# Simulation of Business Systems

**Prof. Deepu Philip**
**Department of Industrial & Management Engineering**
**Indian Institute of Technology, Kanpur**

## Lecture - 24
## Probability Distributions – I

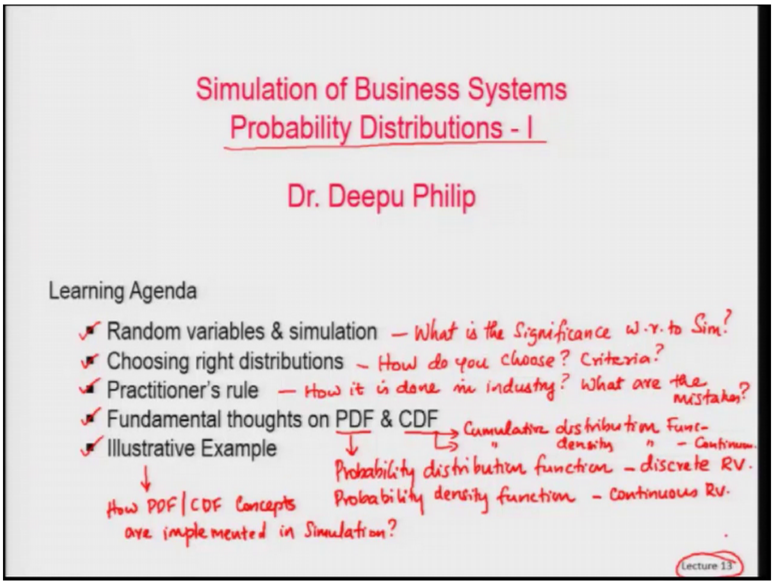
Good morning students. Welcome to some of the concluding lectures of the Simulations of Business Systems course. And we have been going through quite a lot of tools, techniques, and topics that are very pertinent to the business world. And we have seen what is the simulation, different aspects of simulation, different considerations in simulations, some of the software that are available, expensive software, another options are available in simulation.

We also saw some of the simple things like Monte-Carlo simulations, how to use hand simulation to study simple systems, how to use arena to study complicated systems etcetera. And we also seen one of the very advanced software tools called as trigonometric, which is made by Siemens, which is very expensive close to couple crores of licence rupees in licence fees.

And since, it is a short course we are now going to conclude into major aspects of it. And one of the things that we did not cover quite a lot in the class is about probability distributions, we were just using them exponential another things without getting into much details of it. So, we thought it is better to conclude the lecture by giving you a good idea about probability distributions. And hopefully in that process, you will be able to understand the importance of it.

But, I still recommend you guys this is not a probability in statistics course. So, hence there is a limitation to how much of probability distributions can be covered. So, whatever will be covered in this class will be pertaining to what is relevant in the simulation work ok. And I am assuming that somewhere or other you guys will somehow learn the basics of it. And let see what all things I can cover, and when do we use what simulation distribution in the simulation is what I will try to cover today.

(Refer Slide Time: 02:13)



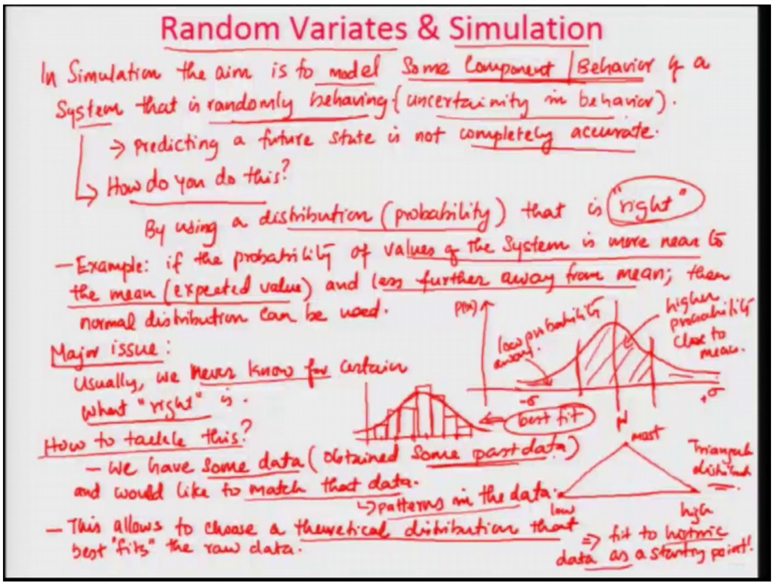

So, if we look in the today's topic what we are talking is Probability Distributions 1. And I will try to bring a second small set of presentation probability distributions too in which we will have some more some more details of the distribution. The major topics learning agenda today is we talk about random variables and simulations or random variates. Why what is the significance with respect to simulation? This is what we will be discussing today.

And then the question is how do we choose second one choosing the right distribution. How do you choose the criteria, what are the criteria that is what we actually will be doing in the second part of it. Then there quite there is lot of books depending upon which book you refer, you will get different value. So, in this case, what we are trying to do is how it is done in industry the practitioner's rule, what do what do the industry do ok? And what are the mistakes ok? So, this is the main aspects that we will be covering here.

Then we talk think about the PDF and CDF, which is probability distribution function. So, this is the probability distribution function, this is for discrete random variables, and also it is probability density function for continuous RV continuous random variables. Similarly, CDF is this is you can think about it has cumulative distribution function which is for the discrete, or it can also be the cumulative density function for continuous ok. So, we will see the fundamental thoughts on PDF and CDF, and how that can be used in the simulation aspects of it.

And then we will look at a simple illustrative example ok. So, this kind of demonstrate, how the simulation. The aim is to how PDF CDF concepts are implemented in simulation implemented in simulation, so that is what we will be covering today in this present lecture. I gave it as lecture number 13, even though (Refer Time: 05:31) one of the concluding lectures. This is kind of the sequence of power point presentation that I have been doing, and you have seen couple of experts, who have anyway come in to the class, and give you different exposure to different other topics.

(Refer Slide Time: 05:50)




So, without wasting much time, let us get into a basic definitions of it. The first thing that we talk about is random variates and simulation ok. And we are talking about random variates sorry especially with respect to the simulation. So, in simulation, we can always say this. In simulation, the aim is to model aim is to model some component or behaviour of a system that is random, or that is randomly behaving.

So, major aim is to model some component or behaviour of a system that is randomly behaving, or you can say that there has uncertainty with it uncertainty in behaviour ok; we are not sure how it will work ok. We cannot really so, this means predicting a future state is not completely accurate. You may predict that the future of the system might look something like this, but it could completely be something else.

So, what we do is in the aim is to model some component or behaviour of a system that is randomly behaving that is random. So, the question then is how do you do this, how do you do this ok. And the answer to this question is by using a distribution what distribution probability distribution a probability distribution that is right. So, for example, if the example if the probability of values or values of the system is more near to the mean or expected value, and less further away from mean, then normal distribution can be used ok.

So, what it actually means is the normal distribution, you have a mu mean here, and here is your standard deviations sigma the plus sigma, minus sigma side. And distribution will kind of have a shape something like this. This means that you have more probability, this is where higher probability close to mean ok. So, you have more probability that are closer to the mean, and this is low probability away from mean ok. So, such kind of a distribution can be used to model such a behaviour of a system.

See if the probability of the values of the system is more towards the mean or the expected value, this is in the mean ok. If the values are more towards this, then you can use a distribution like this, where it is larger probabilities. So, the y-axis of this has probability P of x is available here or f of x you can draw, whichever way you want to draw. And the probability is less, if it is further away from the mean. So, this is the lesser probability that we are talking about both sides. So, if that is the case, then we can use a normal distribution to do this.

But, the problem is what is the issue major issue, major issue is this that. Usually we never know we never know for certain what right is. So, one example is that some people can also model the same thing with using what we call as a triangle distribution. We can say here is a low value, here is a high value, and here is the most observed value, then people can form a triangular distribution like this, and say that this will also replace this. Now, this is called a triangular distribution. Now, whether the normal is better or triangular better or which one is right that is very hard, this is this is why what the right is a tough question for us to answer in this particular regard.

So, how to tackle this ok? The way to tackle, this is we have some data, or obtained some past data ok. So, somewhere we obtained some past data, or we have some data, and would like to match it like to match that data. So, somewhere somehow we obtained some data, some past historic data is available.

And using that, we would like to use that data. And from there, it is plotted using different graphical systems, and try to find out the patterns available with that that is why we studied graphical tools in the previous stuff. And then we decide patterns the data. So, from here, what you are trying to do is this is where you are trying to do patterns in the data ok. So, you are trying to look for how the data actually behaves ok.

So, ideally speaking this allows to choose a theoretical distribution that best fits the raw data best sorry fits best fits the raw data. So, let us say for example, you ended up plotting this using a graphical tool, let say you had various classes like this, and then you plotted it something like this. Let us say this is how you plotted it. Then you can say that to some extent, you can kind of think about it as a normal distribution like this ok, so that is kind of you will say this is somewhat somehow fits the data. So, this is where the best fit theoretical that is what we try to do in this regard.

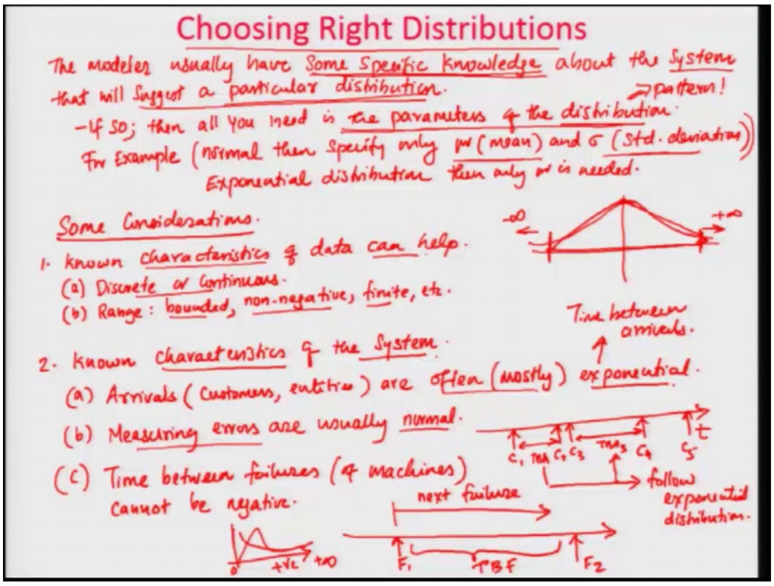
So, assume that, you guys understand what I am trying to say in this particular case. But, the aim here is that you have a you are trying to be study some component or behaviour of a system that is random ok, which means there is uncertainty in this behaviour you have, there is lot of uncertainty that is associated with this. And you are trying to find out you are trying to because there is uncertainty, you are not completely sure what is going to happen in future.

So, to model the system, what you do is by use a probability distribution is the thing that you think, it is the right probability distribution. And like for example, I give you the when people use a normal distribution stuff like that, but the biggest issue in this regard is it is never know for certain what the right is it is quite hard to figure this out.

So, how do people tackle this problem? Usually you obtain some past historic data would like to use a graphical system to plot it, and would like to find the patterns in the data that will actually match the data. So, just I showed you here. When you plot a histogram, and then you can see that maybe sometimes a normal distribution would fit better than a triangular distribution.

So, what happens is once you look for the patterns, then you choose a theoretical distribution that best fits the raw data. So, the aim of the simulation is fit to historic data as a starting point, so you have to start some place. And the starting point in this regard is first start by fitting to the historic data, and then from there do the improvement.

(Refer Slide Time: 16:33)




So, then the next question obviously, is choosing the right distributions ok. Once you plot the graph another things, they are quite a lot of empirical distribution from which you know we can do that. The first thing is that the modeler usually have some specific knowledge about the system, some specific knowledge about the system about the system that will suggest a particular distribution ok.

So, what happens is you may have some very specific internal knowledge about the system ok, some specific system knowledge, which will kind of suggest your particular distribution ok. If you have that particular information is so, then all you need is the parameter of the distribution is the parameters of the distribution. For example, normal then specify only mu, which is the mean, and sigma which is a standard deviation. So, if you think that you are going to use a normal distribution, then all you need to do is specify the mean and the standard deviation.

So, if we have some prior information about a particular distribution, then all you need is the parameters of the distribution. So, this is the advantage of using a distribution is there, once you can say this is the distribution. So, this distribution this is the pattern ok. If the identify this is the pattern, then all you need to do is you specify the parameters of the patterns. For a normal distribution, it is the mu understand the deviation. If it is an exponential distribution ok, then only mu only mu is needed ok, or the mean is needed ok, so like that. So, depending upon which distribution you pick up, you will specify only the parameters. So, this distribution means the pattern, depending upon the pattern which actually comes from some specific system knowledge.

So, some of the rules in this case, some considerations ok, number 1 known characteristics known characteristics of data can help ok. So, if you know some of the characteristics of the data, then it can help. Specific knowledge includes ok, discrete or continuous ok. If you think it is a discrete system, then you may not be able to use a normal distribution, you might have to look at something else. Then some other part is range, which includes bounded, non-negative, finite etcetera.

So, for example is, is your distribution go into infinity, for example in if we draw a normal distribution, this distribution actually we draw like this, but it actually goes to infinity ok. This is the minus infinity side, this is the plus infinity side. But, in somebody says, it is does not go to infinity, it is a bounded distribution, means this is the stopping point, this is a stopping point, it does not go beyond that. And this is the mean value, then you might want to think about a triangular distribution this way as well, it is bonded ok.

So, may be if it is unbounded, you can think about normal distribution; with a bonded bond, you can think about a the triangular distribution as well. So, these kinds of characteristics, whether it is discrete or continuous, whether it is bounded, whether it is non-negative. If it is non-negative, then you might not be using a normal distribution, so you might have use a Weibull distribution ok. Whether it is finite, infinite, so all those kind of things will help you to determine this kind of characteristic of data can help you determine, which distribution to use ok.

Number 2: known characteristics of the system known characteristics of the system. If you know some characteristics about the system ok, so some of the characteristics that can help is arrivals ok, customers, entities ok, so these kind of arrivals. Arrivals are often or you can think about as mostly arrivals are often exponential ok, which means this means time between arrivals.

So, if you think about it if this is a time axis going on like this. This is a particular point, where customer 1 arrived, then there is another customer arrived customer 2, then customer 3, then there is customer 4, then customer 5 like this happening. So, this is the

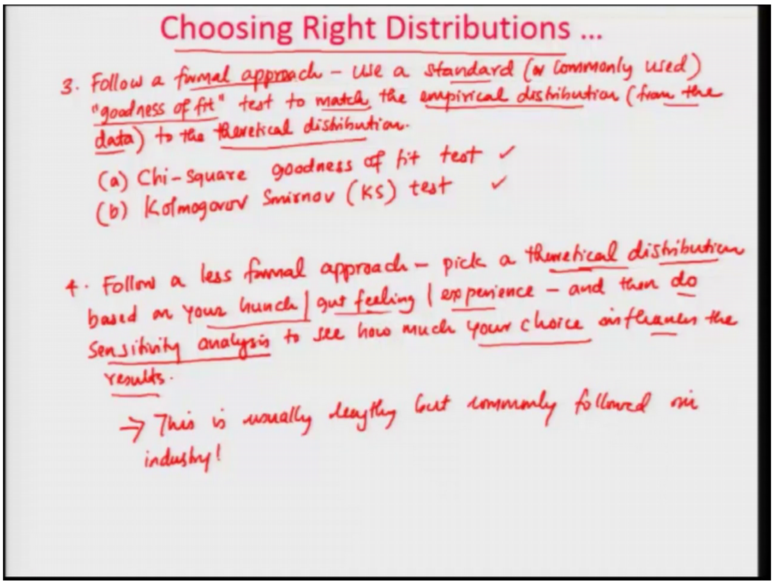
TBA time between arrivals, this is another time between arrivals TBA 3 or something like this. So, these are all these TBAs they follow exponential distribution ok. So, this is what we are talking about here is that arrivals are usually most of the time mostly there mostly exponential right.

B next one measuring errors are usually normal that means, you would use a normal distribution, if you are quantifying somewhere like a quality, (Refer Time: 23:52) where you are checking something and rejecting a product is something like that, then you would like to model this with a normal distribution. Normal distribution where the so, measuring errors means most of them will be close to the normal value or the mean value, and that they will change accordingly ok.

Then the 3d one is time between failures of machines or anything or machines cannot be negative ok. So, you cannot if the machine if you think about this way, if here is a failure machines failure 1, then you cannot have a failure before this. The next failure can only occur; next failure can only occur in a positive direction. So, it could happen some place here as a failure 2. So, this is the TBF time between failures ok. So, this time between failures cannot be negative, so that means, you cannot use a normal distribution here, you might have to use a distribution that is positive in nature.

So, here if you are modelling that, then your graph will probably be this is 0, and this is positive infinity. And maybe you might have a distribution something like this, which you might say the probability of the next failure or you might even say that I might use an exponential does not mater, which some through some care you will have, but the most important thing is this will be a positive value ok. So, the time between failures cannot be negative; so, these kind of internal informations that you have is very important for you to choosing the right distributions.

(Refer Slide Time: 25:41)



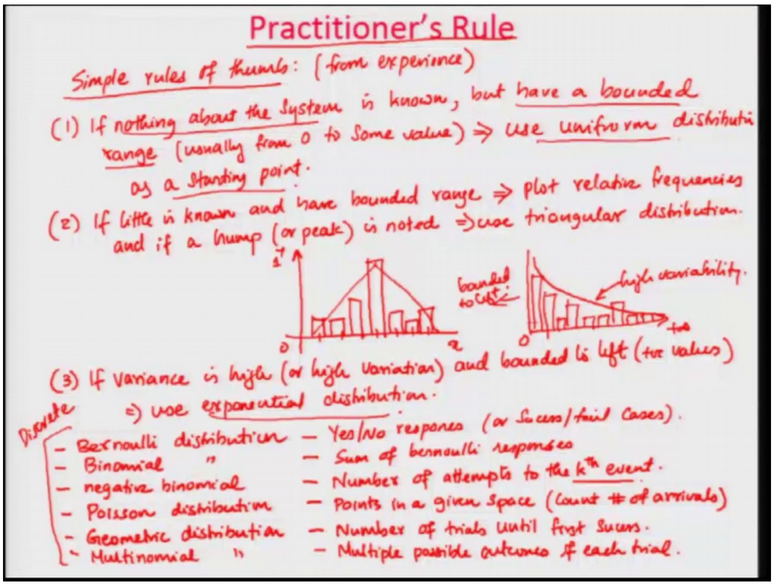

A continuing in the choosing the right distributions process, there is some more thing ok. So, the 3rd point that is usually very prominently used by lot of the practitioners is. Follow a formal approach that is which is use a standard or commonly used, this is a standard or commonly used goodness of fit test goodness of fit test to match to match the empirical distributions empirical distribution ok. The empirical distribution come from the data to match the empirical distribution to the theoretical distribution. So, whatever the data that you are getting the empirical distribution, which is obtained from the data and to match with the theoretical distribution use a standard or commonly used goodness of fit test.

The two most commonly used goodness of fit tests are chi square goodness of fit test chi square goodness of fit test ok. The second commonly used one is Kolmogorov Smirnov test Kolmogorov Smirnov, it is commonly known as KS ok. So, these are the two properly used or prominently used goodness of fit test. So, use either one of this test, and then try to match. You match using this test; you try to match the empirical distribution with that of the theoretical distribution ok. This is the formal approach follow the formal approach, there are people who do this.

Also may be follow a less formal approach. This is also another option of doing it. Pick a distribution a theoretical distribution based on your hunch, hunch, gut feeling or experience you name it whatever you want ok. Pick a theoretical distribution based on your hunch gut feeling or experience whatever it is, and then do sensitivity testing sensitivity analysis or testing to see how much your choice influences the results ok.

So, what you do is you pick a theoretical distributions, some theoretical distribution is picked based on your hunch, gut feeling or experience. People usually use the data to validate the gut feeling or experience. And then perform a sensitivity analysis to see how what is the impact of the choice, how is the choice, the choice of that particular distribution influence the results. This is another approach that is also being followed. This is usually lengthy, but commonly followed in industry ok, we sometimes nick name this as brute force approach to a large extend. Now, we seen that the four kind of methods of choosing the right distributions.

(Refer Slide Time: 30:27)




I will try to give you some practitioner’s rules in this regard ok, how do people who are practitioners, or who are more involved in simulations actually do this. So, I will say it as simple rules of thumb, there is no what you call as scientific backing or anything to this, but these are from experience ok. People who work in the simulation area, these are some of the rules that we do ok.

The first thing if nothing about the system is known system is known, if you do not know anything about the system, if nothing about the system is known, but have a bounded range ok, this kind of a system. You do not know anything about the system, but you know that there is a bonding, you cannot go beyond that ok, usually from 0 to some value.

You talk about the production in a factory ok, the minimum a factory can produce is 0, such bonded in the minimum. The maximum, we can produce is probably in an 8 hour shift 100, 200, 300, but you cannot never producer infinity, there is some level to it ok. If that is the case, then we use is use uniform distribution as a starting point. So, we start with the uniform distribution, you do not know anything about it, but there is a bound bounded range, then we start with a uniform distribution.

Number 2: If little is known, if little is known, and have bounded range ok, then we can say first is plot relative frequencies, and if a hump is noted or peak is noted ok, then use triangular distribution. If it is bounded ok, so let us say you have a minimum of 0, and maximum some values, and you plot it. So, if you do something like this an example would be you take the relative frequencies, so this is like x, and your y is 0 to 1 (Refer Time: 33:46) relative frequencies. And you plot something like this, you get a observation something like this something like this, then we can basically think about it as approximating it with help of a triangular distribution.

Number 3: If variance is high or high variation, and bounded to left, which means positive values only, use exponential distribution; so if you are thinking about a scenario, where you plot the values, and you get something like this ok. Let us say somewhere you get a frequency stuff like this, then you can think about it as a exponential distribution. This shows there is high variability in the system high variability is also shown here ok. And this is 0, and this is positive infinity this way ok, or some particular value, but this is anyway bounded to the left, which means it there is no values here bounded to left. This is the case then you use a exponential distribution right.

So, so I will give you some common distributions ok. So, common distributions, and what are what do they mean ok. So, for example, the first one is Bernoulli distribution. Please read about these distributions Bernoulli distribution, and it is used for yes, no responses ok, or success, fail cases. If we are doing a quality control stuff, where you are doing rejecting and those kind of things, the rejecting bad products, then Bernoulli is a good idea.

Then you have is Binomial distribution ok, it is used for some of Bernoulli responses ok. So, this means that if you inspect 20 bulbs, what is the probability that 2 bulbs are bad, and that kind of questions ok. So, sum of Bernoulli summed up is binomial. Then you have is something called as negative binomial ok, it is used for measuring number of attempts to the k-th event ok. If you want something special something to happen some explicit k-th event to happen, then how many attempts are required, how many attempts are required to reach the k-th event or obtain the k-th event is basically can be modelled using negative binomial distribution.

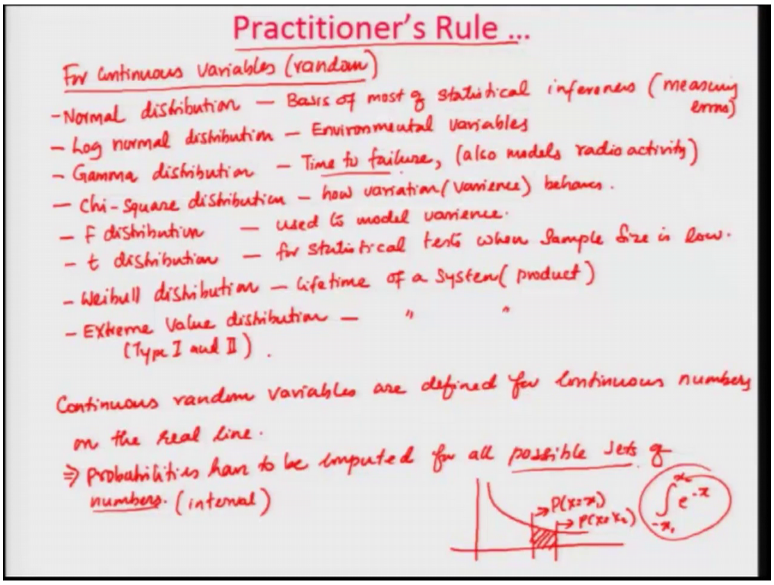
Then you have something called as Poisson distribution ok. And Poisson distribution is typically used for you know points in a given space ok, and many people also called as count number of arrivals, how many people arrived in a particular time period given the rate of what it is. So, we typically called as point in a given space that is a limit, and within which how many points, or how many arrivals, how many customers, all those kind of things are very well done with the help of Poisson distribution.

Then we have is Geometric distribution. I am talking only about the major distributions nothing else. Geometric distribution is basically measures number of trials until first success ok. So, if you are tossing a coin or rolling a dice, and if you consider 3 as getting 4 or more as a success, then how many trails are required before you get a success that can be modelled using a geometric distribution. So, if you are getting a raw material from a vendor, and if you are trying to find out how many parts need to be inspected before you get the first defective part that can also be model very well using a geometric distribution.

And the last one that is pertaining to us I will write today as the Multinomial distribution multinomial distribution. And what it does is it basically models the multiple possible outcomes of a each trail outcomes of each trail. So, what you are doing here is let us say you are in a process of checking the painting of a car, and one painting it can be good, another can be touch up, third can be repaint, forth can be discard. So, every car there is multiple options available to this, and which option will happen in the next one depends upon the way that can be very well modelled using a multinomial distribution right.

Now, continuing in the practitioner's rule, you should notice that in the previous case, all of these all of these can be thought about it as these are discrete cases or discrete distributions ok, we are talking about counting individual each observation is countable.

(Refer Slide Time: 40:21)




Now, when we talk about for continuous cases for continuous variables or random variables, when I say variables, I am talking about random variables. For continuous cases, if we the major distributions and we are talking about is normal distribution ok, we can say it is basis of most of statistical inferences heavily used for measure measuring errors etcetera. So, normal distribution is used whenever you have a scenario, where the probability of occurrence is an unbounded, there is no bound in this case and probability of occurrence is around the mean value is very high, then that is where we use a normal distribution.

Then we there is let say there is a distribution called Log normal distribution. And Log normal distribution is commonly used to study environmental variables ok. We are modelling in studying environmental factors or environmental factors, then (Refer Time: 41:57) tend up using Log normal distribution ok. Gamma distribution ok, where commonly used to study time to failure ok, this also models radio activity, studied about the gamma decay. Gamma function that suggest about how the radio activity actually works, but mostly time to failure is very well modelled as gamma distribution.

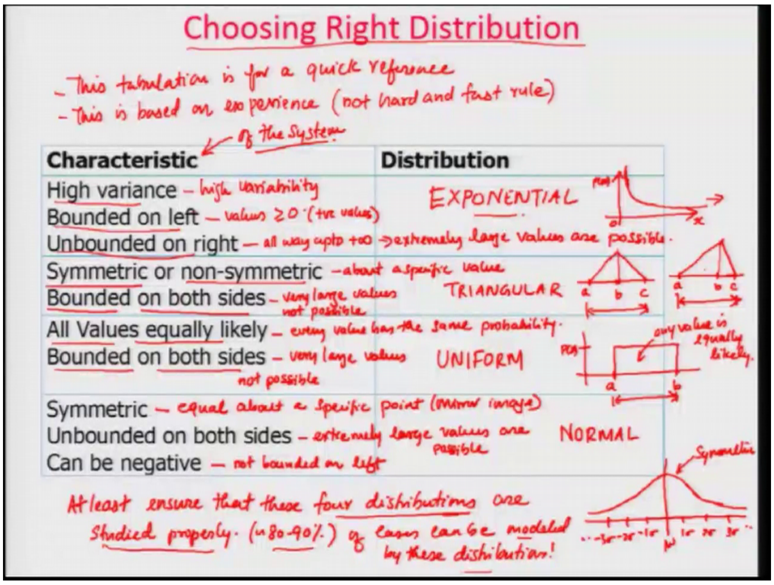
Then there is something called Chi-square distribution, not chai square Chi-square distribution. Chi-square distribution is usually you know it is a family of distributions, but you can say that how variation or variance behaves ok. When we are trying to model variances, then you can use Chi-square distribution. We can also talk about F distribution. This is also another specific case of Chi-square distribution. And this also is used to model variance ok.

Then you have t distribution. This is the student some people called also as the students distribution for statistical tests test, when sample size is low ok, so that is the t distribution aspect, when you have law sample size that is when you use this. Then you have something called as Weibull distribution; Weibull. Weibull is to use a study life time of a system or a product right. If you are trying to study life time of a system, there are product, then you use Weibull distribution. Then you can also talk about as extreme value distribution, it is another one that is used ok, both type 1 and type 2 ok. These are also used for lifetime of a system or a product studies. So, these are all for the continuous random variables (Refer Time: 44:50) talked about ok.

So, what do so, just to say continuous random variables continuous random variables are defined for continuous numbers continuous numbers on the real line on the real line ok. So, the number line when we have the real line real number line that that is where it is. So, here we say that implies the probabilities have to be computed for all possible sets of numbers. So, here we are more look that set of numbers or an interval of numbers interval of numbers ok.

So, you are trying to so, for example, if you are doing a Exponential distribution or Gamma distribution, which looks like this. If somebody asked what is the probability of this particular case probability of x equal to x that does not make any sense ok, we can call it as a x 1. And you can have another probability, probability of x equal to x 2. Between this, what is a probability between this can be calculated by integrating minus x 1 to x 2 e to the power of minus x or what the whatever the function that you want to call it ok, and then we can get the value accordingly all right. So, the continuous probability distributions for the continuous cases, these are the specific things that you actually work on.

(Refer Slide Time: 46:48)




Now, if I tabulate all of these things into a into a particular format ok, choosing the right distribution as for so this tabulation is for a quick reference ok. This is based on experience ok; this is a not hard and fast rule ok. So, consider these as sorry in rule consider these as guidelines ok. These are guidelines for you to decide how which distribution to use ok.

So, this is the characteristics, which means characteristic of the system. What are the characteristics that you are noting ok? So, the first thing that you do is high variance or some people also call it as high variability. There is high variance in the system and bounded to the left, which means values is greater than equal to 0 that is high variability in the system, the values are greater than on 0; unbounded to right, which means all way up to positive infinity.

So, if there is high variance or high variability in the system or the points are highly varying from the mean value is bound bounded to left, which means the all values are greater than 0. This means positive values only positive values. And it is unbounded to right, which means a very extremely so unbounded to right also means, extremely large values are possible ok, so that is what unbounded to right means very large value extremely large values are possible. If that is the case which distribution will use, the distribution to use is exponential ok. You would end up using an exponential distribution, if the you know extreme large values are possible in this regard ok, so as a first rule of them high variability, high variance, bounded to left or positive values, unbounded to right or very large values are possible, then you end up using an exponential distribution ok.

Then either it is symmetric or non-symmetric ok, the so which means symmetric about a about a specific value ok; and bounded on both sides ok, so it is very large values not possible. Then if this is the case, if it is symmetric or non-symmetric bounded on both sides, then use a triangular distribution. So, the exponential distribution if you draw, it will look like this P of x distribution will look like this ok, it will continue like this, so this is 0. So, it is bounded no negative values are allowed and only positive values, and it goes all the way to infinity right.

A triangular distribution, you can draw if it is like this, let us say you have a minimum value, maximum value, and the common value, then you can draw triangular distribution like this, which means it can be symmetric or you could have a triangular distribution, which is a, and you have a b, and you have a let say c, so then you can have a triangular distribution of this fashion, which is non-symmetric. This is symmetric, this is non-symmetric that kind of is, but these are this is bounded both the cases are bounded within the levels all right.

Now, the 3rd one ok: All values are equally likely that means, and bounded on both sides ok. So, there is very large values not possible ok. Every value has the same probability ok. If that is the case, all values have the same probability and very large values are more possible. If that is the case, then you use uniform distribution ok.

A typically uniform distribution is drawn something like this, you have a value a, and you have a value b, and you have a probability of x, and all values between this will have the same probability, any value between this any value is equally likely. And these are bounded. Equally likely means, they all have the same probability of occurrence. So, then the uniform distribution allows you to model a system, which are bounded on both sides extremely large values are not possible, and all values within that are equally likely ok.

Then the last common case that we use is if it is symmetric, symmetric means you know equal about a specific point or symmetric or you can think about as mirror image ok. And it is unbounded on both sides extremely large value is possible large values are possible ok. And can be negative, so not a bounded on left.

If this is the case, you use a normal distribution ok. The normal distribution is something like this you have a mu, and you have a standard deviation, and you draw it as a symmetric distribution about the mean ok. So, this is the symmetric ok. And you have variance on both sides 1 sigma, 2 sigma, 3 sigma like this, minus 1 sigma, minus 2 sigma, minus 3 sigma like this ok, so that kind of a distribution is where we use as a normal distribution ok.

So, these are like so at least at least ensure that these four distributions are studied properly ok, about 80 to 90 percent of cases can be modelled by these distributions ok. So, what I am saying is that if you know these four distributions, if you study them properly, understand them from a good standpoint, then about 80 to 90 percent of the cases can be modelled using this distribution ok.

So, with this what we will do is we will take a quick break from of this today's lecture. And what we will do is we will continue the remaining of the lecture after short break.

Thank you.