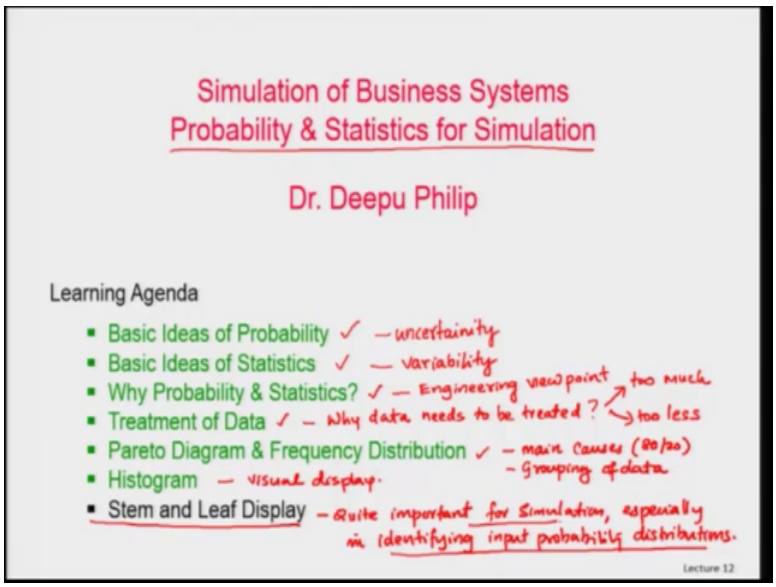
Simulation of Business Systems
Mr. Deepu Philip
Department of Industrial and Management Engineering
Indian Institute of Technology, Kanpur

Lecture - 20
Stem and Leaf Display

Good evening students, welcome to one more lecture probably lecture 12 of the power point sides of presentations of the course Simulation of Business Systems, and we were working on the topics of probability in statistics and we were almost it to the end of the topic. And today we will kind of conclude that major topic today and then we will move on to other aspect validation verification etcetera of simulation systems pretty soon.

(Refer Slide Time: 00:47)



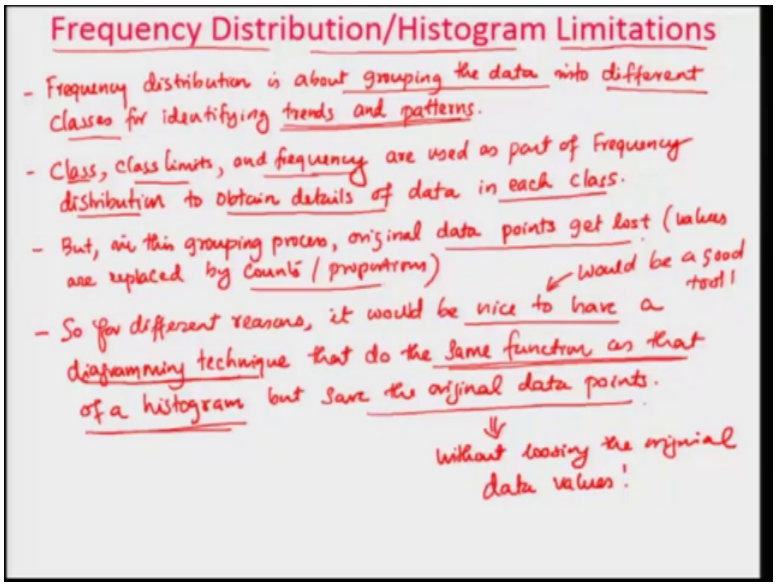
So, today if you look into today's learning agenda, it will be probability and statistics for simulation that was the lecture that we were going through. We have already covered the topics of the basic ideas of probability, why probability and uncertainty we discussed this uncertainty. Then we talked about basic ideas of statistics and we talked about the concept of variability and why statistics is necessary to deal with variability. And we talked about why probability and statistics from an engineering standpoint why it is needed and what do engineers do here in this regard, what is so, important about the engineering.

We talked about the treatment of data, how many ways you can treat the data what are the specific tools and other aspects of it, and why data needs to be treated needs to be treated. We talked about data coming in 2 forms either it is too much of data or too less of data we discussed that. Then we in that regard we also talked about Pareto diagram and frequency distribution. Pareto diagram is identifying the main causes main cause causes that is the 80 20 rule, in the frequency distribution you was about grouping of data and we discussed why it is needed, and we talked about the visual display called histogram which is a visual display of frequency distribution ok.

And today we are at the last topic, which is what we call as the stem and leaf display ok. And it is an important tool usually this is quite important for quite important for simulation people, because it has a uniqueness to it for simulation um. Especially in identifying input probability distributions; input probability distributions.

So, what we are saying here is that, remember if you did arena we actually used constant uniform and those kinds of things. So, how did we come across identifying those probability distributions for the simulation models? So, stem and leaf displays a good way to look at the data and analyze from that.

(Refer Slide Time: 03:26)



So, we already seen the frequency distribution and the histogram, and we saw how it is a good tool. It is a grouping and those kind of things and, but we are now here to look at what are the limitations of the same. So, as we said if you do a quick recap it is

frequency distribution frequency distribution, we support grouping the data is about grouping the data is about grouping the data in to classes in to different classes; different classes for identifying why do we do this? For identifying trends and patterns trends and patterns ok.

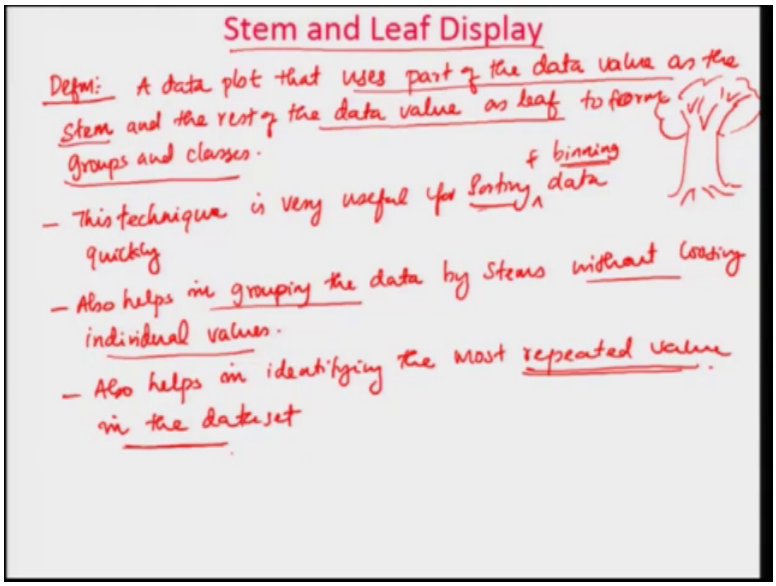
So, you are grouping data you are grouping the data, into different classes multiple classes. So, that you can identify trends and patterns in to this of the data that is one part first part. Second part is that we also saw what is a class ok; class limits and frequency and frequency are used as part of as part of frequency distribution frequency distribution. So, we use classes, class bound limits, frequency boundaries etcetera are used as part of frequency distribution to obtain details of data details of data in each class ok.

So, we know that classes class limits and frequency are used as part of frequency distribution, to obtain details of the data in each class. So, from each category how much of data belongs to that category, we uses these tools to identify that ok. But in this process, but in this grouping process when you are grouping the data in 2 different classes, but in this grouping process original data points original data points get lost or values are replaced are replaced sorry are replaced by counts or proportions.

So, here what happens is, the original data points the original data values get replaced by counter proportions. So, the actual value of the data gets lost. So, for different reasons, it would be nice it would be nice to have a diagramming technique to have a diagramming technique what would this technique do? That function that do the function do the same function ask that of a histogram, but save the original data points.

So, if you have a diagramming technique that can do the same function as that of a histogram ok. Histogram shows you the trends the trend and patterns in the data the grouped data. So, it can do the same thing; show the trends and patterns in the data, but without losing the original data values. So, this means without losing the original data values there is a technique of this. So, if there is a technique if a diagramming technique exists, then that would be really nice it would be a good tool this means would be a good tool ok. So, that is where we are understand, is there a good tool like that.

(Refer Slide Time: 08:36)



So, what we are talking about is stem and leaf display, which is a tool like this which is a tool that do exactly what we wanted to do. So, by definition the definition is that, similarly leaf plot is a data plot is a data plot that uses that uses part of the data value, part of the data value as the stem as the stem and the rest of the data value the rest of the data value as leaf to form to form groups and classes. So, you still doing grouping using classes, but you still uses a part of the data value as the stem and the rest of the data value is used as a leaf that is why it is called stem and leaf ok.

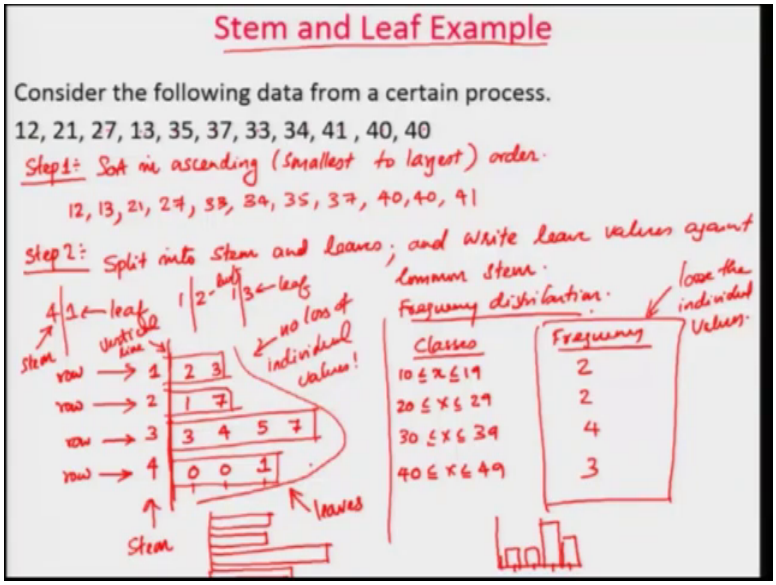
So, if you think about a tree, you can think about a tree something like this which has stems like this and then you have leaves around it ok. So, from a central stem you have multiple leaves going in different directions. So, something similar to that some similar idea like that ok.

So, the major advantage of this is, this technique this technique is very useful for sorting data quickly ok. Sorting and binning data quickly or grouping data quickly sorting and binning can be done quickly also helps in grouping the data, the data by stems without losing individual values ok. Last one is also helps in identifying the most repeated value the data set. This kind of information is usually important when we are doing simulation studies because it is very very important very very necessary.

So, it is a technique that is very useful for sorting and bending, it also helps in grouping the data set without losing the individual values and it can also identify repeated values

quickly ok; so, how because this looks like a very good technique. So, let us now take a look into you know what the technique is all about.

(Refer Slide Time: 11:59)



So, let us take an example so, that I can use this example to demonstrate the stuff ok; so, the stem and leaf procedure. So, a set of data is given to you from a certain process 18 21 like this. So, the first thing when you do is step 1 SoA in ascending order ascending which is smaller smallest to largest SoA in ascending order ok. So, that will be 12 13 then after that it is 21, 27 then it is 33, 34, 35, 37, and 40, 40 and 41 see how 1, 2, 3, 4, 5, 6, 7, 8, 9,10, 11 2, 3, 4, 5, 6, 7, 8, 9, 10, 11; 1, 2, 3, 4, 5, 6,7, 8, 9,10, 11. So, 11 data points all right.

So, now what we are going to do is, if you have to do then step 2 split into stem and leaves ok. In this case what happens is if you have a number let us say 41 then by splitting it in the middle this will become the leaf and this will become the stem. So, if you split all the values in to stem and leaf. So, then you have 12 and 13, those 2 values if you split by them, then 2 and 3 this is a leaf this is a leaf and both of they have common stem 1 split into stem and leaves and write leaves leaf values against common stem ok.

So, in this case you have stems so, 1, 2, 3, 4. So, what you will do is you will create a diagram like this 1, 2, 3 and 4 and in the one the leaf values are 12 and 13. So, it is 2 and 3 you write like this then in 2 you have 1 and 7 then 3 you have 3, 4, 5, 7 3, 3, 4, 5 and 7 and then in 4 it is 0 0 and 1.

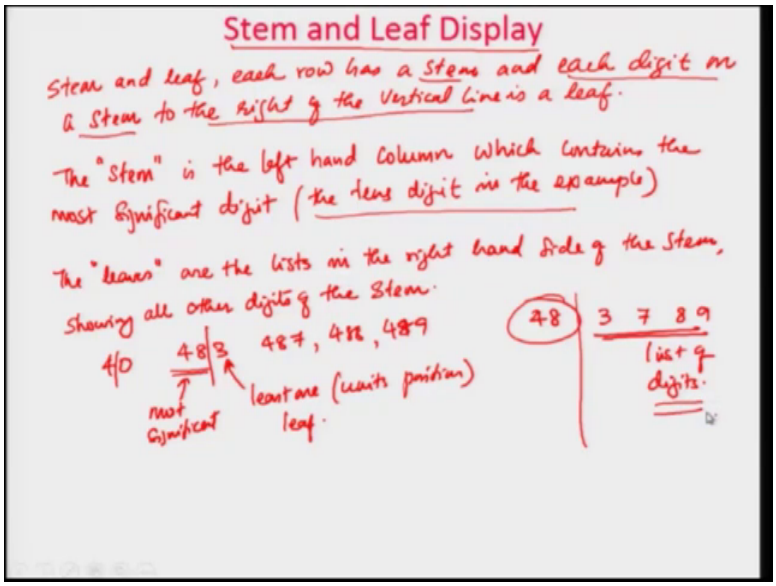
So, if you take a look into this, this is almost like drawing a histogram one way to think about drawing a histogram would be the other way around. So, we think about this as x axis and this as y axis, this is like drawing a histogram of a histogram like this exactly drawing the histogram like this is what you ended up seeing right here if you have to do a ok. If you want to do a frequency distribution I am going to take 4 classes in this regard this 10 to 40. So, I will say classes and frequency.

So, I will say 10 less than or equal to x less than or equal to 19, then I have 20 less than or equal to x less than or equal to 29, 30 less than or equal to x less than or equal to 39; 40 less than or equal to x less than or equal to 49. If I do this in the frequency instead of this I will be writing 2 here also I will be writing 2 this I will be writing 4 now this I will be writing 3.

So, the histogram is exactly pretty much the same if you look into this, but the only difference is that here you do not know once you have this information and you plot a histogram of the same thing which is like this ok. Now 1 2 3 4 you have like 2 and 2 and 4 and 3, and you draw a histogram of this fashion you are not having the individual values. Whereas, here the histograms what we talked about if you use the look at the values, they actually contain the values individual values. So, this way you get to know the pattern of the data, I can say that there is a kind of a lumping towards the end without losing the individual observations or the values of individual observations ok.

So, here you lose the individual values, here no loss of individual values ok. So, that is the main difference that you see in both of the systems ok.

(Refer Slide Time: 17:22)



Now if you look into this there is some more things that additional information that we need to also understand about stem and leaf ok. So, stem and leaf in stem and leaf, each row has a stem each row has a stem and each digit; each digit on a stem to the right of the vertical line.

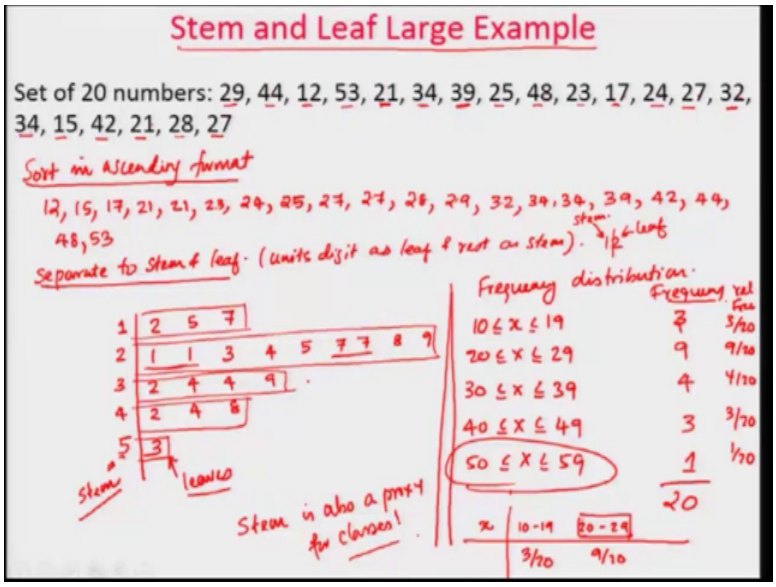
So, what does it actually means is, each row has a stem. So, when we wrote the stem and leaf diagram; if you look back each row this is a row, this is another row, this is another row, this is another row each row has a stem. So, and here is the stem and there are the leaves right. So, each row has a stem and each digit on the stem to the right of the vertical line line line is a leaf ok. So, your leaf is each this is the leaf individually these are the leaves ok. So, each row has a stem and here is the vertical line, and each digit to the right of the vertical line is a leaf thus the first part.

Then the stem or what we call as the stem is the left hand side column left hand column, which contains the most significant digit the tens digit in this case in the example ok. So, this is the, if you look into this, these are the most significant digit in this case ok. So, that is what we are doing here. So, the tens digit is being used in this example, which is the left hand side. The leaves the leaves these are general rules the leaves are are the lists in the right hand in the right hand side of the stem, showing all other digits of the all other digits of the of the stem.

So, for example what happens is if you have 40 it is easy, it you can say 4 and 0. If you have 483, then depending upon how you want to cut? But ideally you would want to cut a stem and leaf like this. So, the most significant digits actually becomes a stem most significant and least one the unit digit units position will become the leaf.

So, all of the time what happens is, whenever you create a stem here if you say you have data 483, 487, 488, 489 89 as the data values then your stem will be 48 and you have a 3 7 8 and 9. So, this is the list of digits for this particular stem ok. So, that is what is meant by in this particular case. So, for any stem the values opposite in the columns will always be the list of the digits.

(Refer Slide Time: 22:07)



So, now let us take a look into another example say larger example. So, that you guys can you guys understand the importance of this once again. So, the large example contains set of 20 numbers. So, let us look at what are the numbers 29 44 like this. So, the 20 numbers 1 2 3 4 5 5 7 8 9 10 11 12 13 14 15 16 17 18 19 20 and the smallest value starts from in this case I believe is from 12 I believe, is 12 right I think so.

And. So, let us first sort in ascending format we are sorting in ascending format. So, we start with 12 this is done, do we have any 13 of this I do not think. So, we have a 15. So, that is 2, then we have is 17 I believe yes we have 17 ok. So, that is done then after 17 we have there is a 21 do we have a 20 the 2 21. So, then you have 20 one 21 2 20 ones

are over then you have is 22 is there, 22 is not there there is a 23 and then there is a 24 that is it.

So, then there is a 23, 24 then do we have 25 this is 25; obviously, only one 25 25 no 26 there are 27, this one 27, w 27 ok so, 27 and 27, 2 27. So, 25 27 27 then 28, this is 28, only one 28, and there is one 29 only one 29 is there a 30 now we dont have a 30. But there is a no 31 that is a 32 there is a 32 then 33, there is no 33, is there a 34 there is one 34 and there is 2 34 34 is there. So, 34 34 do we have 35 no 35 no 37 now the next one is 39. So, we have 39 then what we have is our 40 we do not have a 40; the next largest one is 42 42 and after 42, then it is 44 4 then we have 48 48 the last left out one is 53 ok.

So, we have sorted average numbers the 20 numbers are sorted in ascending order; 1 2 3 4 5 5 7 8 9 10 11 12 13 14 15 16 17 18 19 20 20 ok. So, we have all the 20 with us, now now separate to stem and leaf. So, we take the last digit unit digit units digit as leaf and rest as stem ok. So, if that is the case example we say 12, this is the leaf this is the stem.

So, the stems we have is 1 we have 2, we have 3, 4 and 5 and in one the stems are 2 5 and 7 and in the 2 the stems are 1 1, then you have 3, then you have 4, then you have 5, then you have 7, 7 8 and 9 then in 3 you have is 2 4 4 and 9 and in this one it is 2 and 4 and 8 and 5 it is 3. So, this is the stem and leaf system ok. So, here is your stem this part here are your leaves this part.

Now, if you do a comparing it with a frequency distribution, if we compare with this I can make a frequency distribution of saying 10 less than or equal to x less than or equal to 19, 20 less than or equal to x less than or equal to 29; 30 less than or equal to x less than or equal to 39; 40 less than or equal to x less than or equal to 49; 50 less than or equal to x less than or equal to 59 and here in frequency distribution the frequency that we put here is the count. So, between 10 is 1 2 3 the count is 3; 22 29 the count is 1 2 3 4 5 5 7 8 9. So, 9 then the rest case the count is 1 2 3 4 then the count is 1 2 3 and 1.

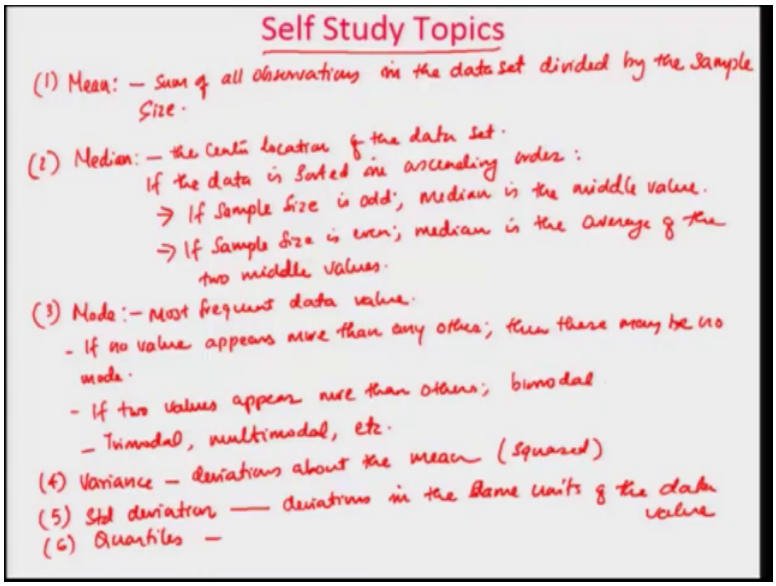
So, if you look into this, this here you can think about as a histogram drawn here without losing the values, but you are still able to see how much of data is grouped into each class ok. So, the leaves you can think about the stem is also a proxy for classes that also you should be able to understand out of this that instead of this class you can think about just replaced by this time ok. So, looks like a good enough examples for you guys to understand and appreciate the importance of stem and leaf.

But now if you want to look into this and why it is important since simulation then you can see that, in a scenario like this you would say that the time would vary between 50 let us say if this was the time taken to process a particular part. You can say that the time will vary between 10 to or 12 to 53, but not uniformly because all values are not equally likely. You would probably say that there are 20 values out of this. So, this will give you a frequency or you can think about as a relative frequency of this. So, that is 9 this is actually 20 data points. So, this is the frequency the relative frequency, in this case is 3 by 29 by 24 by 23 by 20 and 1 by 20.

So, then you can say that or you can say there is out of 20 there are 3 values here the 9 values here, there is you know this 4 values here 3 values here and one value here. So, then you can write a probability distribution saying that the values of x between now 10 to 19, you can say that the probability of it is 3 by 20 value of 20 to 29 is 9 by 20 something like this that is one way to write this, but you can also change here, you can make this further basically saying that the probability of getting 7, 20 ones and 7s are more. So, there are repeated values.

So, that type of adjustments can be done with the help of a distribution here whereas, in this case you will never know that. You assume that all the any value is possible here, but you cannot know here which values are more possible compared to the other values. So, that kind of detailed information allows you to write a better probability distribution for process inputs.

(Refer Slide Time: 30:48)



And this is not a probability in statistics class as I said earlier, but instead is a topic which is necessary for the students to understand and appreciate the use of it in simulation.

So, I am going to ask you guys to do some self-study of the topics. So, the major topics are you should study what is a mean ok. Mostly the what I simply say is as the sum of all observations sum of all observations in the dataset divided by the sample size divided by the sample size.

The next one is what we talk about it does the median. This also you need to study the center location of the data set ok. If the data is sorted in ascending order ascending order ok. If that is the case then number 1; if sample size is odd is odd the median is the middle value median is the middle value ok. If sample size is even even then median is the average of the 2 middle values average of the 2 middle values you can read the details and examples from the textbook I am just giving you important information that is all.

Third you need to study is mode. Mode in a simple sense is the most frequent data value ok. If no value appears more than appears more than any other, then there may be no mode we know there may be no mode ok. 2 if two values appear more than others by model then there is can be tri model, multi model etcetera you need to restart learn and study that ok.

Then you have what is called as variance that is the deviations about the mean ok. It is about the deviations about the mean you to study that, 5 will be standard deviation ok. The, it is a deviation in the same unit as that of the data value in the same units this is actually here you can think about it as squared units of the data. Then you also need to study something called as quartiles ok. These things are all available in a simple probability and statistics textbook and I would expect you guys to read this and understand and go because this by spending time on this we will not be able to spend time on our important topics.

with this what we will do is we will say that we have learned enough to be useful statistics and probability, information for us to be doing very simple simulation models that we are we are building in the class. So, thank you for your patient hearing and please continue to read the books and study things by yourself, do not just limit yourself to the self-studies topics, just study more in probability and statistics, it will be useful to you in simulation.

Thank you very much.