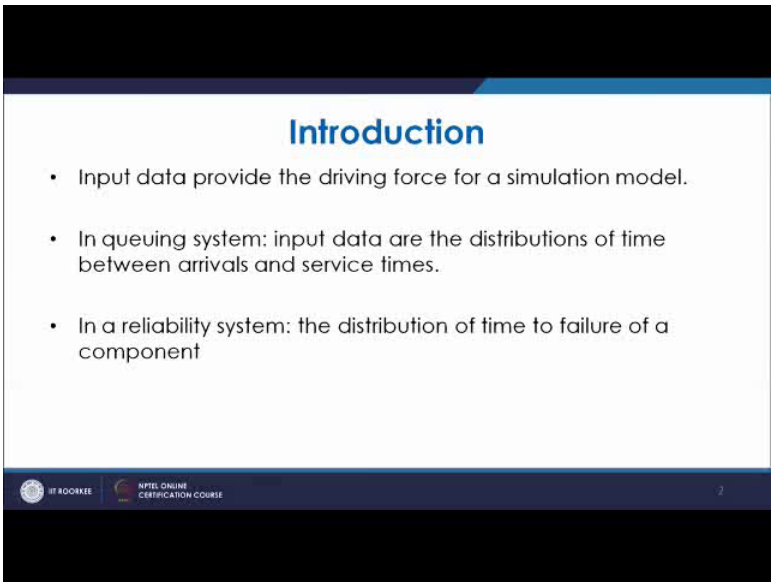
Modeling & Simulation of Discrete Event Systems
Dr. Pradeep K Jha
Department of Mechanical and Industrial Engineering
Indian Institute of Technology, Roorkee

Lecture – 21
Input Modeling: Identifying Distributions with Data

Welcome to the lecture on input modelling. So, in this lecture we will try to have the overview of identifying the distribution with data.

(Refer Slide Time: 00:32)



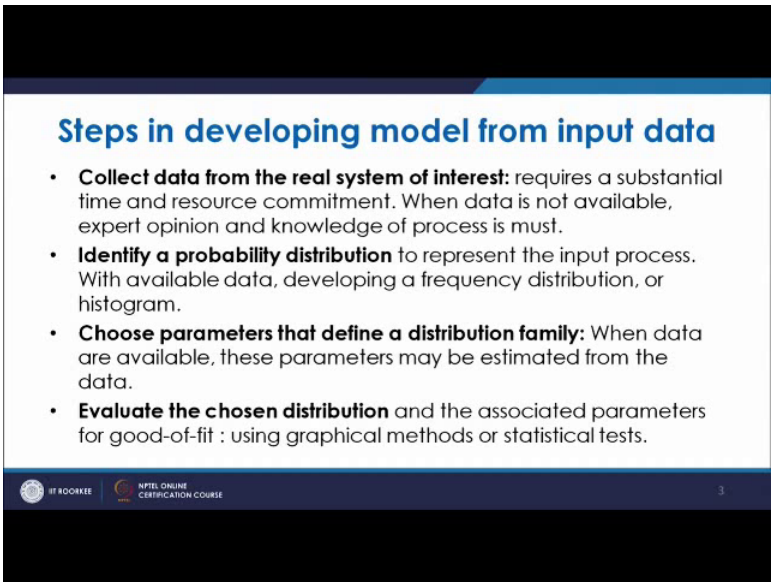
Input modelling is very important part of the modelling process. Now what we do normally in the case of modelling, you have the input data you get it from the environment. You have the sources from where you collect the data. Now this data basically that is why are the driving force for a simulation model. So, basically you need that the more authentic and the more accurate, these data will be the more and more will be trustworthy the model certainly the model assumptions and the methodology is important, but then input data is important.

Just like for example, in the case of queueing system what is the input data? Basically we have the distribution of time between arrivals and the service times these are the input data. So, that you are getting either you can record it live or you are getting it from certain typical sources. So, that you get it similarly in the reliability system the

distribution of time to failure of a component. So, these are the datas which you are generating, you must have certain data set and that you are using.

So, how to model, how to characterise it, how to see that which distribution it fits into, how to use them. So, in this lecture we are going to discuss about them.

(Refer Slide Time: 02:08)



Steps in developing model from input data. Now as we know that the input data is very important. So, you have to have certain steps so that you develop the model from these input data. The first point is that collect the data from the real system of interest. Many a system many a times we are getting the data and we use it for the modelling purpose and whatever result comes we try to interpret it further. Now the thing is that how much accurate the data is, how much pertinent the data is for a particular system.

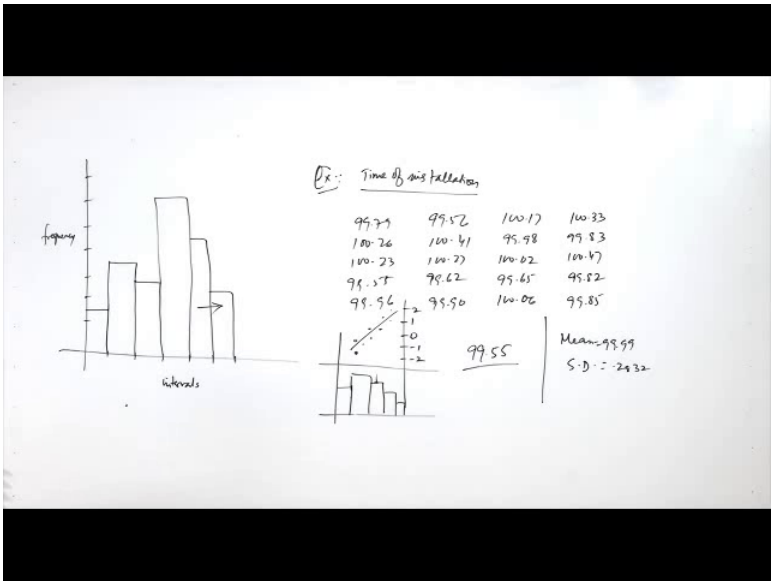
So, if the data is accurate if the data is meaningful it will produce the meaningful results. So, that is why it is said that you must give proper time and you must use proper resource the commitment should be proper. So, that you get the proper data. So, once you get the proper data then only you should use it as an input, many a times the data is not available. So, the data can be generated or data can be suggested by expert persons the subject matter experts in that particular area who has the knowledge of the process.

So, suppose the making of certain unit is to be started. And what will be the time to make certain component of that particular product. Now that may be following certain

distribution that must have being an association with some time most likely time meaning time or a most our maximum time all these data can be given either you get it from the literature or a expert subject matter expert can give you the idea about the you know proper data which have will have the real meaning, and which will make the model more realistic identify a probability distribution. So, once you have the data now with the data you have to identify the probability distribution. So, that will basically we required to represent the input process. So, what you do in that first of all, for that once you have the data you try to see you try to analyse it you do the statistical analysis of that. So, if we have the discrete data you find the frequency of the values which you have got.

Frequency means how many times it has occurred. And then you are basically making a histogram for that. So, basically or you may have the range of data, you range of the interval is there if you have a continuous data. You have the you have to say that some certain data lie between certain intervals. So, whole range has to be defined in certain intervals and then you are making a histogram. So, on the x axis you have the range that is that will be that will be levelled with those range values. And then on the y axis the number of times the data fall into certain range that you have to basically specify. So, for every range for every interval in the range in every interval you will have certain number of occurrences. So, this way you will have a histogram generated for the different intervals of length. Now this intervals and the associated occurrences or frequencies will give you certain idea about the distribution.

(Refer Slide Time: 06:35)



So, what we do is a typical histogram you have the range or interval here. So, we will have interval like this. So, you have intervals and then you have the frequency.

So, in in this case you will have some frequency that particular data, which you have how many number of occurrences it may be 5 10 15 20 or so. And then for every interval you will have suppose you get some kind of frequency distribution, you may get anything like or so. So this way you are getting a histogram. So, histogram is generated. Now if you get a histogram of similar type, then you try to get certain idea from this histogram, as to how it looks what kind of probability distribution function it belongs to. Or if the discrete data is there what is the p m f, how it behaves it may have a shape of something like normal which it looks somewhat it may be something like exponential or it may be something like Weibull gamma or so.

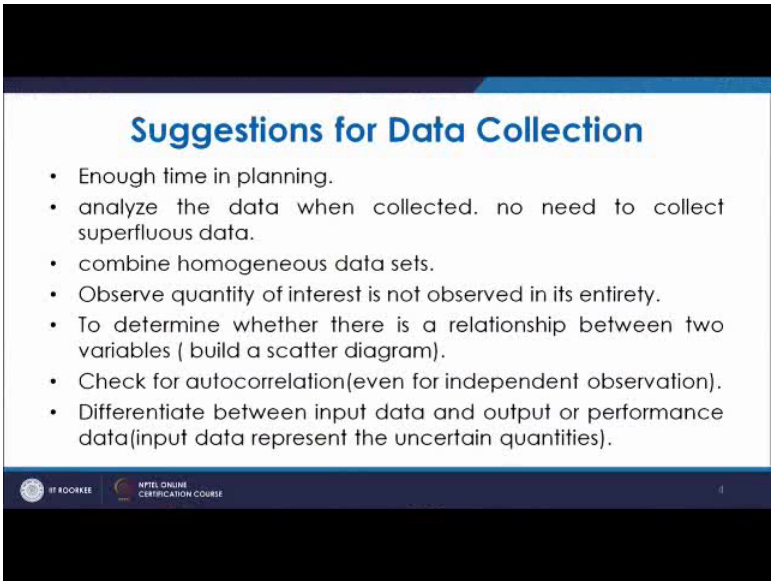
So, this way you are basically identifying the probability distribution. Then the next point is choose parameters that define a distribution family. So, once you see that this is looks like that. So, for the typical that distribution you have the parameter estimation. So, you have the you have the parameter values, like if you have the exponential distribution the rate is to be seen or if you have the normal distribution or any distribution the parameter which is associated like alpha in case of Poisson distribution lambda in case of exponential distribution, you have you know alpha beta and nu in the case of other distribution to have Weibull distribution where you have a 3 parameters along with that shape scale and then location parameter.

So, these are the parameters with the you have to basically define and estimate from the data. Then once you estimate that after that you have to evaluate the chosen distribution and the associated parameters for good or fit the thing, is that once you have you say that it is normal distribution, but then how much it is fit to normal distribution that further it has to pass through certain good or fit test. So, these are graphical test or a statistical test. So, you have to check them you have to test them. And then you can say that yes it passes that test or it does not pass the test it does not confirm to that particular distribution.

If it confirms to certain particular distribution it is fine. If it does not confirm to certain particular distribution you may have to go and see that other kind of distribution function where it basically may fit more properly. And if not in the end you have to take some

empirical distribution function. So, this way you these are the steps which require which help you in developing the model form input data.

(Refer Slide Time: 10:32)



Suggestions for data collection. Now the data which is collected for making the input model, for that there are quite good suggestions, for the you know trustworthy data or for facilitating, you have these are the suggestions. While the data collection enough time in planning. So, you must have to give proper time in planning what you have to collect what is the data which you require. So, it should not be in a hurry it should not be from any source. Whatever you have come I mean basically it is required because ultimately the model output is based on what you are giving the input.

So, if the input is the good one the output will be good, if the input is garbage the output will be garbage. So, garbage in garbage out concept is very much you know applicable here, in this case also that if you have the data which is meaningful you will get meaningful results. Analyse the data when collected no need to collect superfluous data. I mean whenever you are collecting the data you just see, whether this data is required how much it is going to be used how much it is pertaining to the kind of study you are interested in.

So, for that you must analyse the data. There is no need to collect superfluous data. So, the data which is of no use which is outside the limits which does not have any role so that, those data should not be collected combine homogenous data sets, you must know

when to collect the data, which are the data which should be in succession. The data suppose you are collecting you must know that which data should come after which one. Suppose you start the taking the readings on Friday and next reading will be on only on Monday. So, you must know that the Monday data will be coming after Friday, because anyway there is no data in between the 2 days.

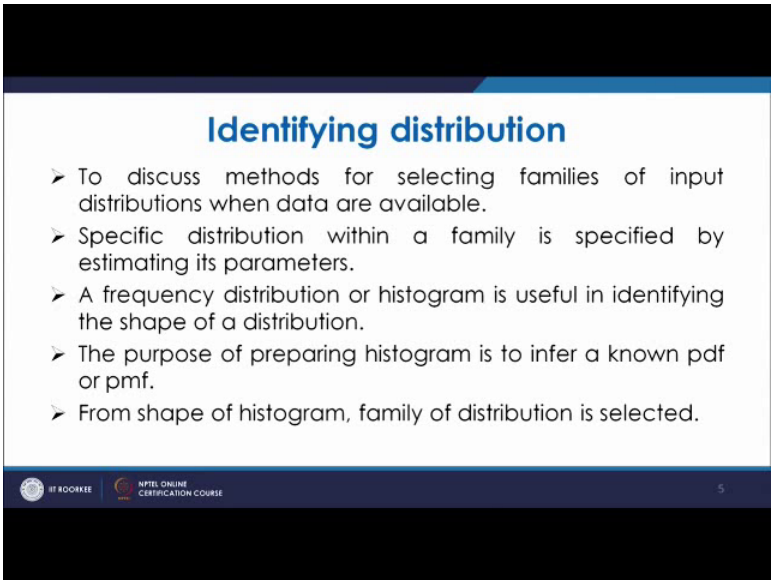
So, the homogeneity homogenous data sets should be combined. So, that you get the data in continuation which has certain meaning which can be correlated with certain pattern or so if required. So, so for that the mean is that you must be able to combine these homogenous data sets. Observe quantity of interest. I mean you have to observe that the quantity of interest is not observed in it is entirety. You must see that the whole approach should be followed here overall you have to see that, the data which you are collecting it should fulfil the desires the goal.

So, in whole hearted way your data I mean quantity of interest it. It should be seen that you are taking these data for the I mean long term analysis or so. To determine whether there is a relationship between 2 variables when you are taking 2 variables using the scatter diagram you can see whether they are related or not.

So, using the scatter diagram you can have the relationship, where they are uniform or they are related in such a way that one increases both increases. The second also increases or one decreases another also deceases like that. So, using the scatter diagram you can see that by looking at the slope of the lines, which you are getting check for auto correlation for independent observations, even you have to see that they are not to auto correlated. Auto correlation can be done by doing the test and you have to see that even the data is looking independent; there may be auto correlation between them. So, you have to also check for that.

Differentiate between input data and output or performance data. Now the thing is that you must know the difference between the input and output data. Now the thing is that input data is uncertain quantity, and output data will depend upon the input data. So, you do not have much control on the input data you are going to get whatever is there. So, you must know that what is the input data and what is the performance output data. So, there must be proper understanding about the datas.

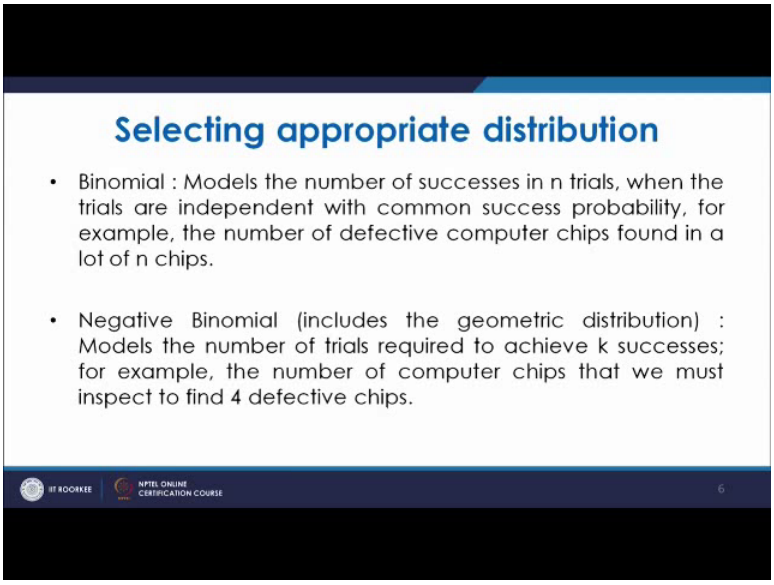
(Refer Slide Time: 16:13)



Now, identifying the different kind of distributions. So, for that we have to discuss methods for selecting families of input distributions, when data is available. Once you have the data available you must know how to put it under a particular family. So, for that once you try to find that this belongs to that family you have to specify or that will be specified by it is parameter estimation it has certain parameters, for every typical distribution you have some parameters as we discussed that you have the frequency distribution of histogram which is basically useful in identifying the shape of the distribution. So, histogram once we draw it tells you the shape of the distribution.

Now, the purpose of preparing histogram is to infer a known p d f, or p m f we have already discussed that we once find the histogram. We see that this is somewhat resembling to certain kind of p d f or p m f. Then from the shape of the histogram family of distribution is selected. So, this way you identify that different kind of distribution. After that we have to see that how to select the appropriate kind of distribution.

(Refer Slide Time: 17:48)

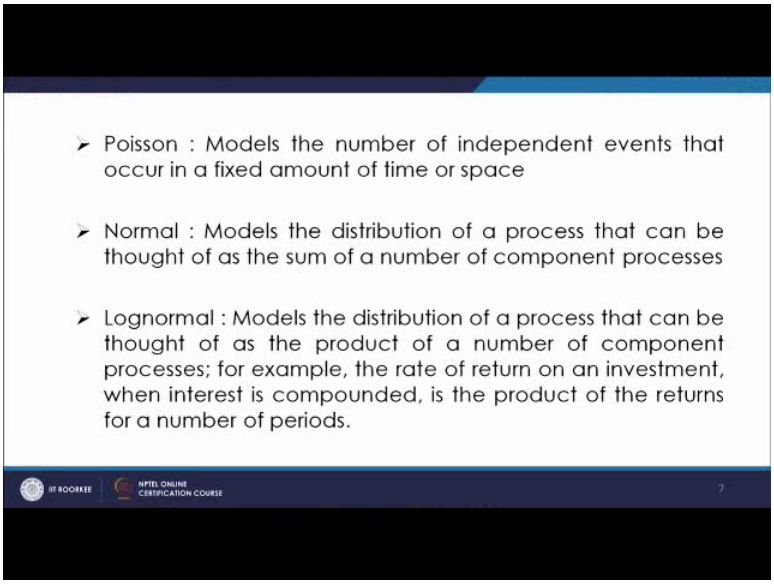


So, we have studied different kind of distribution functions and this will here we will have the knowledge that in what kind of functions, what kind of events what kind of distribution function is normally used; like you have the binomial distribution which we have seen that it is going to help in modelling the number of successes in n trials.

So, whenever we do certain number of trials you will talk about number of successes number of failures even you have only 2 options. So, like you have the number of defective computer chips found in lot of n chips. So, all this kind of studies, whenever we have this we can say that, this kind of if you have the readings it may fit to binomial kind of distribution. Then you have negative binomial distribution which is also something including the geometric distribution, where we are going for iterations or we are going the number of trials till we achieve k number of successes. So, we are going to do the sampling we are going to do, the you know iterations testing till we achieve certain number of you know either successes or failures. So, when we are thus this kind of event in those cases we try to use the negative binomial or the geometric distribution.

Then you have Poisson.

(Refer Slide Time: 19:33)



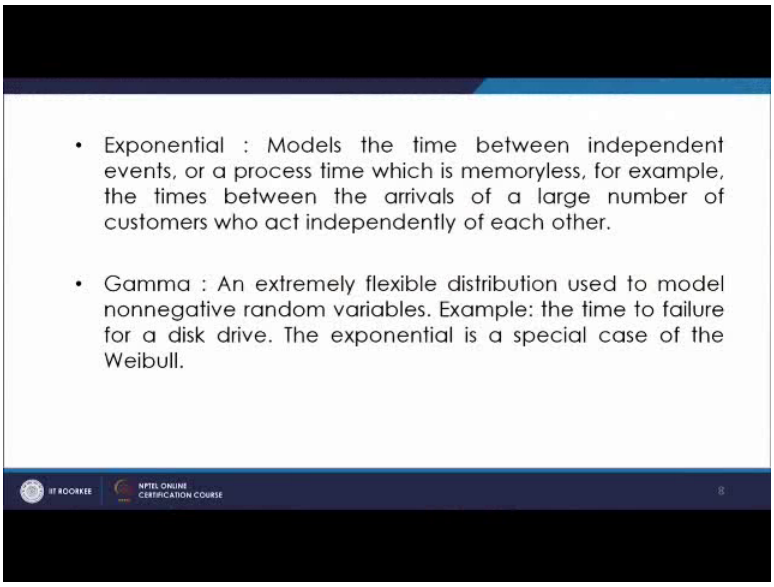
As we know that the Poisson distribution it models the number of independent events that occur in a fixed amount of time or space. So, as we have seen Poisson arrival you have seen that that tells us about how much how many arrivals will be having in unit time. So, number of arrivals in certain interval of time. So, this kind of cases, if we have we can say that it will be somewhat related to Poisson kind of distribution. Normal distribution will be normally it will be modelling the distribution of a process which can be thought of as sum of number of component processes.

Like sometimes we add these processes like final time to prepare certain component is dependent upon the time to prepare n number of components 3 number of components or 4 number of components. So, final time basically it is distribution will be nothing, but the sum of. So, that there I mean if the final time is dependent upon the sum of the individual distribution individual time of some other components. In those cases, in such cases the distribution is normally normal. Because most of the processes will take place in one mean time and then it the other either less or more time that probability will be there, but then that probability will be lesser than for the mean value.

So, this kind of where the you know time is additive, where you see that you have to add this times. So, in those cases this normal kind of distribution normal distribution is used. Log normal is another distribution where the process is thought as the product of a number of component process. When you feel that the process I mean the final product which you get I mean final you have to estimate the final time or so. Or final anything you have to find I mean that is as the product of someone or 2 components or 3

components. In that case the log normal is distribution is seems to be it does seem to be more suitable like the rate of return on investment, when interest is compounded is the product of returns for a number of periods.

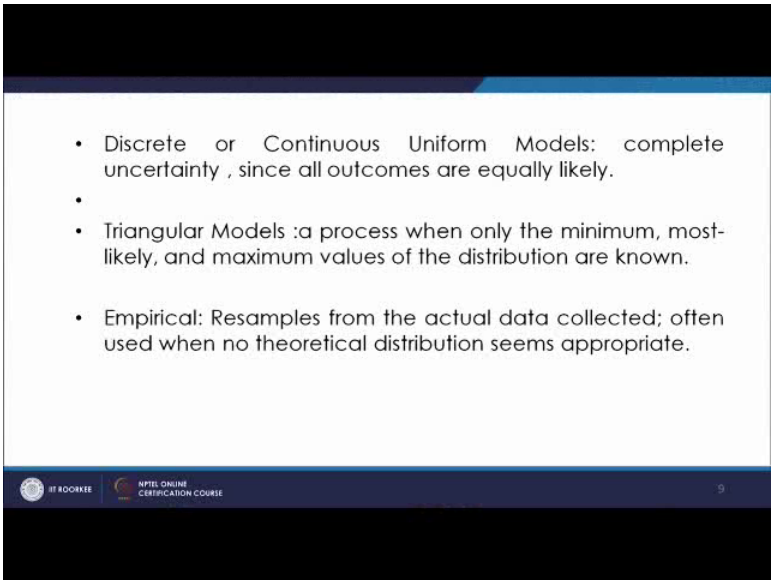
(Refer Slide Time: 22:41)



So, in such cases the log normal distribution holds good exponential it is a very popular distribution which models the time between independent events. So, time between arrivals. So, of a large number of customers react independently. So, you know that it is a it has a memory less property and whenever we have to model such times or we have those times, which are about the time between independent events then in that case we try to fit to the exponential kind of distribution.

Then next is one of the important distribution is gamma distribution, which is very flexible kind of distribution and used to model the non negative random variables. Like time to failure for a disk drive. So, basically in this case you have a gamma function and as we know that it is special case is also the exponential distribution.

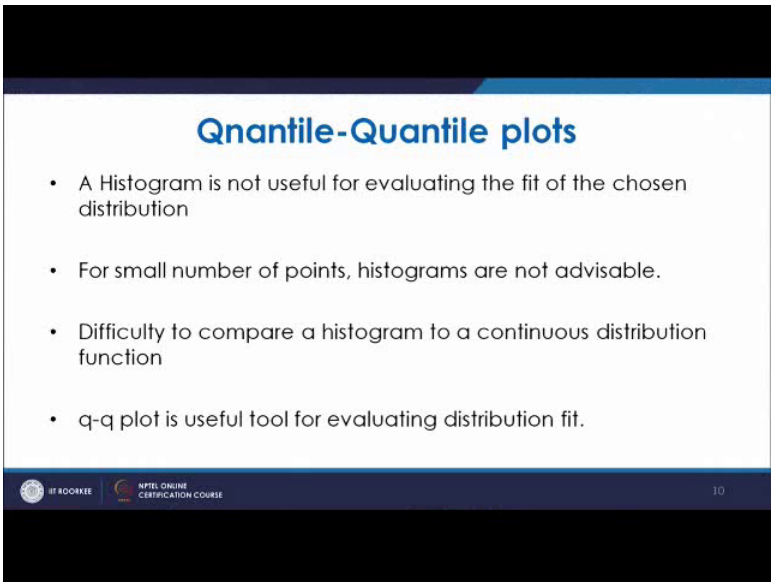
(Refer Slide Time: 23:52)



Then you may have the discrete or continuous uniform models, where the complete uncertainty. Since all outcomes are really equally likely. So, in that case you have uniform continuous or discrete model is used. Triangular models where you have 3 kind of data 3 kind of you know values are there like minimum most likely maximum values.

So, in those cases you can use the triangular models. And we go for empirical kind of models where you have to have the data from the actual corrected one. You have to further sample it and most of the time it is used when we see that if no other distribution is appropriate. In those cases, we go for empirical kind of distributions. So, these ways then you have other distributions also. You have distribution like Weibull is there then you have distribution like beta. So, we know that you have different kinds of distribution functions which can be used for used as different you know areas where they are applicable.

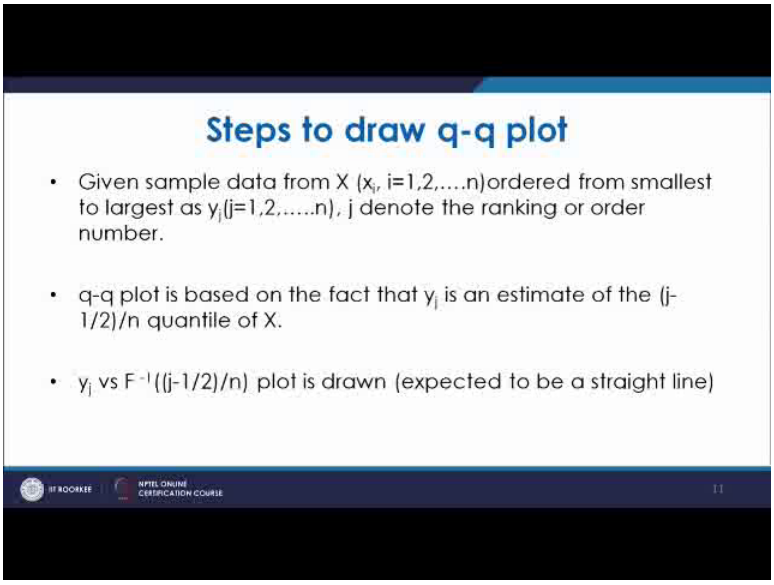
(Refer Slide Time: 25:24)



Now, coming to the qnantile plots, we have to see that how much this distribution is fit, because histogram is useful for evaluating the fit of the chosen distribution. For a small number of, but for a small number of points no when you have large number of points you make it. And then you finally, say that this is how the histogram looks like. And further you can estimate the parameters. And you can find it, but then when the there are small number of points. These histograms are not suitable the histogram will not be it will be invert (Refer Time: 26:15) mode if very small number of points are there. So, there is difficulty when you compare the histogram to a continuous distribution function.

So, there is another tool that is q-q plot which is useful, when you have to evaluate the distribution fit. So, what we do in this? In the cases in the case of q-q plot what we do is once we suppose have made the histogram. Then now we feel that you should see whether how much it is uniformly distributed or how much it is close to the normal distribution, how much it is fitting that. In that case the quantile q that is qnantile quantile plot is drawn.

(Refer Slide Time: 27:02)



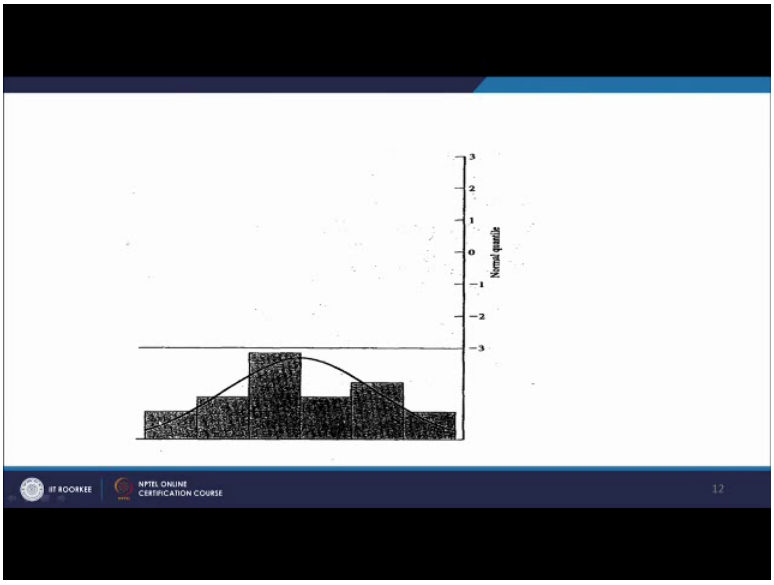
So, how it is drawn, that suppose you have the data sample of data which is there you have suppose hundred number of data is there or fifty number of data is there. So, all the data will basically we first ranked in order. And then you are basically finding the y j, y j means your suppose you have x i as the values x i means x 1 x 2 x 3. And they are not in the any order. So, you are putting in certain order like in a ascending order.

So, earlier you have the data which is not in order now you are putting in the order. And that is y j and then y j versus f inverse j minus 1 by 2 and by n this plot is drawn. So, what is done is suppose you have some data, if you try to see that if you have some data like you have the time of installation for a machine and if you have some data given like the time is 90 9.7 900.2 600.23 99.55 99.96 99.5 600.4 100.27 99.6 2 99.90 100.17 99.98 100.02 99.65 100.0 600.3 3 99.83 100.47 99.82 and 99.85.

Suppose you have these 20 times which you have got from the sample. And you have to see that how much it fits. So, what you can do is first of all you will make the histogram and you will see that what we see is the minimum value in this is 99.55. Similarly, you can get the maximum value and what we see this that may be 100.47 is the maximum value. So, in that case we make the you know histogram for that histogram will come of certain set.

So, if we try to see the histogram.

(Refer Slide Time: 30:13)



We get the histogram like this, where this is 99.4219 1.6 99.18 190 900.2 100.4 and 100.6 something like that. So, this way we can get the histogram which tells us that it is somewhat close to the normal distribution. Now once we get that, once we get the normal distribution in that case it how much it is fit to that normal distribution that has to be drawn I mean seen by using the q-q plot. So, q-q plot in that what we see is we refer to the table. So, in that what we see is f inverse j minus 1 by 2 divided by n. So, what we see is when j is one. So, y j. So, once we get some kind of suppose whatever we get now in that, now I am in this you have q-q plot. So, what we see is normally if you look at this f inverse j minus 1 by 2 by n. It has ideally it has to be straight line ideally; it will be a straight line because j keeping 1 1 minus 1 by 2 that is 0.5 by 20 is 1 by 40. Next it will be 2 minus 1 by 2 by 20. So, it will be 3 by 40. So, 1 by 43 by 45 by 40 or so, this will come.

Now, this value when you will get now the thing is once you have these values suppose 99.55. So, for that what we get is we get the z estimate as 99.55. So, for this sample estimate if you calculate you get mean as 99.99 and standard deviation is 0.2832 that is calculated. So, x minus mu by sigma that is suppose 99.55 minus 99.99 that is 0.44 that is minus 0.44 by 0.2832. So, that is 1.6. So, that will be the minus 1.6 value on the quantile. So, it will. So, here you are 0 minus 1 minus 2 and. So, similarly 1 and 2, this will be quantile value this, quantile value will come suppose here. Here it may come like this. So, this way all these 20 q-q values will be done and they will be seen if they are forming something like a straight line we can say that they are fit in a good manner and if not in that case we say that it is not fitting. So, basically they should be error and

below certain straight line or so. Ideally it has to be a straight line you will not get the ideal case, but you may get to see that they are nearly in the form of a straight line. So, this way q-q plot is plotted to see the that they are good fit.

Thank you very much.