is the categorical. So, what is a numerical distribution is say it is the frequency distribution distribution.

Where data are grouped data are grouped according to size to size or what we can call it as the numerical value ok.

Here the data is grouped according to it is siz or the numerical value. What is the categorical distribution? It is again another frequency distribution frequency distribution. Where the data values where the date are grouped according to some quality or attibute. So, here we are looking at some of the attribute of the data. Like for example, I can divide the color we can divide the data according to some group by the similarity of color in something. stuff like that where is here you can think about grouping in the form of numerical. So, in this class, we are go focus more on the numerical distribution, not too much on the categorical distribution. Categorical distribution is used for other aspect and this outside the scope of this 20 of course.

(Refer Slide Time: 10:23)



So, the first step how do we do this the first step the first thing that you need to decide is how many class. So, the question here is decide or instead of this practitioners challenge ok. One of the things that practitioners are do here is ok, decide how many classes to use? And choose the limits of limits for each class in making the frequency distribution. The practitioner how to decide how many classes to use? And then also decide the limits for each class ok.

The class limits what does it mean? This means that from where to where each class is to go? So, the idea is this is somebody says that young or youth is an age between 18 to 29. So, then 18 to 29 is a class that class is called youth and the class limits are 18 years of age. So, the lower limit in this case will be 18 and upper limit will be 29 like this.

So, that is what the class limits are, then what do we do? Once class limits are decided we tally a observations that fall into each class each class, providing providing the class frequency which is an. So, what is class frequency? Which is the number of observations in each class in each class? So, once you decide where from where to where each class to go? Or we decide the upper limit and lower limit of each class.

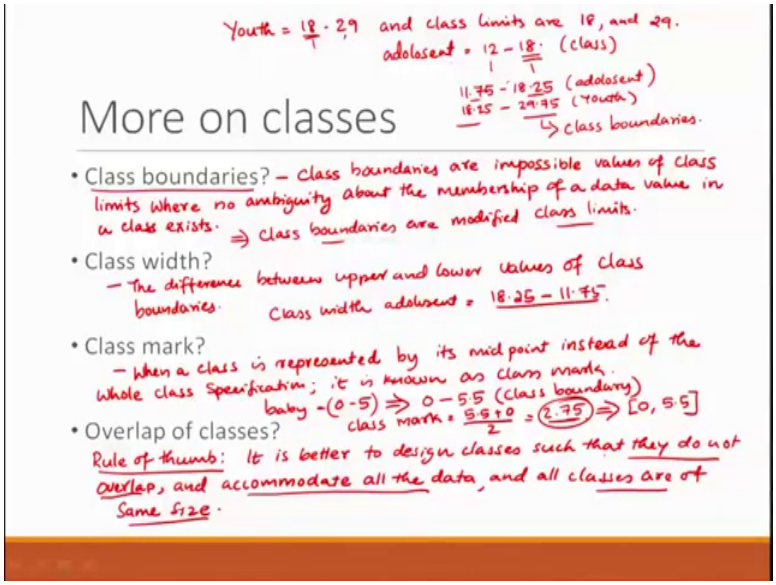
So, then from once class limits are decided, Then we tally the observations that fall into each of these class and by doing that we get the class frequency class frequency means the number of observations that fall into each class so; obviously, you how to make a decision. So, general rule there are many rules are available the number of classes

typically depends on the number of observations that is rule number 1 and roll number 2 it also depends on the range of the data ok. Range of the data is the difference between the largest and the smallest value ok. So, range equal to largest data value minus smallest data value ok. That is what the range is all about.

So, so the 2 things the number of the classes depends on the number of observations how many data values and the range of the data values. Ok the practitioners rule is that even though people say lot of other things the typical rule is that there is of no use it is seldom of any use for having the number of classes less than 5 or greater than 15. So, the number of classes it typically between 5 to 15 ok.

This will help you the. Know the best way to identify this. So, it can think about is minimum number of classes equal to 5, less than is not desirable ok. Similarly, maximum number of classes equal to 15, more than is not advisable. So, as practitioner you trying to decide the number of classes between 5 and 15 which is less than 5 your grouping too many things and you are losing lot of information if it is greater than 15 that you are not been able to identify as appropriate patterns out of that.

(Refer Slide Time: 16:10)



So, let us think about some more other definitions of this ok. Which are related to this, first one is called as a class boundaries and we talked about earlier our class limit ok.

So, we defined the class youth equal to 18 to 29 and class limits are 18 and 29. So, then what the class boundary? Class boundaries are impossible values of class limits where no ambiguity about the membership of a data value, in a class exist exists. So, we think about it if you let us think about it as adolescent; adolescent is the age from 12 to 18 let us say that and this is one class and the youth is 18 to 29. So, the question; obviously, is where would you put 18 would it adolescent or would you put it in the youth

So, in this case in the class limits the this case is 12 one 18 is the class name is here is 18 to 29 is a class limits. So, this creates a problem. So, sometimes we can make it has 11.7 5 to 18.25. We can say this is the adolescent and 18.25 to 29.75 is youth. Let us say you put it that way. If that is the case then you can say that these new values these new revised values which are impossible values for that matter. Because you cannot really the data that you collected does not have that type of values you only have 11 years 12 years only the absolute value of age is what you have.

Then these will ensure that 18 will fall here it removes the ambiguity. So, these are called as class boundaries ok. So, class boundaries are class boundaries are modified class limits when you do frequency distribution you do change class limits into class boundaries. Then the second one is class width. What is class width? The difference the difference between upper and lower values not of class limits, but it is values of class boundaries ok.

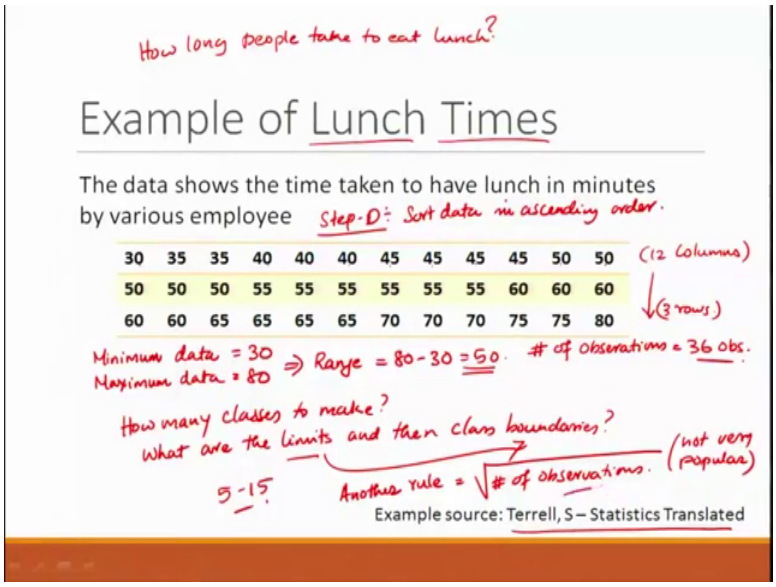
So, the class width of the adolescent will be equal to 18.2 5 minus 11.7 5, this will be the class width ok. Then class mark what is the class mark, when a class when a class is represented represented by it is midpoint midpoint. Instead of the class, instead of the whole class whole class specification, it is known as class mark. For example, let us define a class called baby as the age from 0 to 5 ok.

So, then this is the class limits class boundaries you could probably say the 0 to 5.5 ok. As the class boundary, the class mark will then be the midpoint which will be class mark equal to 5.5 plus 0 divided by 2 plus 5.5 by 2 will be that is 2.75 ok. So, if you when you say that the 2 .7 5 represents the class. So, in your if your drawing a diagram or something then you can basically just use 2.7 5 instead of the 0 to 5.5. So, this replace is this 05.5 specification of the class. So, 2.75 represent this particular class ok.

Class overlap what happen in class overlap if the there are overlap classes, but also rule of thumb ok. Let me say rule of thumb or general rule is that it is better to design classes. Such that they do not overlap and accommodate accommodate or the data and all classes are of same size.

This is what we call as the desirable design of classes. So, we can say that yes you can design a class with an overlap, but it is better to design classes such that they do not overlap and you can accommodate all data do not loose any data and classes are of the same size. So, the size of the class the width of the class that would be the similar you can see that the baby class is a one bed then the adolescent is of another bed and then the youth is of different bed. So, instead of doing that it is advisable that you make all the classes of the similar way.

(Refer Slide Time: 23:46)



So, what are the best way to look into solve this kind of or understand this is? Like let us look into what we call as the, an example problem, what we are doing here is looking at the example of lunch times how. So, question here is how long people take to eat lunch? Ok and people collected data and this data is given to us data shows the time taken to have lunch in minutes. So, this is in minutes by various employees.

So, this is done in some factory and the source is the steven terrell books statistics translated this is the example actually take from that book. So, here you can see that the

minimum data value is 30 maximum data value is 80 you can this is arrange in the ascending order ok.

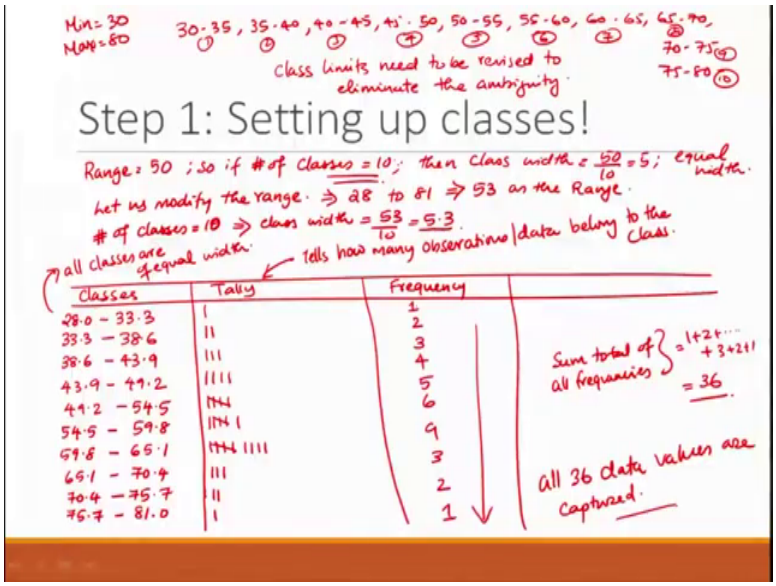
Sometimes when data is not available in the ascending order it is advise that you sort in ascending order. So, step one I will say or step 0 sort data in ascending order this is important ok. So, the minimum and maximum data is given to you as 80 and 50. So, which implies the range of the data is 80 minus 30 which is equal to 50.

So, one is the range of the data the frequency distribution the number of classes depends upon the range of the data and also on the number of observations number of observations how many are there. So, there is 1 2 3 4 5 6 7 8 9 10 11 12 12 observation. So, 12 columns and this way you have 3 rows. So, 12 times 336 so, 36 observations you have ok. So, these 2 things the choice the now the question here is your practitioner choice is, how many classes to make? And second is what are the limits? And then class boundaries ok.

So, you do class boundary. So, that ambiguity the limits get translated to the class boundaries. So, ambiguity is removed and the rule of thumb that we said is 5 to 15 there is another rule is the square root of number of observations ok. This is not very popular only used in the case of small data sets.

So, we will not follow this, but some people do use square root of the observations to decide the number of classes if this is more than 5 and 15 then you round it 15. If it is less than 5 then you round upto 5 that is the way people do this, but is not a very popular rule, but I am still mentioning that this rule is available ok.

(Refer Slide Time: 27:16)



So, the first thing we need to do is, ah. So, since we know that the number of classes. So, the range was found out to be 50 and so, if number of classes is equal to 10 then class width will be 50 by 10 equal to 5 equal widths. So, you would have 10 classes of the equal width.

So, we think about it. So, then the minimum value starts it is the minimum is 30 maximum is 80. So, your classes will be 30 to 35, 35 to 40, 40 to 45, 45 to 50, 50 to 55, 55 to 60, 60 to 65, 65 to 70 then 70 to 75, 75 to 80. So, you will have 1 2 3 4 5 6 7 8 9 and 10 classes , but; obviously, when you do this and if you look at the data values you will see that where will you put 40. Let us take the data value of 40 will the data value of 40 go here or will go here

So now, you can see that the class limits need to be revised revised to eliminate the ambiguity the ambiguity ambiguity of where the 40 belongs to or where 45 belongs to like that. So, to do that there is one way is that since we taken the number of classes of 10 for easy division and other things. one way to do it is I can always or we can always say that let us modify the range ok. The range now start from 28 to 81 giving result to 53 of the range. So, that is given by 81 minus 28, 53 number of classes equal to 50 is equal to 10 which implies class width equals 53 by 10 equal to 5.3.

So, since you have a decimal point and all your data values every data value here is an absolute value without any decimal point. By using this decimal point, you can ensure

you can eliminate this ambiguity ok. So now, with this n you can have a new classes. So now, our next step is to create the frequency distribution.

So, I am going to do it like this I will said as class classes then the classes will be I will start at 28.0 and I will move to if you add 5.3 to this it will become 33.3 then the next one will be 33.3 it will move to 38.6 then from 38.6 it will move to 43.9 then 43.9. It will go to 49.2 then fort 9.2. If you add 5.3 with it you will actually get it as 54.5 52.5 if you add 5.3 with it you will get 53.8 then 59.8 you add 5.3 you get 65.1 and 65.1 you are 5.3 you get 70.4 and 70.4 you add 5.3 you get 75.7 and 75.7 you are 5.3 you get 81.0 .

So, this is the classes that you have ok. Now, one way to do the frequency distribution the simplest way to do the frequency distribution is you do what we call as the tally tally mark as ok. So, what we do if you go back to the data and see between the range of 28 to 33.3. How many data values are there? So, if you go back and we check we have 30 is there that is only one data. That we have.

So, then we come here and we say value of one from 33.33 to 38.6. How many data values are there? So, that is 35 to 35s of there 40 is the outside range there is no ambiguity in that. So, we basically say 2 than 38.6 to 43.9 we go back we will have 1 2 3 3 40s between that value. So, we go here and write 1 2 3 then 43.9 to 49.2. So, between that only 45s will come 1 2 3 4 4 of them. So, then we basically do 1 2 3 4 then 49.22 to 54.5 go back and say that will only you get is 50 1 2 3 4 5 5 50s ok.

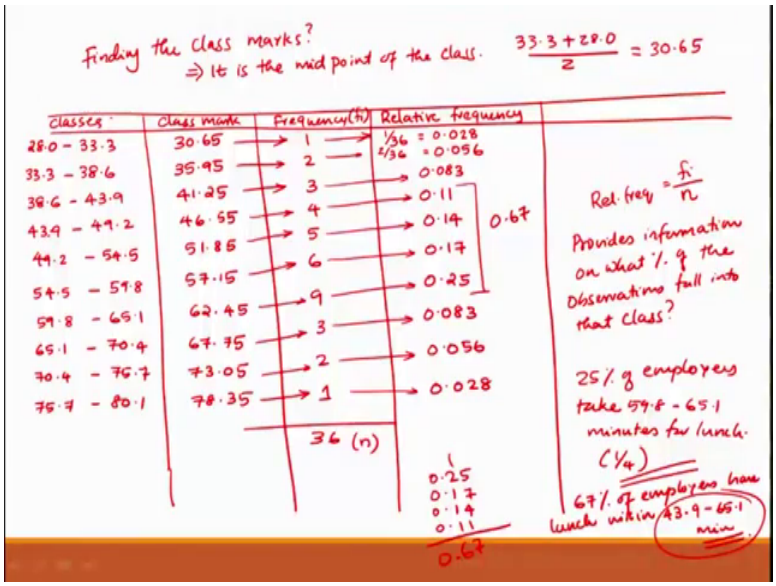
So, we will go is it will be 1 2 3 4 5 then 54.5 to 59.8. So, that will be if you go back you will actually get only 55. So, 1 2 3 4 5 6, 6 55s are there 60 will not come because it is outside that. So, we go back. So, that is the 6 1 2 3 4 5 and 6 then 59.8 to 65.1. So, 59.8 65.1 ok. So, we are going there. So, 59.8 and 65.1 means both 60 and 65 will come in. So, 1 2 3 4 5 6 7 8 9 9 of them, then we go back and do the tally 1 2 3 4 5 1 2 3 4 that is 9 of them similarly if you look at this you have 3 of these 3 have 2 of these and one of these. So, this tally is what we just call us know and then you write that. So, here this tally tells how many observations or data belong to the class ok. So, then the next one we can create out of this is called as the frequency ok. The frequency is more like counting this numbers one this is 2 this is 3 near 4 near 5 is 6 then your 9 then your 3 then your 2 then your one ok.

So, this is your frequency. So, if you some all the frequency so, if you say that sum total of all frequencies if you do that that is equal to 1 plus 2 plus all the way up to 3 plus 2 plus 1 do that that gives you how many observations? it gives you total of 36 observation of this ok. So, this is the sum 36 so, this 36 observations 3 plus 2 5 plus 5 plus 1 6 6 plus 9 15 20 26 30 33 35 36.

So, 36 observations you get solve the observations are put it into each one of the class marks. So now, if you look at this you can see that the individual observation. We lost the individual values these are no longer there instead what we have is which class or how many of them will belong to each class is what we have ok.

So, one another thing that we can do out of this is quickly that can be done out of this is the second step is finding the class marks ok.

(Refer Slide Time: 36:39)



So now let us see finding the class marks. So, what is the class marks how do we find it it is the midpoint of the class ok. In our case, we can see that all classes are of equal width are of equal width and all the data points. Here all 36 data values are captured ok. So, in that regard we can say that we have basically fulfill the rule of thumb that be predicted which basically says classes should be of equal width and classes should not have any ambiguity. And all the value should be captured.

So, you think about it this is a good frequency distribution that is being built. Now, if you look into this as I said find in the class mark is the midpoint of the classes. So, let us draw the table little bit more nicely this time we want to the tally and other stuff. So, we will say class width or not of class width classes then we have our class marks class mark then we have is our frequency let us do so far.

So, the classes where once again 28.0 to 33.3 33.3 to 38.6 or we were adding 5.3 which has the class mark than 38.62 43.9 43.9 to 49.2 and 49.2 to 54.5 54.5 to 59.8 and 59.8 to 65.1 and 65.1 to 70.4, 70.4 to 75.7 and 75.7 to 80.1 ok. And the class mark ideally is the midpoint class mark is the midpoints. So, the first-class mark will be 33.3 plus 28.0 divided by 2 which should ideally give you the number of 3.65 ok. So, the class mark here is 30.65.

So, when you said 30.65 that literally means it represents the class 28. . So, it is represents this class similar the other one is 38.6 plus 33.3 which will give you 35.95 and similarly 43.9 and 38.6 if you find the midpoint of it will be 41.25 similarly 49.2 143.9 if you do this you will get 46.5 5 then it is 51.85, 57.115 62.45 then 67.75 73.05 and 78.35 ok. And the frequency of this is as I said earlier 1 2 3 4 5 6 9 we have 3 2 and 1 ok. So, this is the frequency and the sum of the frequency was studied up to 36. Then the next concept that is applicable in part of this is what we called as the relative frequency ok. So, relative frequency is calculated I will be writing here the relative frequency equation is equal to ah. So, if you call this as f sub I individual frequency f I divided by n and this is your n ok.

So, the first relative frequency will be 1 divided by 30 6 I will come to 0.028 the next will be 2 by 36 ok. So, that will give you 0.0 5 6 then the third one 3 by 36 will give you 0.083 then the 4th one will give you 0.11 then you will get is 0.14 then 0.17 0.2 5 0.0 8 3 then you have is 0.056 then 0.028. These are the relative percentages in a way think about this is actually tells you that this class the 62.45 class mark for 59.8 to 605.1 contains 25 percentage of the data values in this. So, the relative frequency kind of tells you the relative frequency provides information on what percentage of the observations fall into that class ok. So, that information is made available to us with the help of you know relative frequency.

So, now with this data one of the ways to think about it is you did it. So, what did actually help? So, if you want to find out in this particular time where will be the maximum amount number of time people will be taking lunch you want to find out of this time period this particular time period I know it is not kind of straight forward to read. So, I will kind of do this.

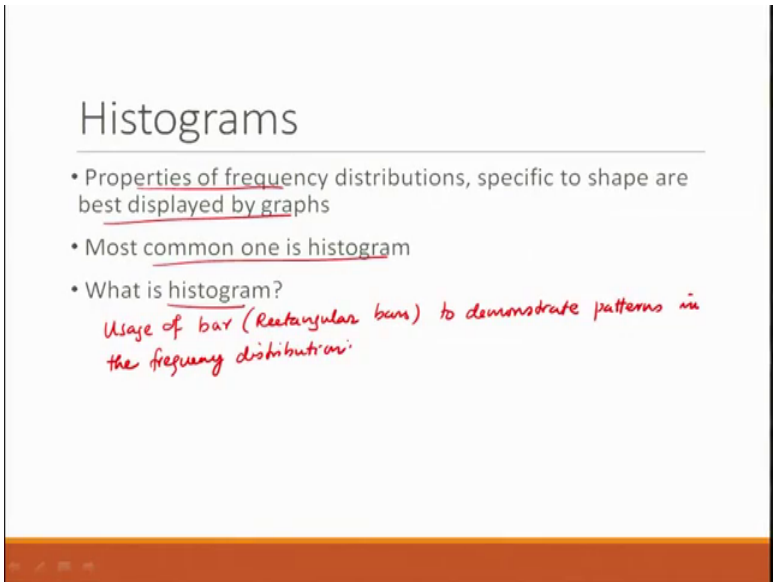
So, that you can understand sorry my writing is not very a in that particular order. So, you know what values we are talking about you can see that this particular case let me put it this way kind of wrote in a different goofy fashion. So, you will understand what I am trying to do in this case you can see that about 25 percent of the people one thing is 25 percent of employees take 59.8 to 65.1 minutes for lunch .

So, it is one 4th of the total crew actually take somewhere close to 60 between 15 about one hour for the lunch there are people who finish the lunch early also, but very small number and one other way to look at it is you can start cumulating this frequencies also, but we are not going to work on that we can do that can do many fancy things with this stuff.

So, that allows you lose exactly what was the lunch times of individual people in this you do not know that, but you get this new vital information that yes one 4th of the total crew wants this take lunch in that then you also know that 17 percentage of them take the lunch between 54.5 to 59.8 time period and then 14 percent of them takes typically between 49.2 and though you do not know the exact values of it, but you kind of know. So, you can see that this much time period if you look into this you have this which is you know point 25 you have 0.25 then you have 0.170 14.1. You think about this that is for 5 17 1 tripple 3 4 5 6 6 ok. So, about 67 this is point 67. So, about you can say that 67 percent of the people take lunch between one other way to think about it is 67 percentage of the people 67 percentage of employees have lunch within 43.9 to 65.1 minutes ok.

So, this is an important information the majority of your people this is how much time they will take to have lunch. So, that information is of vital importance to us in this regard now; obviously, the question that we have to ask is you know.

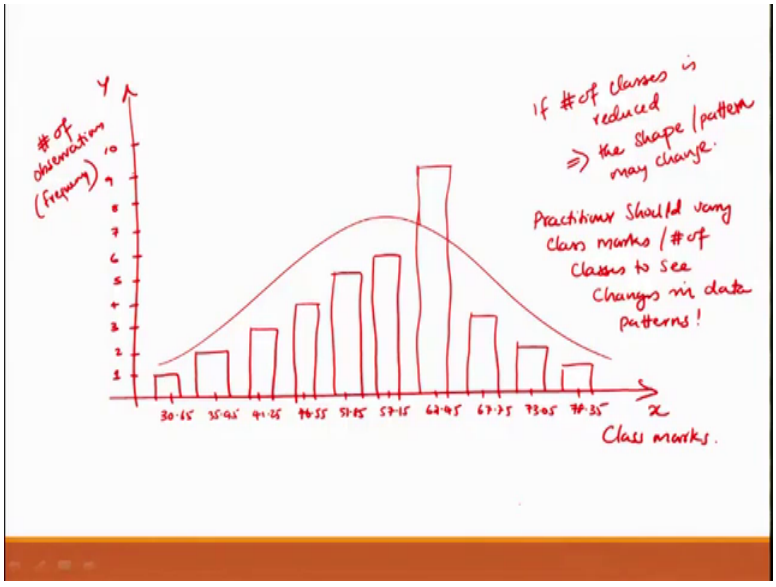
(Refer Slide Time: 46:36)



Why is this, an important thing? The important thing is once you have this frequency distribution created then the next the easiest way to do it is this frequency distribution can be best displayed with the help of a graph. And the most common graph is what we call as a histogram and what is the histogram? Histogram is a usage of bar or rectangular bars to demonstrate patterns in the frequency distribution ok. Many of the software excel and other things will do this for you.

So, what you will try to do is you will try to plot a histogram here ok.

(Refer Slide Time: 47:34)



So, there are multiple ways you can plot here is your X axis here is your Y axis and X. We will say it was the class marks for the time being can you use the class width also, but this is where it is easy for you 1 2 3 4 5 6 7 8 9 10. So, you have a 10 classes . So, I will say it as 3.65 this will be 35.95 this will be 41.25 46.55 then you have is 51.85 then is 57.15 and 62.45 then 67.75 then 73 point of 5 and 78.35. So, these are the class marks of the 10 classes you have and you can say let us put the counts here. So, we will say 1 2 3 4 5 6 7 8 9 10 stuff like this.

So, for the first class we had a frequency of one and then the second class the frequency was 2 in the third class the frequency was 3 ok. I am trying to draw this properly , but you will understand 3 then the next one is 4 ok. So, we draw it as the 4 then we had is 5 apologies for the ugly drawing 5 then we had is 6 then we had is 9. So, it is like end of right here when you draw actually you draw with software or something then we have is a 3. So, somewhere here then you have say 2 somewhere here then you have you say one somewhere here. So, if you look at that this kind of a graph.

So, here is where the number of observations or what you can call it as frequency ok. So, you can kind of see that there is kind of a you know some people can will probably say this kind of shows bell shaped curve kind of a thing or the data certain to show. You know normal distribution or something like that you can also say that you can also if this number of classes if you reduce a number of classes to.

Let us say if number of classes is reduced then it implies that the shape or pattern may change ok. So, one other things has the practitioners do practitioner should vary class marks or number of classes to see changes in data patterns ok. This is something that the practitioners is expected to do because by doing this the practitioner will be able to will be able to find which is the best number of classes available for this ok

So, with this today, we complete our lecture and I hope that you understand the concept of histograms and as well as the frequency distribution. And how frequency distribution can be use histogram can be used to pictorial depict the frequency distribution. And I advice you guys to use Microsoft excel to do this because excel has a very graphing tools and are can also do this for you, but it is very easy to do this in excel. And hope that you guys learns something out of it and see you guys in the next class.

Thank you.