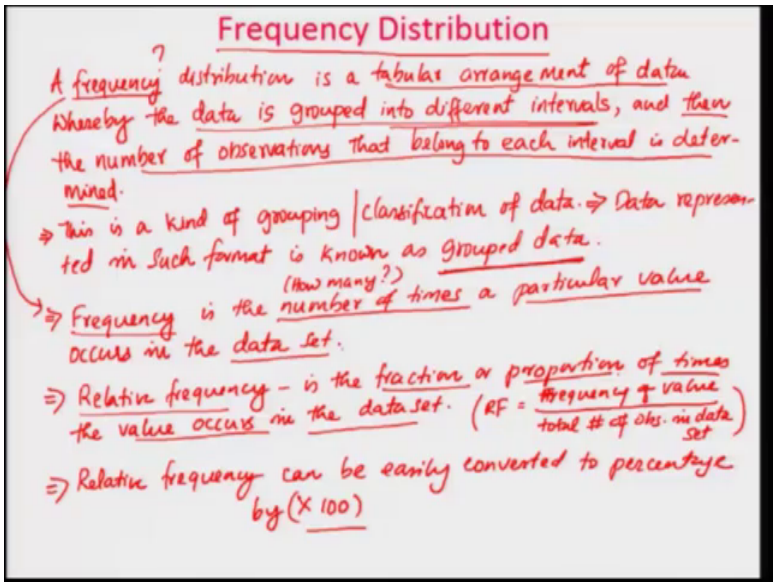
Simulation of Business Systems
Prof. Deepu Philip
Department of Industrial & Management Engineering
Indian Institute of Technology, Kanpur

Lecture – 15
Frequency Distribution

Welcome back to the continuation of the lecture of probability and statistics for Simulation of Business System course. We been going through the application of the basic criteria of probability in statistics and those kind of basic tools that are required to do simulation experiments, experiments with simulation models and to analyze its data. So, we have been going through various things and one of the thing major thing that we finished the so, far is the Pareto diagram. And then we also decided that we will do the next tool which is called as Frequency Distribution.

(Refer Slide Time: 00:54)



So, if you look here the frequency distribution today, what we are going to discuss it. And, we have to find out and we will do a small example or to demonstrate to you how to do frequency distribution and stuff like that. So, by definition a frequency distribution is a table, a frequency distribution is a tabular arrangement of data; is a tabular arrangement of data or it is a table; table with data in it data arranged a particular format data arrangement of data whereby, the data is grouped into different intervals; whereby, the data is grouped into different intervals and arrangement of data whereby, the data is

grouped into different intervals. So, first thing is a tabular form or is a table of data where the data is group into different intervals. And, then the number of observations then the number of observations number of observations that belong to that belong to each interval is determined.

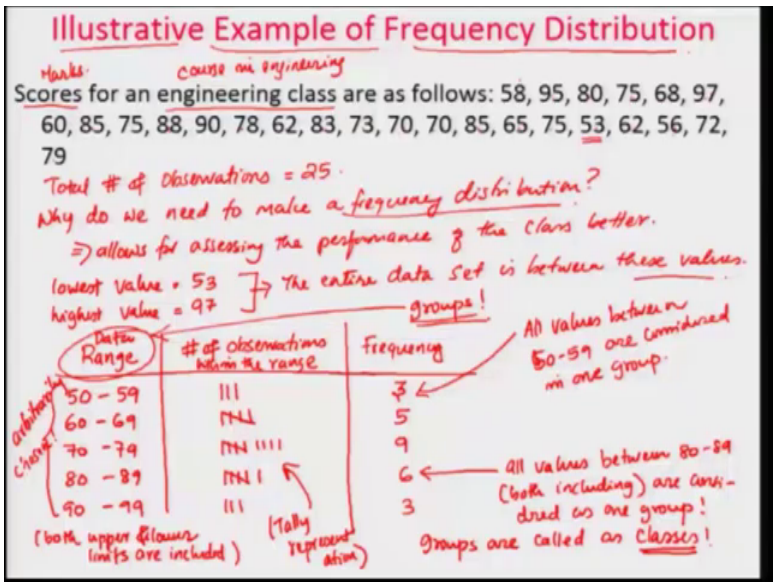
So, you arrange the data in a tabular format, a tabular arrangement of data where data is grouped into different intervals. So, you have different intervals. So, which means that you are taken different intervals of data and everything within that interval is counted as belonging to the same category and then the number of observations there belong to each interval is determined. So, which kind of in a way this is a kind of grouping or classification of data. Now, implies data represented in such format data represented in such format is known as grouped data. So, we are talking about when the data is it is kind of a classification or grouping of data. And hence if you do this, when you use different intervals grouping data into different intervals then such data is called as grouped data.

And then there is another term that we talked about this is frequency. So, what is frequency? So, the frequency distribution what is this frequency. So, the frequency that is what we are defining here, a frequency is the number of times is the number of times a particular value occurs in the data set; occurs in the data set. So, how many times a particular value number of time so, this is how many times you are answering the question how many, how many times a particular value occurs in the data set that is called as frequency.

Then when we talk about another term which is called as relative frequency I have discuss this in the previous case especially in the Pareto diagram and I did not define it at that point I told you guys that we will define it when we reach here. And, relative frequency it is defined as is the fraction or proportion of times the value occurs in a data set occurs in the data set. So, what we are saying is the fraction of times the value occurs in a data set. So, typically that is as it is given by relative frequency can be given by the frequency of value divided by total number of observations in data set. This is another way of looking into this, the mathematical way of looking into it. Then we can talk about it as relative frequency a multiplied with 100 in the previous one.

So, relative frequency can be easily converted to percentages percentage, by doing what? By multiplying 100. So, because, it is a fraction say proportion fraction or proportion we can multiplied with 100. And, we can easily find what is the percentage of the time or what percentage of the total data set this particular data value occurs in the data set. That can be calculated in percentage by multiplying by 100 really quickly. So, we seen what is a relative frequency, also we seen what is a frequency also. We saw what is a frequency distribution and we also talking about what is a grouped data. So, this much concepts we have cover in this particular aspect.

(Refer Slide Time: 07:23)



So, now let us do a illustrative example or let us work through an example so, that you understand how the frequency distribution is created. So, the exam is a scores for an engineering class are as follows a particular engineering class. So, this is a scores in the sense, scores can be called as marks, engineering class is a course in engineering. So, here with the marks for the particular course in engineering it starts with 58, 95, 80, 75 like this and all the way up to 79. So, how many observations do we have? 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25. So, total number of observations equal to 25. So, the 25 observations as part of this.

So, now the question is why do we need to make a frequency distribution here? This is the big question, why do we need to make the make a frequency distribution? Why cannot we deal with individual values? The advantage of making the frequency

distribution is it allows for assessing the performance of the class better or it tells you how the performance in the class is spread out. How people can be grouped into people who are doing excellent in the class, doing average in the class, doing poorly in the class that kind of thing. So, what I am going to do is let us say that if you have no other way to group this.

So, let us start finding out which is the lowest data value unless the so, unless that finding which is the highest data value. So, the lowest data value of this is there is a 58 here and then we go through there is 56 here. So, the lowest value is lowest value is 53, I believe it is 53 and here is a 53 and then the largest value or the highest value is 90. So, 95 right here is a 97 write here, I believe it is a 97 I believe that is a largest value highest value is 97. So, the entire data set is between these values and we are not able to figure out other than what is the minimum and the lowest we are not able to identify anything else between this. So, I say you know for a simple way to do it is then there is people with this scores of 50, there are people with scores of 90.

So, let me say that I take rangers. So, let me create a table, simply a table with the range where I am saying that the first range is 50 to 59. Everybody who gets a score from 50 to 59 and the second range is 60 to 69, third is 70 to 79 next is 80 to 89 and the 90 to 99. Let us say I make the range or let me put is the data range and then I start doing the number of observations that fall within these ranges, within the range; let us say I am trying to calculated something like this. So, I will say from 50 to 59, I will count there is 1, there is 2, there is 3. So, 50 to 59 I have 3 observations, typically people do this is 1 2 3 they do this kind of a mechanism is called tally, we do tally representation of this.

Then 60 to 69, let this count how many are there is 1 2. So, by the way in this 50 to 59 both sides are included 50 and 59 values are included, 60 and 69 values are included data range. So, if this range where we can say both upper and lower limits lower limits are included. So, consider with that make. So, from 60 will be then included in the one. So, let us say there is a 68 1 2 then 3 4 and 5. So, there are 5 of them so, 1 2 3 4 5, 5 observations.

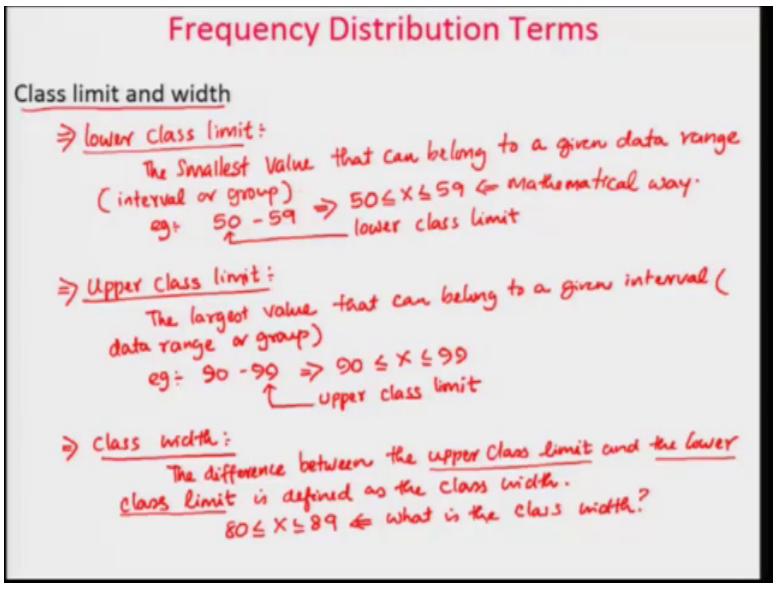
Then 70 to 79, if I do this there is 1 2 3 4 5 6 7 8 and 9 so, there are 9 observations. So, 1 2 3 4 5, 1 2 3 4, then 80 to 89. So, let us count how many 1 2 3 4 5, I believe it is 5 no 6 actually so, 6 1 2 3 4 5 and 6 and then 90 to 99. So, we will have 1 2 3 only 3 so, 1 2 and

3. So, this kind of representation of data this is called as Tally representation. So, representing data like this is called as the tally representation. So, then I can called as frequency, frequency is the numerical value of this tallies. So, there are 3 observations so, I have 3 frequencies within this range; 60 to 69 I have 5 then it is 9 then I have 6 and then I have 3.

So, what here we have done is so, if you look into this, this we are basically saying all values between 50 to 59 are considered in one group. Same way this, all values between 80 to 89 both including the other case also the both are including are considered as one group. So, one way to do it is we can basically looking at the data we can come up with some criteria or some way or some data ranges or some. So, this is also another way you can think about it thus, these are groups. I use the data ranges to develop the or generate what we called as the groups in this case.

So, these groups are sometimes there are technical terms that are used for this and we need to understand that terms and go forward with that. So, these groups are also called as classes; groups are called as classes, classes is the right term for groups, groups is the (Refer Time: 16:13) usage, but it is also called as the classes. And, everything when we talk about frequency distribution we are going to talk about the classes. So, let us look into the details of the classes and other things now.

(Refer Slide Time: 16:25)



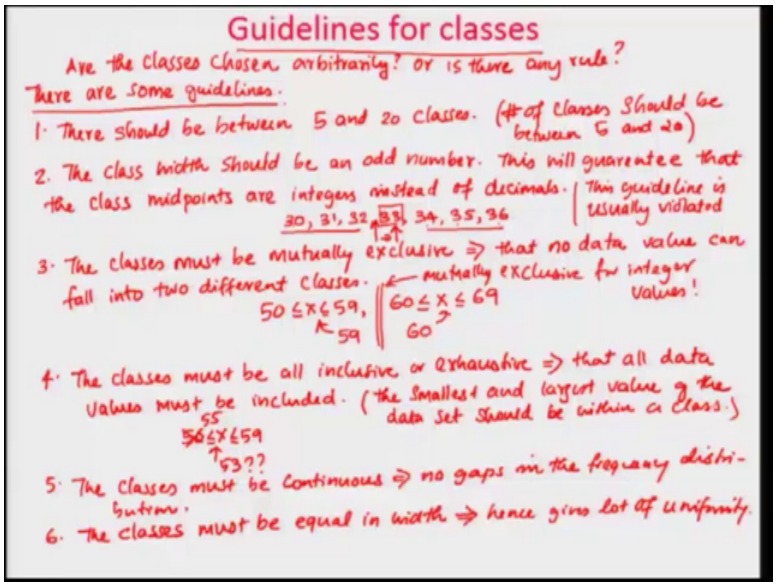
So, the first term that we need to get into understand is the class limit and class width. So, the first one is the term called lower class limit so is the first term. So, what is it means the lower class limit is defined as the the smallest value that can belong to a given data range or data range is also known as or interval or group. So, the smallest value that can belong to a given data range is called as or the group or the interval is known as the lower class limit.

So, when you say example 50 to 59, this 50 is the lower class limit. So, if you have write this is means that 50 less than or equal to x less than or equal to 59 this is the mathematical way. So, you saying that all values including 50 and 59 between this falls into this particular group and of which 50 is the lower class limit of the group; then the other one is the upper class limit. So, what is the upper class limit? Same thing just different the largest value, the largest value that can belong to a given interval or interval is also known as data range or group. So, these are the two things. So, now let us take an example; when you say 90 to 99 which implies 90 less than or equal to x less than or equal to 99 and this 99 is the upper class limit alright. So, you now understand the two terms.

So, in any class that you require you require a upper class and lower class limit. Then if this is not then the third term that we need to know is what we call as class width. So, what is class width? The difference between the upper class limit the upper class limit and the lower class limit is defined as the class width; is defined as the class width. So, in a class of 80 less than or equal to x less than or equal to 89 what is class width. I think it is a simple question that you guys can answer. So, class width this question please remember and I expect that you guys answer this as part of your homework.

So, I told you what is an upper class limit, what is a lower class limit and I given you the definition that it is the difference between the upper class limit and the lower class limit if there is a difference then what is the class width of class where 80 less than or equal to x less than or equal to 89. Now, one other thing that you might noticed from this case is that these data rangers, this arbitrarily chosen. So, I choose it by myself, I did not use any rules or other things. So, the question is, are this one question that we need to ask here; obviously, is are the classes chosen arbitrarily or is there any rule. This is the big question when it comes to the guidelines for making classes.

(Refer Slide Time: 21:27)



And, I will give you 6 major set of rules. There are some guidelines, there are some guidelines and what are those guidelines? And these are usually followed this are not rules, they are guidelines but this will help you. The first one, there should be between 5 and 20 classes. So, the number of classes; number of classes should be between 5 and 20. So, if you notice how many classes I had? 1 2 3 4 5 so, I had 5 classes; so, I exactly hit the lower number of desirable number of classes in this case. Number 2, the class width should be an odd number. So, this is a tool to the answer of the question. What is the class number? Should be an odd number.

This will guarantee; this will guarantee why is it an odd number, why is recommended to be an odd number guarantee that the class mid points the class midpoints are integers instead of decimals. So, for example if you have a class of let us say if the class is of 30 to 30 35 30 31 32 33 34 and 35. So, what is the class width in this 1 2 3 4 5 6. So, if we have the class width of these as 6, the midpoint of this class is here 32.5. So, that it is equally on both sides, but if it is 36 is there, if the class number is class width is 7 then the midpoint will shifted towards here. So, then now you have 1 2 3 equal width on both sides of the class. So, this is one way of looking into this alright.

Now, 3 the classes must be mutually exclusive, but by the way this guideline is usually violated, that is which is not very easy to follow this guideline. The classes must be mutually exclusive; classes must be mutually exclusive. What does it means? Implies

that no data value can fall into two different classes can fall into two different classes. So, let me put it this way, if you write if you say I have two classes; one class is 50 less than or equal to x less than or equal to 59. Another is 60 less than or equal to x less than or equal to 69, then it is mutually exclusive.

Because, the value of 59 will fall in this class, value of 60 will fall in this class. If you ask me what is the value of 59.5 then; obviously, we were already dealing with integers in the previous lecture. So, the so, this is a mutually exclusive mutually exclusive for integer values alright; otherwise you have to make the classes accordingly different. Then the fourth rule is the classes must be all inclusive or exhaustive. What does this means? Implies that all values all data values all data values must be included.

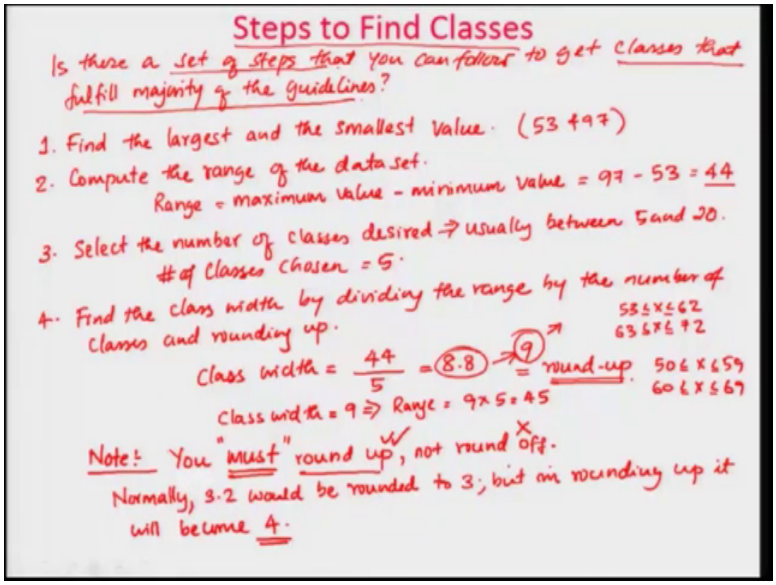
So, one of the things that you have to ensure that the classes should be inclusive or exhaustive so, if we go back to the previous data, if in this 99 that we used, 99 we will ensure that the largest value that we got which is 97 is included in this. So, and same way 53 because, the lower limit is 50 will ensure that the 53 is included in this. So, another way to say it is the smallest and largest value of the data set should be within a class. So, that is the important aspect in this regard. So, 5 the smallest value if I make the classes, first class is 50 less than or equal to x less than or equal to 59 will include 53. If I made this was 55, then this is not; this will not be included. So, that is what we call about it as a mutually exclusive class.

Then the other thing the fifth rule that we talk about is the classes must be continuous, must be continuous which implies, no gaps in the frequency distribution. So, what we are saying here is that since these are all integer values, if you look into it 50 to 59 is one class both sides included then 60 to 69 is another class. So, it is continuous after 59 the next value you can get is 60, but this in a different class. So, ensure that these are exclusive, but also continuous for that particular. So, since it is all integer values this classification is fine or classes is fine.

And the 6th one which is also another thing that sometimes people violate, the classes must be of equal width the classes must be equal in width. So, this implies that this gives you hence, gives lot of uniformity. So, what we are talking about it is, in this case if you look into it this is 50 to 59 having the equal width. 60 to 69 having the same width, 70 to 79 having the same width, 80 to 89 having the same width, 90 to 99 having the same

width. So, which what is that width that you have to calculate and then you have to answer that question as part of your homework.

(Refer Slide Time: 30:47)



So, since I given you guidelines then are there some steps available to find classes can we find a way or is there is a there a set of steps that you can follow to get classes that fulfill majority of the guidelines. So, what we are trying to do here is, we are trying to find the set of steps that you can follow. So, that you get classes that fulfill majority of the guidelines that we listed previously. So, that is what we are trying to do. So, some of the steps are the first one is, the first step is find the largest and the smallest value, find the largest and the smallest value.

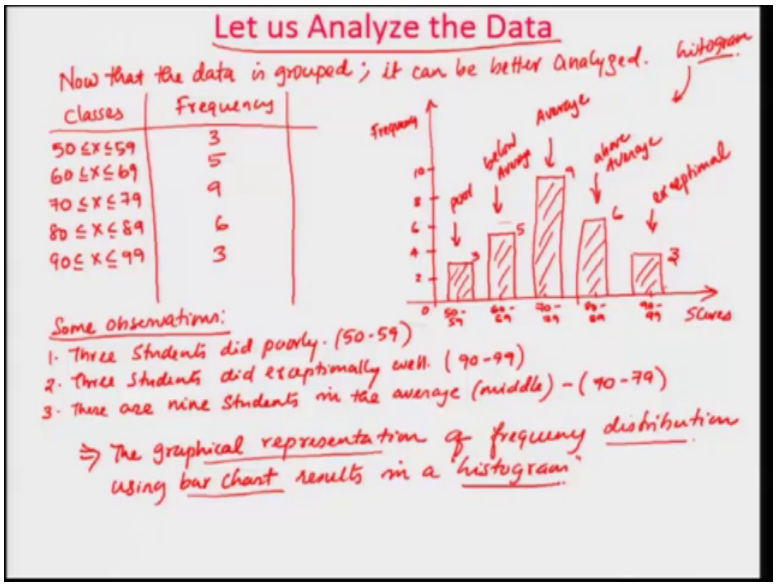
We already get this with the previous data set. So, if you remember that we got the largest and the smallest data values are 53 and 97. So, we calculated as 53 and 97 right. Then second one is compute the range of the data set, range is defined as maximum value minus minimum value. So, in this case it will be 97 minus 53. How much it will be? So, that will be 4 and 44 yes so, 44. So, the 44 is the range of the data. So, from 53 to 97 you find the difference between the largest and the smallest alright. So, once you calculate that then 3 select the number of classes desired. This is usually between desired; usually between 5 and 20, you have to pick a number. So, I picked so, number of classes chosen was equal to 5.

Then fourth one, find the class width by dividing the range the range by the number of classes and rounding up. So, one way to do this is this you can say it as so, class width is equal to the range 44 by number classes is 5. So, that is 8 point what 8.8 that is a class width. So, if you if I round up I can do multiple things, if I round up to 9 then this is rounding up. Then I get the class width will be if I this up, then the a range will be class width will be, if the class width is equal to 9 implies the range equal to 9 times 5 will be 45 right.

So, then this should be fine, I can do this and the class width will I can pick a class in a very unique fashion. What I can say is that I can start with the minimum value of 53. So, the first class will be something somewhat little weird I can start with 53 less than or equal to x less than or equal to add 9 with this so, that will be 62 something like this, then 63 less than or equal to x less than or equal to like that 72. So, I can do classes like that, if I make the class width as 9. Or another thing I can do it is I know this is 9, I can maybe round it up to in a different number I can do it as. Another way of doing it is I can do it as 50 less than or equal to x less than or equal to 59, I can do that.

Then 60 less than or equal to x less than or equal to 69 like that, I can do that kind of a class width as well. Once I get this then I can decide how will I do this. So, if I have 5 classes of 9 width each, then I would be able to get to what I what I want to do. But the trick here is you should remember, note the important part here is that you must you must roundup, not round off round up always you must round up not round off. So, for example is normally 3.2 would be rounded to would be rounded to 3, but in rounding up it will become 4. So, the idea here is that always here round up, you whenever you get the class width you should roundup. So, that you can get to the good thing and remember we rounded up to the next nice looking odd number so, that our problem as actually get solved.

(Refer Slide Time: 38:15)



Now, once this is done now, let us go back to a data and let us analyze the data. Now, that the data is grouped now, that the data is grouped what do I have it can be better analyzed. So, what I can say is that so, I have made the so, we will make the graph once again classes data, not the graph we will put the table classes then we say frequency.

It was 50 less than or equal to x less than or equal to 59, 60 less than or equal to x less than or equal to 69, 70 less than or equal to x less than or equal to 79. 80 less than or equal x plus less than or equal 89, 90 less than or equal to x less than or equal to 99 this were classes and the frequency is where pretty much they were 3, they were 5, 9, 6 and 3 alright. So, if you think about it think about the frequency that we drawn, let us try to see I can say from here saying that some observations from the data. Number 1, three students did poorly who are the three students you want to say poorly? I can say 50 to 59 right. 2, three students did exceptionally well, who are those people? 90 to 99 right; there is also there are nine students in the average or middle which is what is the middle 70 to 79.

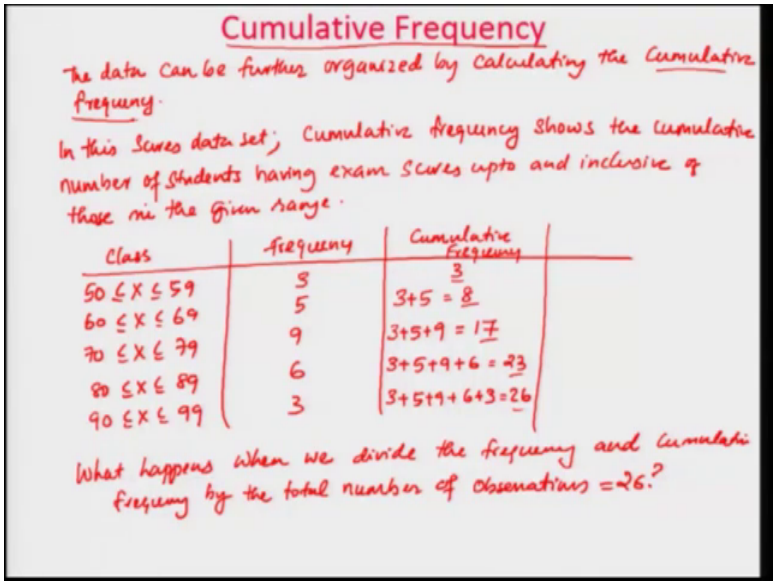
So, this analysis that we just talked about, this is good way of analyzing this is a good way of analyzing the behavior or trends in the data. So, let me let us see whether we can draw this data in graph. This ones are not very easy to draw straight lines, but let us called this as the scores or and I have 1 2 3 4 5. And, this is 50 59, 60 69, 70 79, 80 89, 90 99 and then I put on the y axis, let me put frequency put a 0 it is a 2 4 6 8 10. So, the

50 to 59 we have 3 peoples so, somewhere here we draw a bar. Then 60 to 69 we have 5 so, it is somewhere here. Then 70 to 79 we have 9 so, this will be somewhere here. Then 80 to 89 we have 6 so, somewhat here; my graph drawing skills are not very good, then we have 3 right.

So, if you look into this it actually shows that there are no, if I say this is the average, exceptional, if I call this as the poor. Then this I can call it as the above average, this I can call it as the below average. So, you can see how nicely we can analyze. So, there are it is a class of there are we can say 3 exceptionally good students, then there is 6 above average students, 7 not 7 sorry 9 average students. Then about 5 below average students and 3 poor students. So, this tells you how to analyze or how to even draw.

So, the graphical representation of frequency distribution, using bar chart bar chart results in a histogram so, this diagram that we just mentioned about. So, this is our histogram the using bar chart the graphical representation of the graphical tool of frequency distribution bar chart is called as the histogram. These are the bars, some people shaded like this do some stuff like this, you can do that it is up to you alright.

(Refer Slide Time: 44:59)



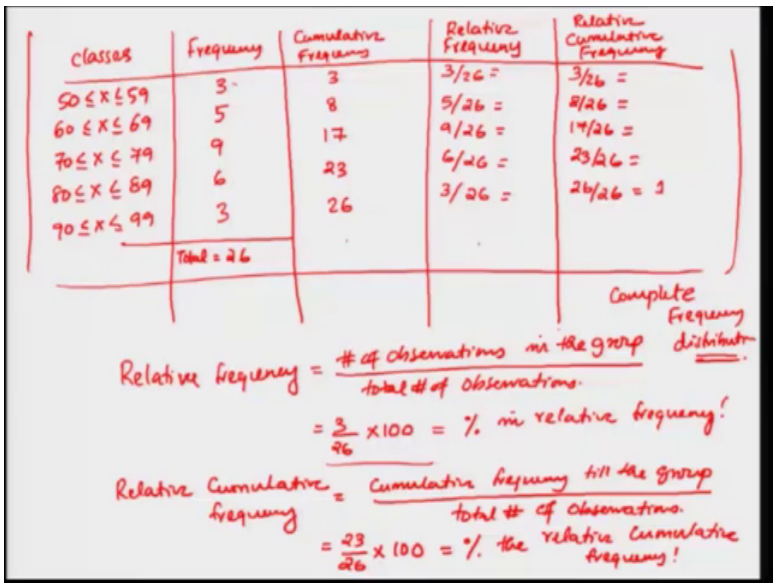
Now, let us talk about cumulative frequency right. So, what is cumulative frequency? Let us, the data can be further organized by; the data can be further organized by calculating the cumulative frequency. So, if we calculate the cumulative frequency then you can organize the data better. So, what is cumulative frequency? In this data set in this scores

data set cumulative frequency shows the cumulative, it shows the cumulative number of students; cumulative number of students with scores up to student number of students not with having exam scores up to and inclusive of those in the given range.

So, what I am trying to do this is do here is so, if I have a class 50 less than or equal to x less than or equal to 59, 60 less than or equal to x less than or equal to 69, 70 less than or equal to x less than or equal to 79, 80 less than or equal to x less than or equal to 89, 90 less than or equal to x less than or equal to 99. And, I have the frequency it has 3 and then 5 and then 9 and 6 and 3. Then the cumulative frequency the cumulative frequency is for this one it is 3, next one it is 5 plus 3 is 8, next one is 8 plus 9 is 17. Or another way to say it is instead of doing that so, this is 3 and then the previous one is 3 and then the frequency that is belong to the current one, which is 8.

This one will be 3 and then 5 and then the current one, that is 8 plus 9 17, this will be 3 plus 5 plus 9 plus 6 that is 23. Then it is 3 plus 5 plus 9 plus 6 plus 3 26. So, this is what you call as a cumulative frequency 3, 8, 17, 23, 26, where you have all the values up to that mark of the class; the upper limit of that particular class is what we are talking about. So, now if this is the case then, what happens when we divide the frequency and cumulative frequency by the total number of observations which is equal to 26, what happens? So, let us think about that in a second.

(Refer Slide Time: 49:42)



So, I will redo this table once again we will have classes which is 50 less than or equal to x less than or equal to 59, 60 less than or equal to x less than or equal to 69, 70 less than or equal to x less than or equal to 79, 80 less than or equal to x less than or equal to 89, 90 less than or equal to x less than or equal to 99. And, we had the frequency and that was 3 5 9 6 and 3. And, then we had cumulative frequency that was 3, then 8, then 17, then 23 and 26 and here the total is also 26.

So, then there is a term that I told called relative frequency. Relative frequency is calculated by dividing the number of it is relative frequency for you to remember again, will be number of observations in the group divided by total number of observations. So, in that case we will have if we look at the relative frequency of this will be 3 divided by 26, this will be 8 divided by 26, this will sorry my bad again 3 divided by 26. These numbers we are talking about and then it is 5 divided by 26, then it will be 9 divided by 26, then it will be 6 divided by 26, then it will be 3 divided by 26. So, these values so, for example when you say 3 divided by 26 multiplied by 100 will give you the percentage in relative frequency alright.

So, the relative frequency can be easily converted to a percentage by multiplying it with 100. Same way what we call as relative cumulative frequency, when we say that, then that will be 3 by 26 the first one, then it will be 8 by 26, then it will be 17 divided by 26, then it will be 23 divided by 26, last will be 26 divided by 26 which will be 1. So, if you have the relative cumulative frequency is the cumulative frequency till the group divided by total number of observations. So, it can be also converted to percentage for example, it will be 23 by 26 multiplied by 100 will give you in percentage the relative cumulative frequency.

So, this whole thing, this whole distribution that we saw this much, you can call this as a complete frequency distribution. So, frequency distribution will have the frequency classes the frequencies, the cumulative frequency, relative frequency and relative cumulative frequencies summarized and appropriate diagrams histograms that you are going to draw about this, will give you the behaviour and you can make nice conclusions about the data. Because, once you group it then you will have start seeing patterns or specific behaviors within the data that is why human beings are good at grouping things. So, if we have different type of vegetables in a bag you will take that and then group them by the type of vegetables, by the size and shape, same exact idea.

So, I hope that all of you guys have understood the concept of relative frequency and what we will do continue in the next class is about getting back into little bit of probability distributions and type of common distribution that we use. We remember we cannot cover all sort of probability and statistics that is required the skills that is required for you in this class, this is just an overview a quick overview of probability and statistics. So, that you have some idea or some of the basic tools that are required to do better or do the analysis of it. Please do refer the probability and statistics books and read better. So, these are some of the tools that I just demonstrated to you guys in the class, continue reading the text book and we will see you in the next class, the next week.

Thank you for your patient listening, good night.