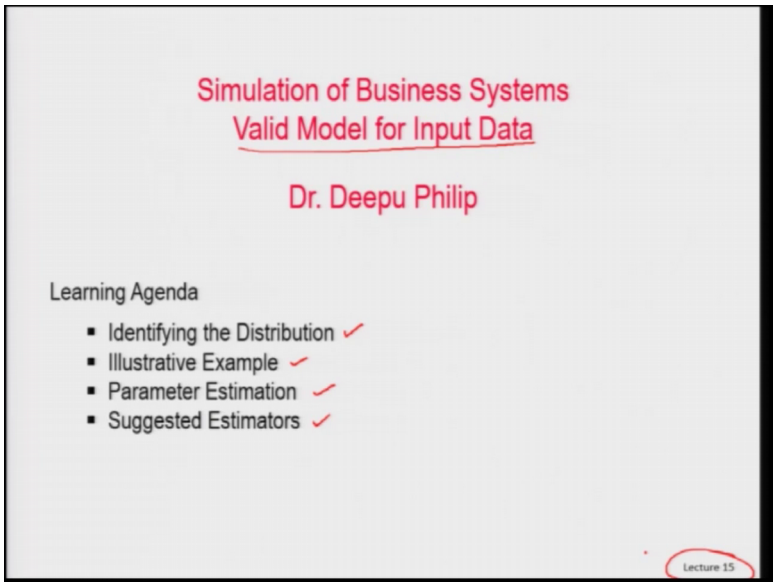
Simulation of Business Systems
Prof. Deepu Philip
Department of Industrial & Management Engineering
Indian Institute of Technology, Kanpur

Lecture – 26
Valid Model for Input Data

Ok students good evening, once again welcome to the conclusive lecture of Simulation of Business Systems course and today we are going to look into how to create the in how to deal with the input data for the models and how to use different parameters for the various input models and also about a quick recap of what we did in the course and stuff and other aspects to this. So, that we have a reasonably good idea of how the course is and what is the purpose of taking this course and how can you take this course, the learning's that you heard in this course forward.

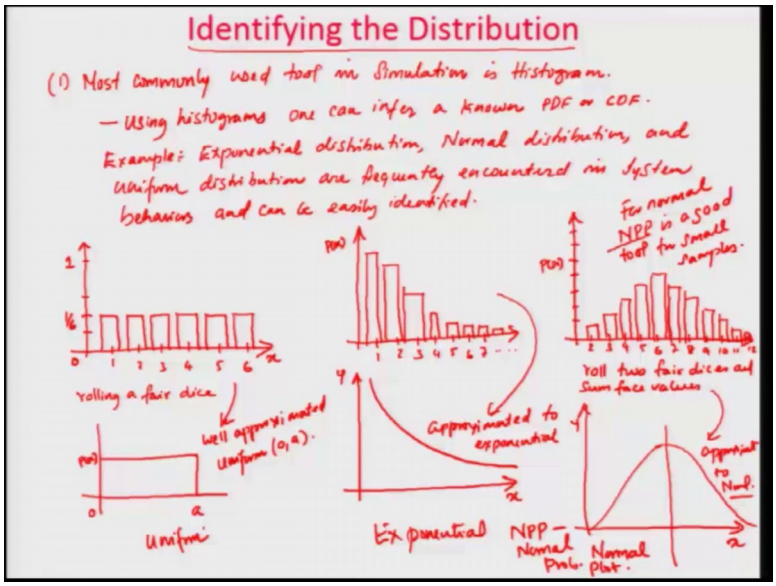
(Refer Slide Time: 00:58)



So, without wasting much time, we will actually look into today's presentation. And today's data is about the valid model for input data, how do we create a valid models for it. And today's learning agenda is the 4 parts. How do you identify a distribution? Just look at an illustrative example because again, people had issues and confusion and doubts about histograms and probabilities. So, we will take a another quick example.

Then we talk about parameter estimation and suggested estimators and this is kind of the lecture 15 of the course.

(Refer Slide Time: 01:34)



So, the first question is identifying the distribution ok, we already talked about this how do we identify the distribution and the first and foremost. Most commonly used tool in simulation is histogram you can using histograms, one can infer a non PDF or CDF ok.

Example ok. Exponential distribution, normal distribution and uniform distribution are frequently encountered in system behaviours and can be easily identified ok. So for an example, if I say that if you look into this, let us say here is x and you have values 1, 2, 3, 4, 5 and 6 and this is let us talk about rolling a fair dice and you have let us say, here is a 1 by 6 as a probability and all the way to 0 to called as 1

So then, in this case let us not do 1 by 6 here, we do 1 by 6 somewhere here ok. So, for each case you will have a distribution something like this ok, you have all the cases the probability will be exactly the same ok. So, then we can easily infer that for this particular case, you can say that the distribution from here, you can identify something like this, you have a distribution like this 0 to a, a particular value and you can say that and here is the p of x ok. So, this is a well approximated uniform distribution, uniform 0, a distribution ok.

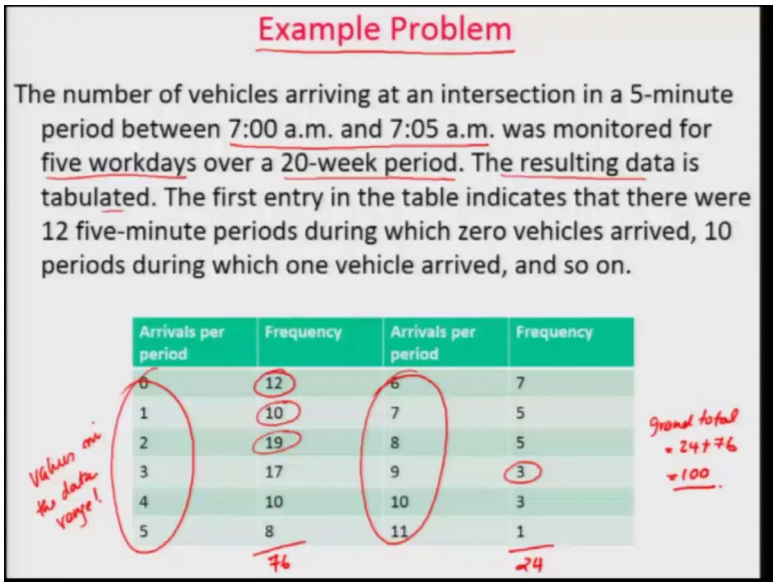
Let us say, think about another scenario where you have a distribution like this, here is your let us say here is a probability and here is your 1, 2 3, 4, 5, 6, 7 all the way like this and your histograms looks something like this. Let us say, you see a scenario like this then, you can very well from here, you can say that here is your x and here is your y and

we can say that distribution like this ok. So, this can be approximated to approximated to exponential ok.

Another example, you can think about it is we discussed this in the class earlier, if you roll 2 fair dice, another process is like roll 2 fair dice and some face values ok. So, then you will have 2, 3 4, 5, 6, 7, 8, 9, 10, 11, 12 like this and you have different probabilities here like this and you will get a probability distribution something like this, 4, 5, 6 a distribution like this, where you can say that an approximated version of this can be easily said like a bell shaped curve ok, where you can say that x and y you can say that this approximated to normal ok.

So these kind of things, you can identify using the histogram, you can identify appropriate distribution. So, these 3 normal uniform, the uniform this is uniform this is exponential and here is normal, these distributions are quite easy to do or develop the, identify through the histogram also for normal ok. For normal NPP is a good tool for small samples. So, NPP stands for normal probability plot ok. So think about this NPP is normal probability plot for small sample size less than 20, 50 or so, you can easily use NPP as a good tool instead of the histogram also ok.

(Refer Slide Time: 08:44)

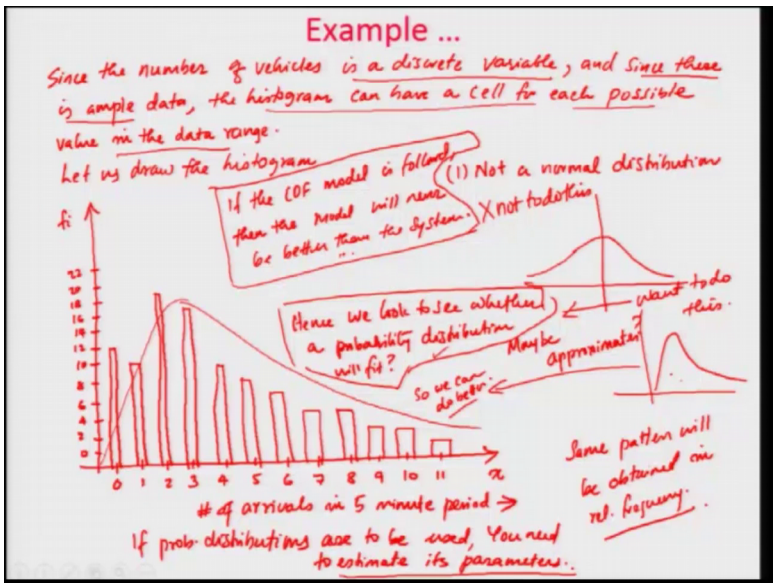


So let me show you an example, an example problem this will probably help you to understand some of the stuff quickly. The number of vehicles arriving in a at an intersection in a 5 minute time period between 7 am and 7 of 5 am was monitored for 5

workdays over a 20 week period. So for 5 workdays so Monday. So, this is Monday, Tuesday, Wednesday, Thursday, Friday, 5 workdays for 20 weeks, you standard an intersection and 7 am to 7 of 5, the 5 minutes in the morning and you observe the number of vehicles arriving at that intersection.

The resulting data is tabulated it is given in a table below. The first centre, in the table indicates that there were 12 5 minute periods during which 0 vehicles arrived. So, this 12 this number 12 means of this total these numbers will total to these frequencies will total 24 and this will total to 76. So, the grand total is 24 plus 76 equal to 100 ok. So, these 100 a vehicle the frequencies that you observed of the 12 frequencies that you saw 0 vehicles arrived, the 10 that you see here means 1 vehicle arrived at an intersection 19 times, you saw 2 vehicles arriving in the intersection, this means 3 times, you saw 9 vehicles arriving in the intersection and so on ok.

(Refer Slide Time: 10:29)



So, if that is the case then ok.

One way to think about it, is since the since the number of vehicles of vehicles is a discrete variable, discrete means they are countable separated out, not continuous and since, there is ample data the histogram can have a cell for each possible value in the data range. So, what we are telling is that, number of vehicles are discrete, where countably discrete and since there is ample data, then the histogram can have a cell for each possible value in the data range.

What are the values in the data range? The values in the data ranges are this;these are the values in the data range ok. So, if you do that. So, let us see let us draw the histogram ok, if we draw the histogram. So then, we can do it this way here is let us take it as fi the frequency, x the number of vehicles, we have 1, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 ok. These are the numbers that, we got to 11 and there are values here 0, 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22 like this ok.

And we saw that for 0, the number of value. So now, we are going to plot this frequency. So, these are your fi frequencies ok. So in this case, 0 it was 12. So, we instead of the histogram, we will brought a small histogram, I was trying to avoid drawing more lines, this will be 12 then, 1 was the previous data was 10 ok. So, this will be 10 somewhere here, then 2 was 19. The next one is 17. So, 19 is somewhere here right, 19 then the one after that it is 17. So, somewhere between 16 and this one then 4 is 10. So, I will get to 10 here ok. 4 is 10, 5 is 8, 6 is 7, 5 is 8, 6 is 7. So, slightly less than that then, 7 is 5, 8 is 5. So, we go to 5, 7 is 5, 8 is also 5 then 9 is 3, 10 is 3, 11 is 1. So, 9 is 3, 10 is 3, 11 is 1 ok.

So, we draw a curve like this, what can you say? Ok can you well this is not a; obviously, a normal distribution ok. So, you can; obviously say, not a normal distribution because, normal distribution ideally would be something like this right. You would have like a symmetric curve, but you can think about a curve something like this ok, maybe this approximates maybe approximate ok. A curve like this maybe people are looking at something like this, a curve like this possible, it is possible ok.

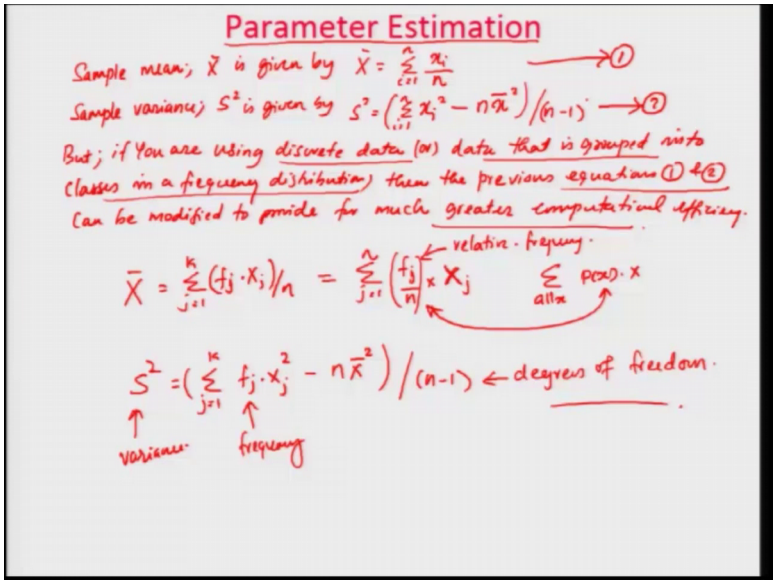
So here, if you look at the number of arrivals in 5 minute period is the x axis and the frequencies or you can do the relative frequency also you get the same graph ok. Same pattern will be obtained in relative frequency ok. One way to do this is you calculate the relative frequency and calculate the CDF and then do the uniform 0 1 or otherwise you try to create this. Why do we try to create this? Because if you go with if the CDF model is followed then, the system will never not the system the model will never be better than the system, this was the problem we studied this earlier ok.

If we use the CDF format ok, we convert this into frequencies and the CDF then, the model will be never better than the system. So, hence we look to see whether, probability distribution will fit. So, our question is we are trying to see whether a probability

distribution will fit to the data so that, why? So, we can do better we might be able to do better than the actual data. So, that is the reason, why we try to look into patterns and try to generate these, kind of probability distributions ok.

So, if you do not want to do this. Let us say, not to do this and you say want to do this then, what you need? Ok, if probability distributions are to be used you need to estimate it is parameters ok. So, the difference in the previous case is here, you do not need any parameters in this case, you do not need any parameters whereas, in this particular case parameters are required, you need to estimate your parameters ok.

(Refer Slide Time: 18:59)



So, parameter estimation, how do you estimate the parameters? Ok. So, the basic thing is first thing is sample mean X bar is given by X bar is equal to sigma i equal to 1 2 n Xi divided by n so; that means, sum all the X values then individual by observations and then divided by the number of observations, you will get the X bar. Similarly, sample variance S square is given by S square is equal to sigma i equal to 1 2 n X i square minus sorry the bracket is not here my bad. X i square minus n x bar square divided by n minus 1 that is one option, you can do ok.

So, what you are trying to do is here is that, you are basically trying to estimate the. So, here. So, you are you are looking at the variance here, in this case and then go from there. So, the idea here is that you can calculate this, the mean and the variance out of this particular case, but if you are you are using discrete data or data that is grouped into

classes in a frequency distribution then, the previous equations, let us call this as equation 1 and equation 2 ok.

Then the previous equations 1 and 2 can be modified can be modified to provide for much greater computational efficiency ok. So, if you are using the discrete data or you have data that is grouped into a classes, in a frequency distribution then the equation 1 and 2 can be modified for greater computational efficiency.

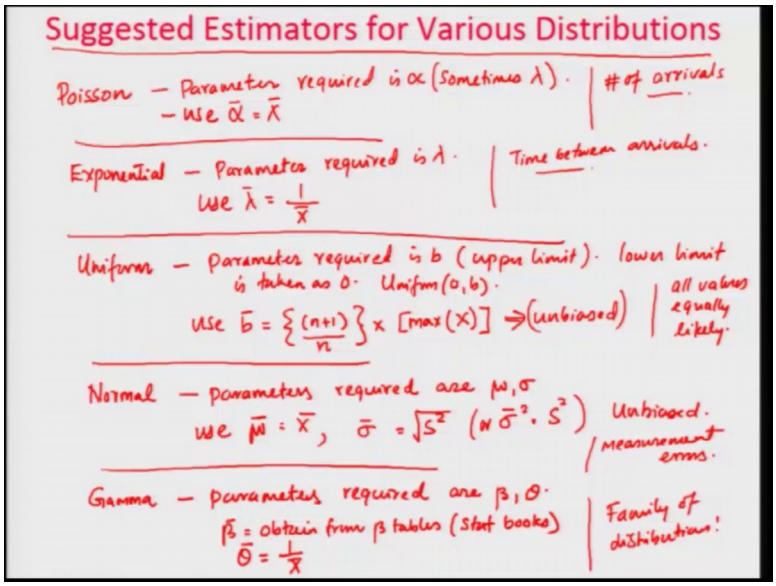
So, what are those? So, X bar will be given by sigma j equal to 1 2 k, f j times xj divided by n. So, what you are doing is individual frequencies multiplied by the value individual x js and you multiply that and divide by n, you will get the X bar this can also be said as sigma j equal to 1 2 n, you can say think about this whole thing as you know f j over n times X j Xj ok.

So, in this case what you doing is this is your relative frequency, this way also you can calculate you might see this is equal to saying sigma all x, P of x times x ok. So, you can see these are the estimates of probability, in this one particular case ok. Similarly S square, you can estimate it as in a particular fashion, you can estimate it as sigma j is equal to 1 2 k fj times x j square minus n x bar square right divided by n minus 1 ok.

So, this is once you calculate X bar then, from there you can calculate the variance can be calculated using this equation, this is your f j is the frequency, this is your frequency x js individual observations, you use n minus 1, this is the degrees of freedom this 8 minus 1 is also exactly the same the degrees of freedom right.

So, I hope you guys understand that. So, this way if you know the mean and the standard deviation then quite a lot of distributions can be dealt with ok.

(Refer Slide Time: 24:27)



So, the suggested estimates estimators for various distributions are in this are given ok. So, I will give you some examples of this, some of the commonly used distribution. So, if you use Poisson distribution ok, this is for the number of arrivals parameter required is alpha or some people call it as lambda also sometimes lambda ok.

The rule is use alpha equal to X bar ok, you are using the mean for the alpha can be estimated as or alpha bar can be used as your X bar ok. So, that is one case exponential distribution ok. Parameter required is lambda ok. So, use a lambda bar is equal to 1 over X bar ok. So, because, Poisson and exponential are related to each other, you can think about lambda as 1 over alpha or 1 over mu, you can think about it either way. So, lambda bar average can be used as 1 over X bar. So, take the reciprocal of the mean.

Then another distribution you can think about it is uniform ok. Uniform distribution parameter required is b upper limit lower limit is taken as 0. So, you are talking about uniform 0 b that is the case then, you can use b bar is equal to n plus 1 divided by n ok. Let us put it this way n plus 1 divided by n times max value of x ok. So, for b bar you use n plus 1, the number of observations n plus 1 divided by n multiplied by the max value of x will give you, the b bar the upper limit and this will be an unbiased estimator ok. This is where unbiased one, that is one thing to use the uniform all right.

Now, for normal parameters required are mu and sigma use mu equal to mu bar equal to X bar and sigma bar equal to S square root both of these can be or sigma bar square is

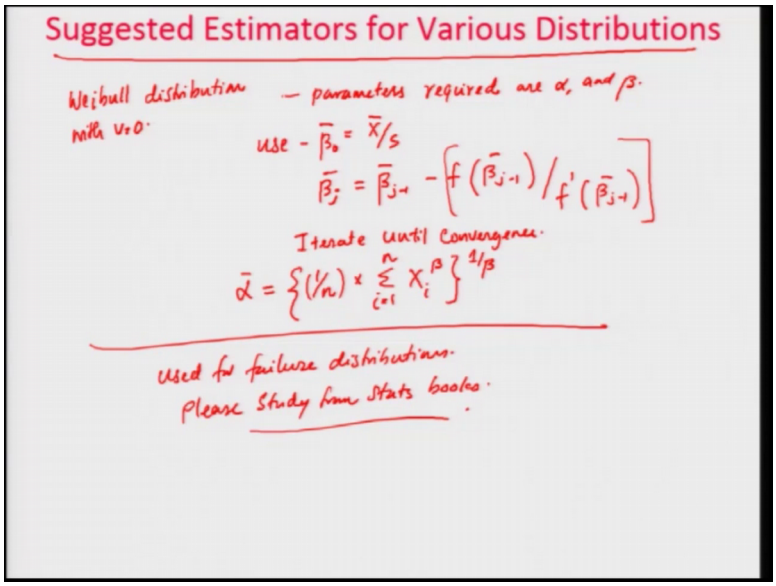
equal to S square one of either one of them you can use ok. So, this will give you the unbiased estimator for this is also unbiased ok.

So, we can see that for Poisson distribution, this is the number of arrivals usually, for to do that you require parameter alpha and use X bar and uses there is an estimator parameter. The exponential distribution is usually time between arrivals and for that you use lambda bar is 1 over X bar then uniform is typically think for all values equally likely and from this case, you use a 0 b the minimum value 0 maximum value is n plus 1 divided by n multiplied by max value of X, maximum value of observation right. Normal this is used to measurement errors, that is what we are using normal distribution for and measurement errors or things, where you know it is a symmetric distribution both negative infinity to positive infinity, we talk about that ok.

So, for that you use mu bar equal to X bar and sigma bar equal to root of S bar S square or sigma bar square equal to S square then, there is one more distribution, which is not that prominent, but it is called as gamma and gamma distribution parameters required are beta and theta beta bar obtained from beta tables, stat books, statistical textbooks will tell you the beta, beta tables from there you can obtain, the value of beta bar ok.

The theta; however, can be calculated as or theta bar can be calculated as 1 over X bar ok. So, theta bar you can is similar to that of lambda bar, but beta values you obtain up to beta bar, you obtain from the textbooks ok. In statistical textbook typically, table eighth table a 8 is typically, what actually tells you the gamma distribution ok. So, this is a family of distributions, even you can think about a long Weibull, they are all kind of part of this kind of a distribution and in a way all right.

(Refer Slide Time: 31:01)



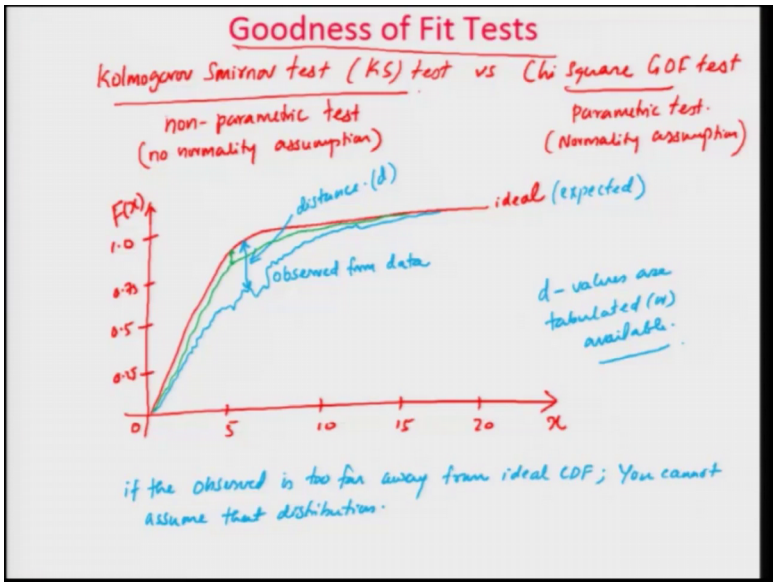
So one other thing that will lot of people ask to us, how do you do the, calculate for Weibull distribution? So, lot of the Weibull distribution is used for failure. So, let me Weibull distribution with v equal to 0 ok. If that is a case parameters required are alpha and beta ok. So, the Weibull distribution with v equal to 0 is parametric, where these alpha and beta. So, first thing you do is use beta 0 bar is equal to X bar divided by S ok.

So first, you calculate beta 0 then beta j bar, beta j bar is calculated by beta j minus 1 bar minus f of beta j minus 1 bar divided by f dash, the differential f dash of beta j minus 1 bar ok. So, what you do is beta j minus 1 minus this ratio, this particular ratio you calculate right.

Then iterate until convergence ok, so you keep on iterating, until you get 2 consecutive values of beta j converging to a particular value beta j minus 1 right then your alpha, alpha bar is calculated as 1 over n multiplied by sigma i equal to 1 2 n X i the power of beta for a 1 over beta ok. So, from beta 0, you can keep on doing this you can do beta 0 calculate that there from there, you calculate beta j and you keep on iterating until, you get the convergence of the value of beta.

Once you get the value of beta then, you calculate alpha bar using that particular value of beta that you obtained ok. So, this is used for failure distributions, time to failure another things, but this is not very common, but you, but please study from stats textbooks. I have already mentioned the name of the textbooks ok, this you how to do it by yourself.

(Refer Slide Time: 34:04)



Then the last point, that people have asked these because, I earlier told about goodness of fit tests and people asked question about Kolmogorov Smirnov test ok. It seems that people have no issues understanding the Chi square goodness of fit test, but lot of people had issues with the called a particular test called Kolmogorov Smirnov ok. It is also known as KS test, identified by some Russians ok.

The biggest difference is versus Chi square goodness of fit test ok. So, this test is a nonparametric test and this is a parametric test., This the normality assumption, it requires a normality assumption, which means the data do follow, some normal distribution here is no normality assumption, you do not require any normality assumption, in this particular case ok.

So what Kolmogorov Smirnov test, ideally do? Because people having this issue with the distance. So, typically what it does is, it has an x like this sorry it is looks like this and then here you have F of x, F of x is the cumulative. So, you have 0, 5, 10, 15, 20 like this the values will go and you have 0, 0.25, 0.5, 0.75, 1.0 like this and your probability cumulative will actually might look like this will come to 1 ok. It will not go above 1 it will just come to 1 right.

The actual this is the ideal and the actual one that you will see might do something like this something like this ok. So, this is the observed from data what the KS test is actually doing is or this is the ideal or expected. The KS test it is doing is, it is calculating this

distance ok. It is this is the distance that, we talk about ok. So, if the curve is far away ok. If the observed is too far away from ideal CDF, you cannot assume that distribution that is the fundamental idea. So, this K whatever the distance or not K d actually is that d that we talk about the d values are tabulated or available ok.

So, if this particular d value for a particular case is less than a particular term, you can say that fine this the CDF this graph is. So, if you have another graph in this regard and you get a another set of data and this actually shows something like this, which is close to this then, you can say that this actually this green line fits more close to the ideal distribution than, the blue line because this distance is a much smaller ok. So, this is the fundamental idea behind the KS, test which is a nonparametric test because, you do not require the normal normality assumption to be satisfied as part of this.

So, considering that today now we come to the end of our presentation and also towards the end of our course, but one of the things that I wanted to tell everybody about this is this is not the end of this vast topic called simulation. Simulation is never taught in 20 hours, it is pretty much taught over 4 to 5 semesters, couple of years.

You need to work on this to get expertise in this and the reason, why lot of the lot of the places the simulation courses are not offered is purely, because of the extensive cost of the softwares and other aspects that are available for it, but I think as far as you guys are concerned, you have been exposed to a critical topic, which is very useful for lot of the students and as well as practitioners. It is a good skill to have because you can actually model and solve complicated problems out of this.

So, I hope that you guys understand this and you guys actually would use it in your life to some extent, maybe in the coming time period in NPTEL, we were probably coming up with an advanced course of this, where you would be having more advanced topics and advanced concepts of simulation being covered in that. So, till that time, I hope that you guys will all have a good time study well and do well in your exam and try to ensure that, you get all get that good certificate. So, that you can show it to people that look, I have acquired this skill, thank you very much and have a happy learning.

Thank you.