**Essentials of Data Science with R Software- 1**
**Professor Shalabh**
**Department of Mathematics and Statistics**
**Indian Institute of Technology, Kanpur**
**Lecture 42**
**Exponential Distribution**

Hello friends, welcome to the course Essential of Data Science with R Software- 1, in which we are trying to understand the basic concepts of probability theory and statistical inference. So, now in this lecture we are going to understand the exponential probability density function or in general exponential distribution.

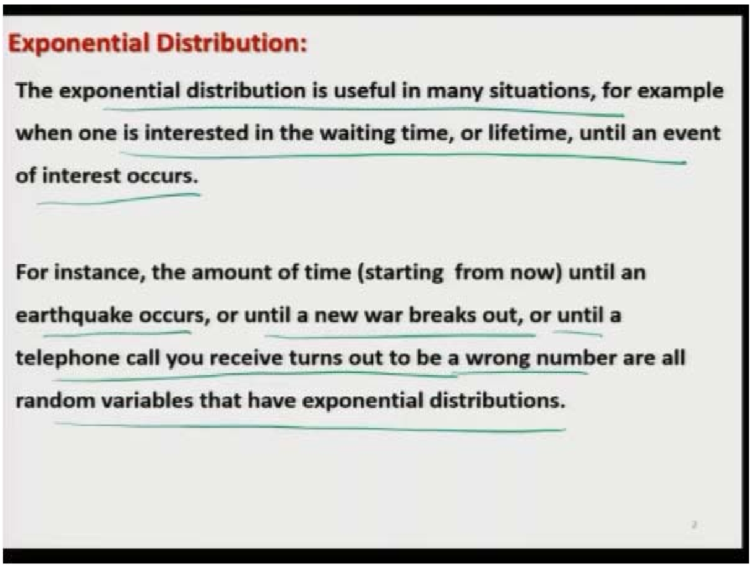
Now, you know as soon as I take the name of a new distribution you know what are we going to do. Number one, where it is going to be applicable, under what type of condition you can use it, then what is its PDF, certain properties, how to implement it in the R software, how to compute different types of probabilities and then how to compute those probabilities inside the R software also.

So, now there is no issue, so let us begin our lecture and first we try to understand that under what type of conditions you are going to use this exponential distribution. Tell me one thing, whenever you go to a shop to buy a bulb or a led bulb or an led light, what do you ask the shopkeeper, what is the life of this bulb, what does this mean?

This means you want to know that if you buy this bulb and if you start using it, how long it will go or when the first error will come because as soon as the error come the bulb become unusable, that will not remain as useful and similarly, if you want to buy any equipment you are always interested in finding out after what time the first error will happen.

Now, when it will happen that you do not know but you can quantify it in terms of probability. So, in case if you feel that the probability, that the first error will come, say after a long time then you will buy the equipment, but if the shopkeeper says that the bulb is not going to last for more than 2 days, that means the first error will occur only after say this 18 hours of the purchase of the bulb, then are you really going to purchase it, but you do not know whether the error is going to be after 18 hours or not so you simply try to compute the probability, that what is the probability that the bulb will have the first error at the $18^{th}$ hour. So, now these types of instances, these types of situations can be modeled through the exponential distribution. So, let us try to understand it.

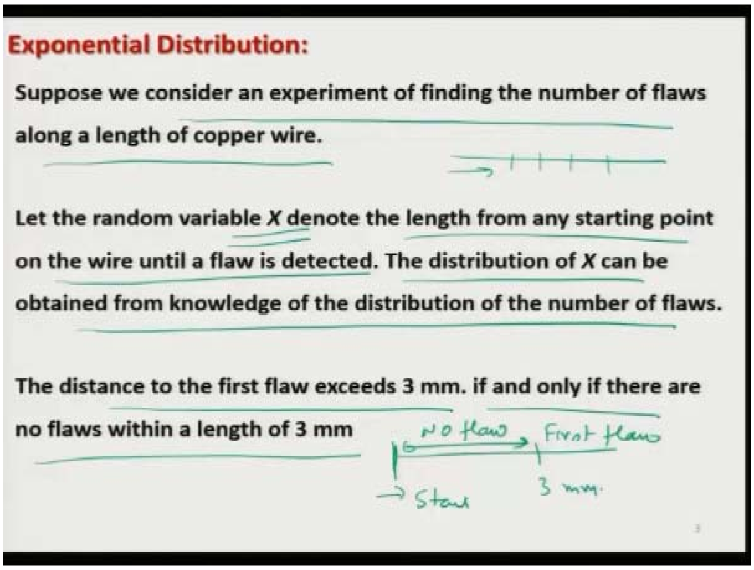
(Refer Slide Time: 2:59)




So, this exponential distribution is useful in many situations, for example, when one is interested in the waiting time or lifetime until an event of interest occurs. For example, the amount of times starting from now until an earthquake occurs or until a new war breaks out or until a new telephone calls you receive and that turns out to be a wrong number, these are all random variables, why they are random variables?

Because if you try to take this example that you have to wait until the telephone call you receive that turns out to be a wrong number, you know that when it is going to happen that you do not know, that may happen even after 1 second, 1 minute, 1 hour, 2-hour, 3 hour you do not know. So, in that sense this is a random variable and such random variables can be modeled through the exponential distributions.

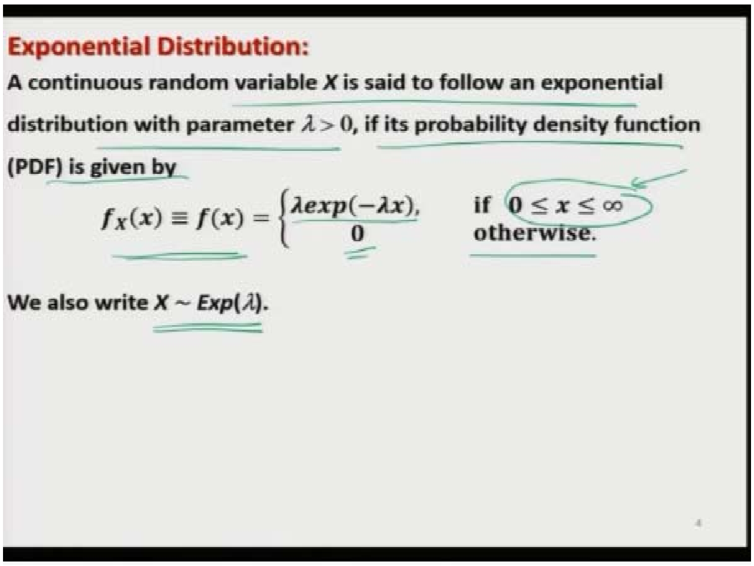
(Refer Slide Time: 3:59)




So, let me try to take one more example, suppose, we considered an experiment of finding the number of flaws along the length of copper wire. So, now suppose this is here the wire, now you start looking into the wire, the first flaw comes either here or here or here or here you do not know, because you have not seen the wire earlier. So, in that sense this is going to be a random variable.

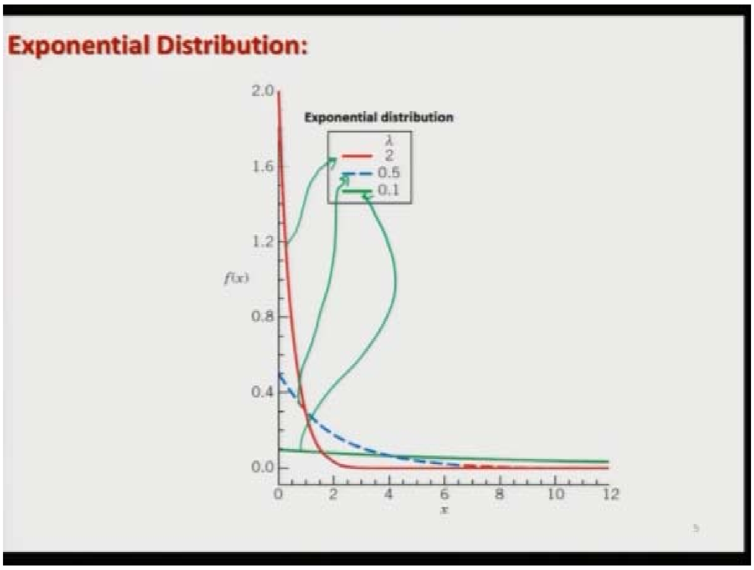
So, let this capital X be the random variable that is indicating the length from the starting point on the wire until a flaw is detected. And the distribution of such an X can be obtained from the knowledge of the distribution of the number of flaws, the distance to the first floor, for example, exceeds 3 millimeter that can happen if and only if there are no flaws within a length of 3 mm. For example, if you are starting from here, you start here and now this is here the point of 3 mm. So, there is the first flaw here, so this will occur if there is no flaw before that. So, under these situations the probability density function of exponential distribution can be very well used to characterize such phenomena and compute such probabilities.

(Refer Slide Time: 5:29)



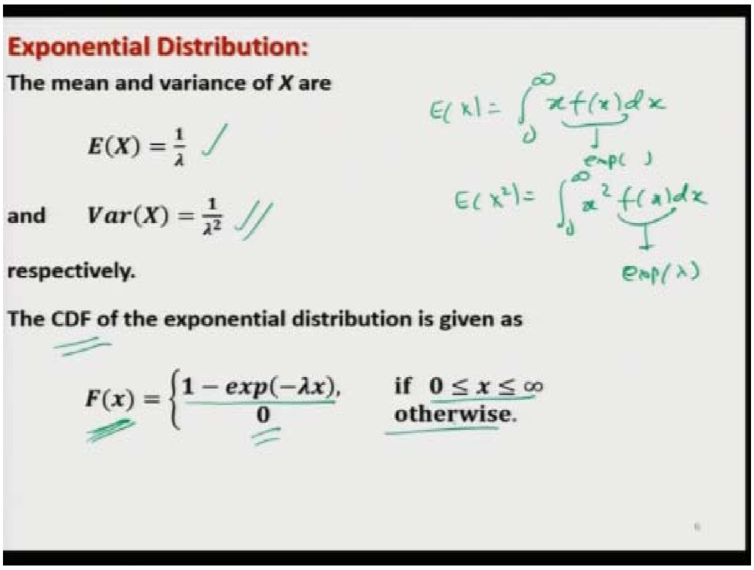

So, a continuous random variable X is said to follow an exponential distribution with the parameter λ greater than 0 if its PDF is given by like this, $f(x) = \begin{cases} \lambda exp(-\lambda x), & \text{if } 0 \le x \le \infty \\ 0 & \text{otherwise.} \end{cases}$. And that is obvious when you are trying to find out the lifetime like quantities, then definitely they cannot be negative. So, that is why you can see here the range is also indicating the same that X is greater than or equal to 0. We indicate such a PDF by writing X ~ exp(λ), so that is indicating the parameter λ.

(Refer Slide Time: 6:19)




Now, in case if you try to plot this PDF you can see that for different values of λ this curve has a different types of shape, for example, this red color is indicating the curve for λ is equal to 2 and this blue color is indicating the curve for λ is equal to 0.5 and similarly here this green color is indicating the curve for λ is equal to 0.1. So, you can have an idea that how the curve will look like.

(Refer Slide Time: 6:51)

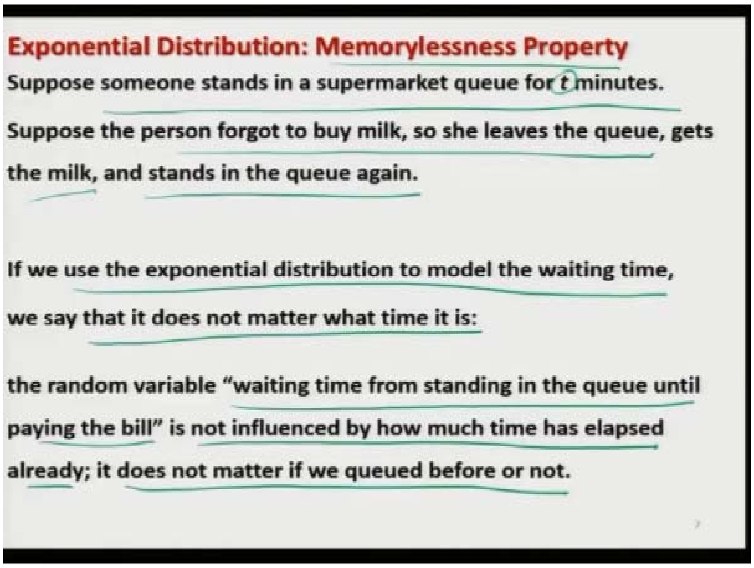


Now, in case if you try to find out its mean and variance the mean will come out to be $1/\lambda$ and the variance will come out to be $1/\lambda^2$. There should not be any confusion and then finding out these things is very difficult, for example, if you want to find out the expected value of X that is simply going to be integral 0 to infinity x into f(x) dx, now you have to simply substitute here the PDF of exponential distribution and solve it.

Similarly, you can find out expected value of x square 0 to infinity x square f of x dx and then use simply here substitute here the exponential PDF with parameter $\lambda$ and then you simply solve it and then substitute the values, you will find these values and similarly if you want to find out the cumulative distribution function that is the CDF of the exponential distribution, then it is coming out to be like as this

$$F(x) = \begin{cases} 1 - exp(-\lambda x), & \text{if } 0 \leq x \leq \infty \\ 0 & \text{otherwise.} \end{cases}$$

(Refer Slide Time: 8:00)



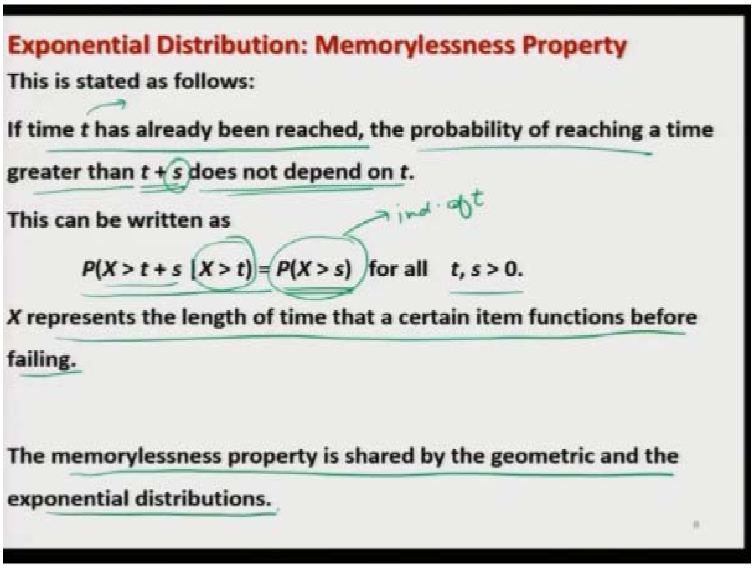

And now this exponential distribution has got a very important property Memorylessness Property and if you remember you also had seen that such type property is also present in the geometric distribution and there itself I had told you that there is one more continuous distribution which also has this property.

So, now that is the time where we are going to discuss about it. So, the first question comes here what is this memorylessness property, what is the meaning and what is the interpretation of this? So let me try to take a very simple example to explain it. Suppose, someone stands in a supermarket queue for t minutes and person has collected the things whatever he or she has to buy and while he or she was in the queue to make the final payment, the person realizes that he or she has not bought the milk. So, what will happen? The person will leave the queue, get the milk from the shelf and once again come back to the queue and stand in the queue once again.

So, now the question is this we always try to find under such conditions that what is the probability of the waiting time, waiting time means the waiting time for making the payment, so that after that you can leave the shop. So, we always try to think that if you have come late, then it is going to take a longer time, otherwise I would have been at the second place or third place by this time. So, this property is related to that type of event.

So, if we use the exponential distribution to model the waiting time, we say that it does not matter what time it is, the random variable that waiting time from standing in the queue until paying the bill is not influenced by how much time it has elapsed already and it does not matter if we queued before or not. It looks little bit surprising but it happens, that is the property of this exponential distribution.

(Refer Slide Time: 10:30)

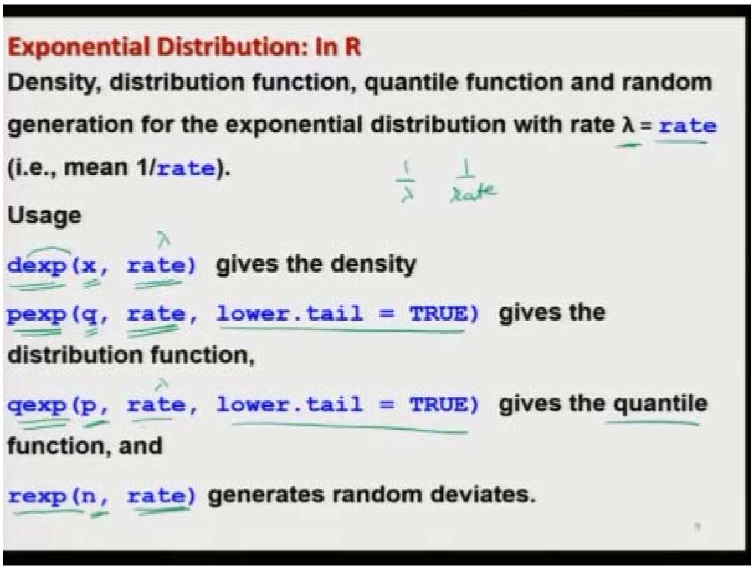



So, in case if you try to write down this property in a statistical way, then we can say like this, if time t has already been reached, the probability of reaching a time greater than t plus s does not depend on t, means you can say that t was the time for which the person was standing in the queue and then the person leaves the queue, gets the milk and again stands in the queue and after that it is going to take s units of time. So, t plus s will not depend on the time t.

So, this can be indicated in a probabilistic statement as $P(X > t + s \mid X > t) = P(X > s)$ for all $t, s >$. If this is given to us that X > t, then the P[X> t+s] this does not depend on t and that is why it is written as probability that X is greater than s. So, the right-hand side probability is independent of t and that is why we are trying to say that it does not matter that if the person has queued before or not.

So, here this X represent the length of time that a certain item function before failing if you try to take any other example. And I said that the memorylessness property is shared by the geometric and the exponential distribution. So, geometric is a discrete distribution and exponential is a continuous distribution.

(Refer Slide Time: 12:19)




Now, the next question come how are you going to implement it inside the R software? You can see that earlier first I was trying to explain you the PDF, PMF and then in the next lecture I was trying to take the implementation in the R, because earlier you were trying to understand it, you were trying to learn it, but now if you try to see you have understood each and everything, so now it is not difficult for me to explain you these things very quickly.

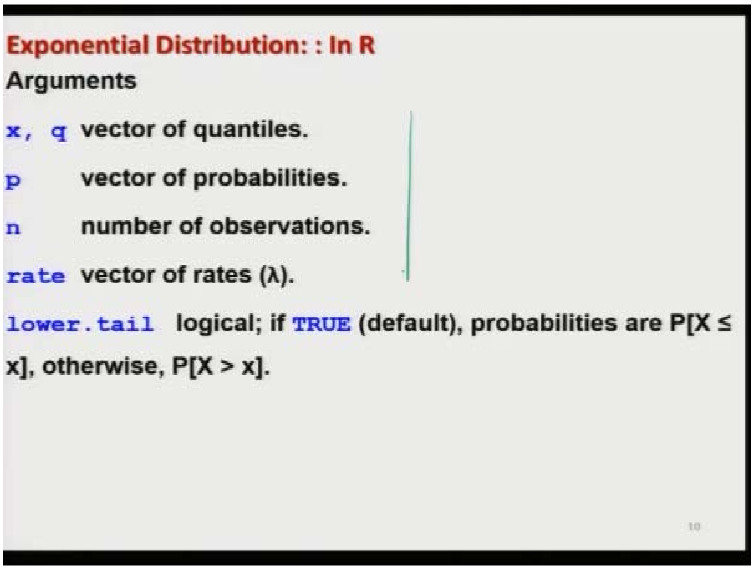
And the fact is this that after that if I take any more distribution that will take a very short time and that is why I have decided that after that I will not consider any univariate probability mass function or probability density function and I will leave up to you that how much you want to learn in this topic.

So, when you want to come to the R software for the implementation of exponential distribution, you have to keep one thing in mind, that here this parameter $\lambda$ is given by here rate. So, in case if you for example, want to find out here the mean for example, the mean was given by here $1/\lambda$, so you have to write down here 1/rate like this.

So, if you want to compute the density then you have to use the command here dexp, that means density from the exponential distribution and then you have to give here the value of x and then the value of here $\lambda$ as your rate and then if you want to compute the CDF then you have to give here the pexp, then the vector here q for specifying the value at which you want to find out the CDF, then you have to give the value of here λ and then you have to give here the value of lower dot tail is equal to TRUE or FALSE depending on your requirement.

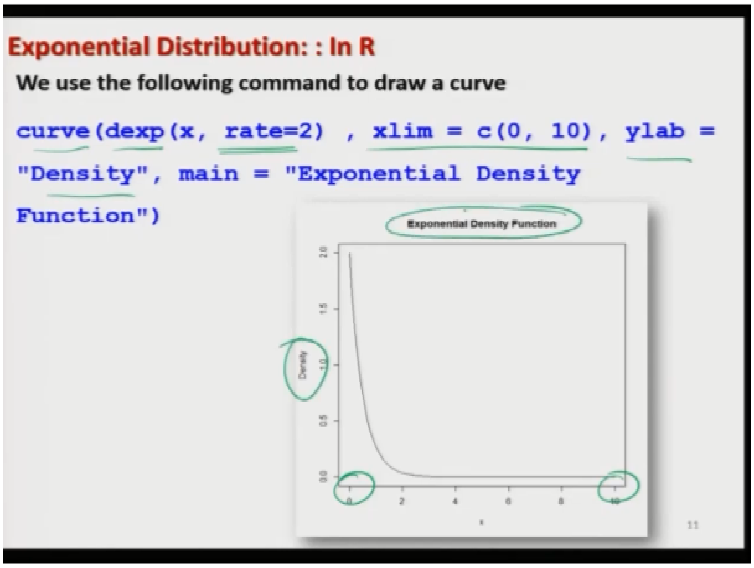
Similarly, if you want to find out the quantiles of this distribution then you have to simply give qexp with the value of the quantile that you want to find with the value of a λ as rate and then lower tail is equal to true or fall, depending on your requirement and if you want to generate the random numbers from this distribution you have to write down rexp with n, n is the number of observation that you want to generate and then the value of rate and then this will give you n random numbers from this distribution.

(Refer Slide Time: 14:45)



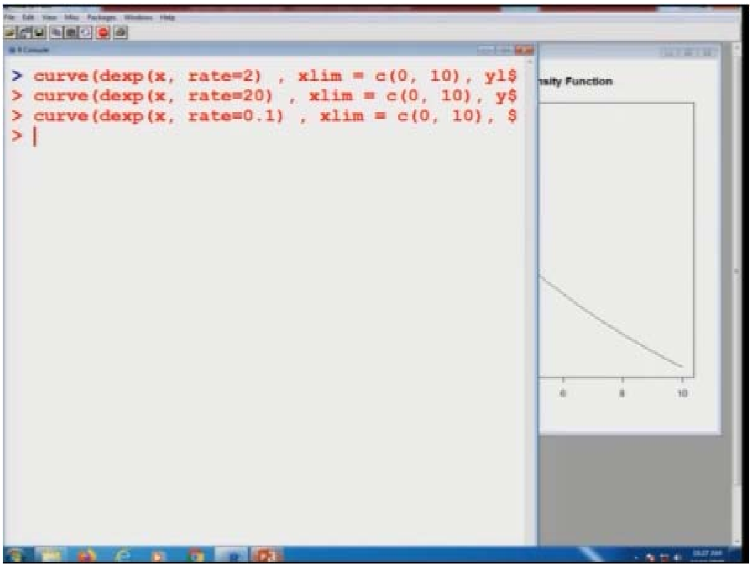

And these are the details of the same parameter that I just explained you.

(Refer Slide Time: 14:51)



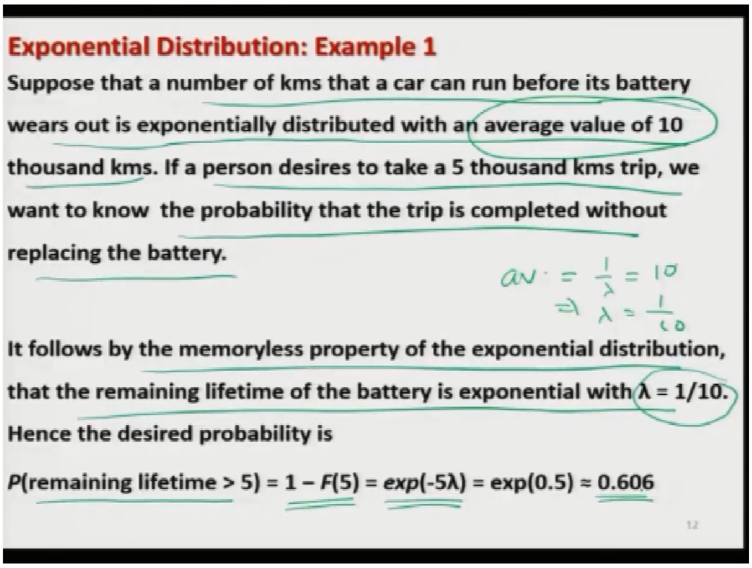

So, now let me try to show you that how are you going to plot this curve, if you want to see it for any particular value of λ, for example, if I try to write down here curve and then the dexp and suppose I want to plot here the exponential distribution for λ is equal to 2 this will come out to be here like this by writing rate is equal to 2 and then xlim that means I want to plot the curve between 0 and 10 on the x axis and on y I want to print that density like here and the main title exponential density function.

(Refer Slide Time: 15:30)



So, in case if I try to plot this curve on the R console it will look like this you can see here, and if you try to change the value of the rate here you can see here how the things are changing, if you want to say here rate is equal to suppose 20, you can see here the curve is changing like this and if you try to increase this value or try to decrease this value to suppose here 0.1 you can see here this value is like this, this curve will look like this. So, definitely this type of things will help you in having an idea that how the curve will look like so that the understanding becomes better.

(Refer Slide Time: 16:16)

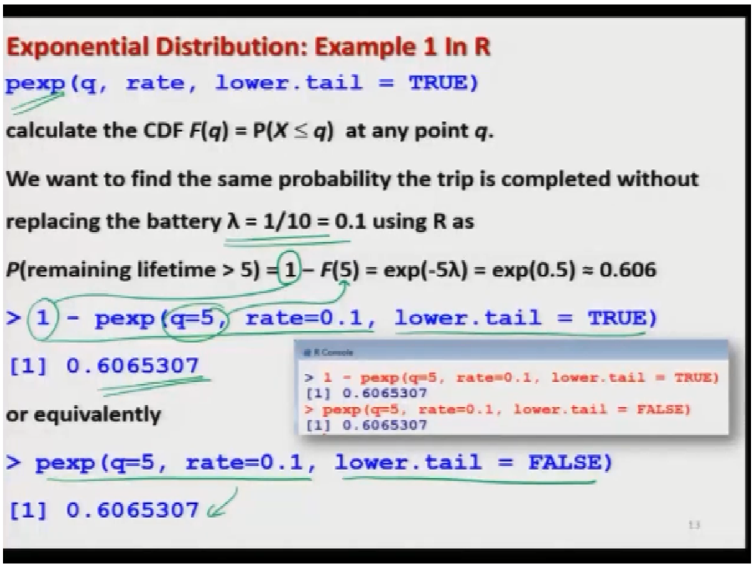



Now, let me try to take some examples, first I will try to solve those examples theoretically and then I will try to show you them on the R console also. So, suppose the first example is suppose that the number of kilometer that a car can run, that is the mileage, before its battery wears out is exponentially distributed with an average value of 10,000 kilometers, and if a person desires to take a 5000-kilometer trip we want to know the probability that the trip is completed without replacing the battery.

So, that is what we want to know here, so now it follows by the memory lessness property of the exponential distribution that the remaining lifetime of the battery is exponential with the parameter λ is equal to 1/10, because here you are given the average value, so the average value is 1/λ. So, average value here is 1/λ and it is given as 10, so this implies that λ is equal to 1/10, that you have to be careful.

So, in this case if you want to find out the probability then the probability that the remaining lifetime is greater than 5 that will be simply here 1 minus CDF at 5 and this is simply your exp(-5λ), λ here is 1/10. So, this value comes out to be close to 0.606, so there is 60 percent probability that the trip can be completed without replacing the battery.

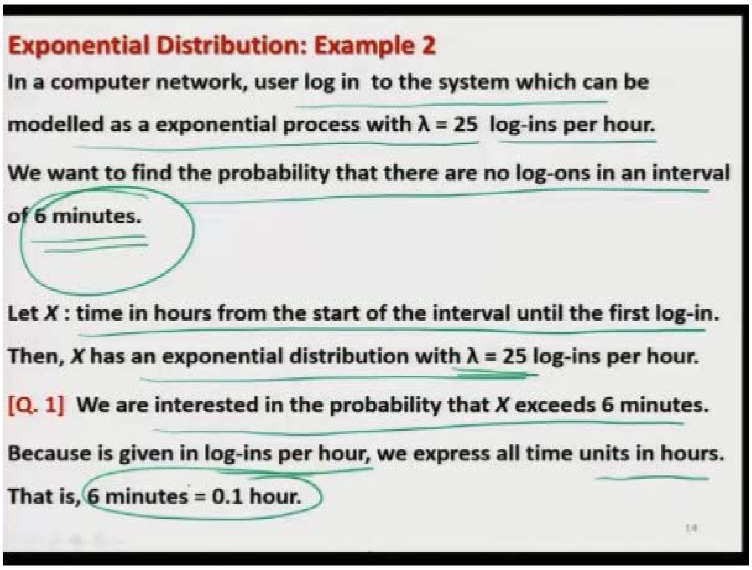
(Refer Slide Time: 17:53)




Now, if you want to solve the same problem through the R software how are you going to do it? You simply have to compute the CDF. So, in order to compute the CDF, you have to use the command here pexp and you have to use here λ is equal to 0.1 and so you can see here I write here 1 minus pexp q equal to 5 and rate is equal to 0.1 and lower dot tail is equal to TRUE. So, you can see here this 1 is corresponding to this one the probability that we have just found and this 5 is related to this F5.

So, now if you try to compute this probability this comes out to be simply here 0.606 and in case if you do not want to use this option of 1 minus CDF, then you can also use here directly the command that lower dot tail is equal to FALSE and this will also give you the same value here 0.60. So, you can see here computing such types of probabilities are not difficult at all in R.

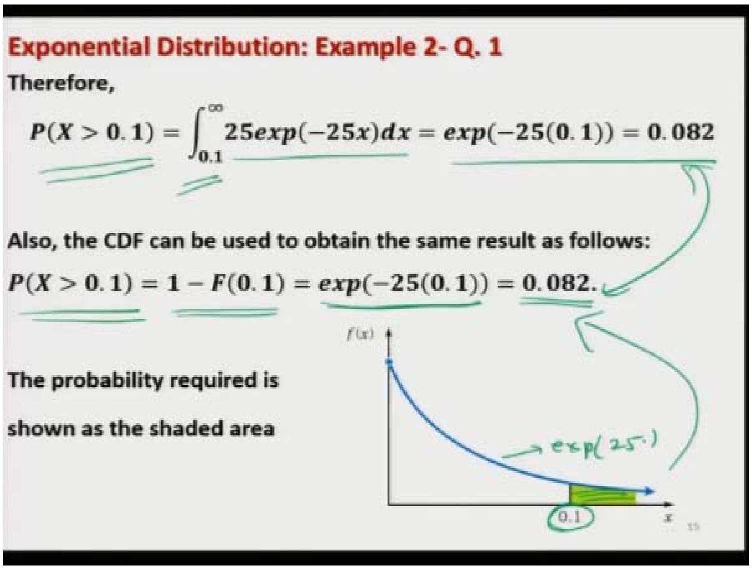
(Refer Slide Time: 18:56)




Now, I try to take here one more example and try to compute different types of probabilities. So, in a computer network a user login to the system and this process can be modeled as a exponential process with $\lambda$ is equal to 25 logins per hour, and we want to find out the probability that there are no logins in an given interval of 6 minutes, that means nobody logs into the computer for 6 minutes.

So, let capital X be the time in hours from the start of the interval until the first login occurs. Now, X has got an exponential distribution with $\lambda$ is equal to 25 logins per hour, because that is an average value. Now, the first question that we are going to find is that suppose, we are interested in finding out the probability that X exceeds 6 minutes and because the logins here are given in the units of logins per hour, so we would first like to express all the time units in hours. So, if you try to see here in the question, you are trying to find out here the answer for the 6 minutes, so the 6 minutes are going to be equal to 0.1 hour.

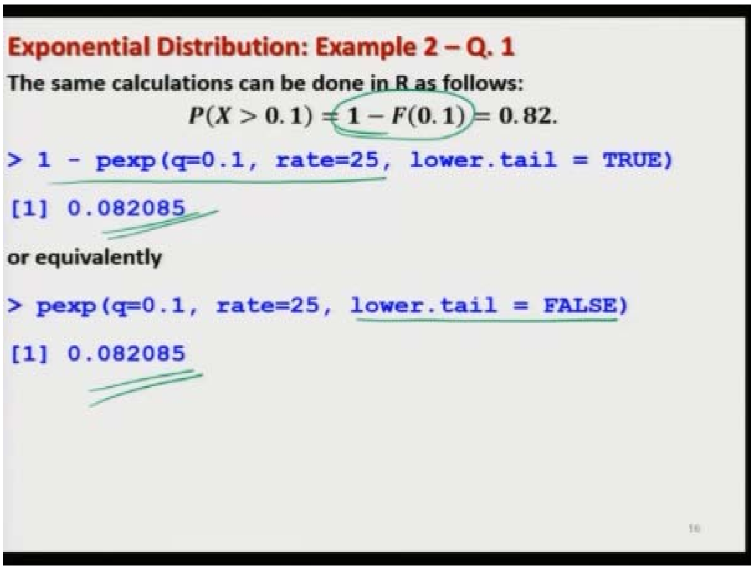
(Refer Slide Time: 20:20)




Now, I try to find out this probability. So, you can see here probability that X is greater than 0.1 that will become 0.1 to infinity integral and the CDF that is λ exponential of minus λ x will with the λ is equal to 25 this comes out to be here after solving is 0.082 and in case if you want to find out the same probability using the concept of CDF, then the probability that x is greater than 0.1 can be written as 1 minus f(0.1) and just by substituting the value of this 0.1 in the CDF, here like this you get here the same value 0.082 and you can see here both the probabilities match. So, this is up to you whatever rule you really want to follow to find out different types of probabilities.

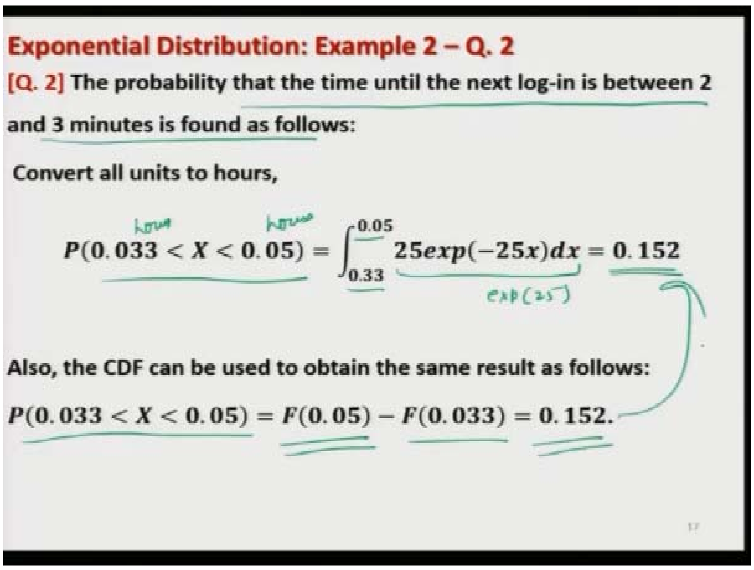
Now, in case if you try to understand what this probability is really indicating on the curve of this exponential distribution. So, you can see here this is the curve of exponential here 25, then this is the value somewhere here for example, 0.1 so we are trying to find out that X is greater than 0.1, so this area, this in the green color that is indicating the area or the probability that we have found here.

(Refer Slide Time: 21:45)




Now, in case if you want to find out the same probability using the R software that can be obtained by the command 1 minus pexp, it is from this probability for 1 minus f(0.1) and once you try to solve it, now this is very simple for you, this will come out to be 0.082085 and equivalently if you do not want to use this 1 minus CDF concept then you can simply use here the option lower dot tail is equal to FALSE and the probability can be computed once again using the pexp function and this will come out to be the same.

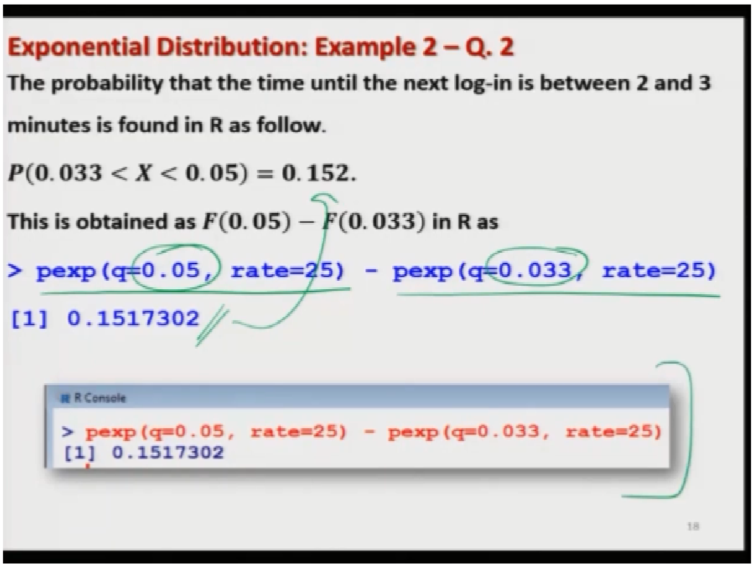
(Refer Slide Time: 22:23)



Now, I try to take here the question number two in the same setup and we want to find out here the probability that the time until the next login is between 2 and 3 minutes. So, for that that means we are simply trying to find out that probability that X is lying between 2 and 3 but then 2 and 3 minutes we have to convert it in the terms of hours. So, we try to write down here that X is lying between 2 to 3 minutes or X lies between 0.033 to 0.05 hours and if you try to find out this probability that will be the integral between these two points into the PDF of exponential.

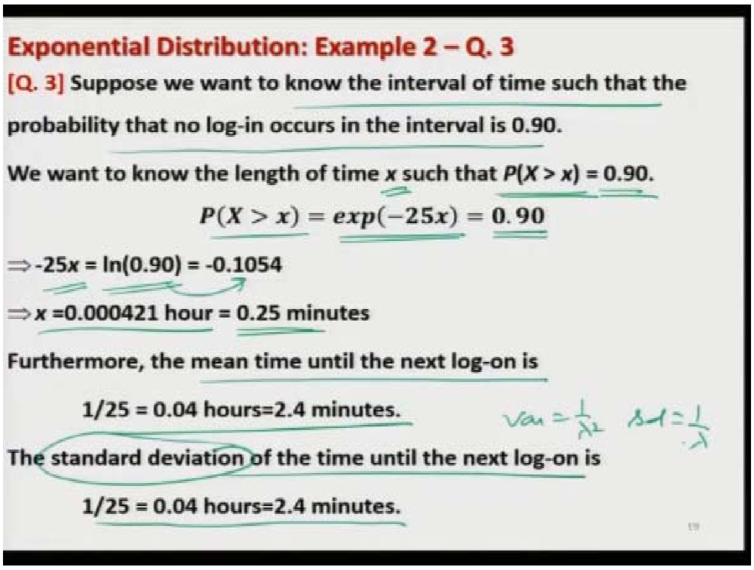
So, this here λ is equal to 25 and this integral can be computed to be 0.152, computing integral is very simple for you and similarly if you want to compute the CDF, to compute the same probability we can write down here the probability that X is lying between 0.033 and 0.05 that will be simply here F at 0.05 minus F at 0.033. So, this is the difference of the value of the CDF at 0.05 and 0.033 and this will come out to be the same as 0.152.

(Refer Slide Time: 23:55)




And now let us try to do the same example in the R software also, so you know that this probability the difference of CDF can be expressed as here is the value of CDF at 0.05 minus the value of CDF as 0.033 and if you try to compute it here this will come out to be 0.1517302 which is very close to 0.152 and this is here the screenshot of the same operation.

(Refer Slide Time: 24:26)



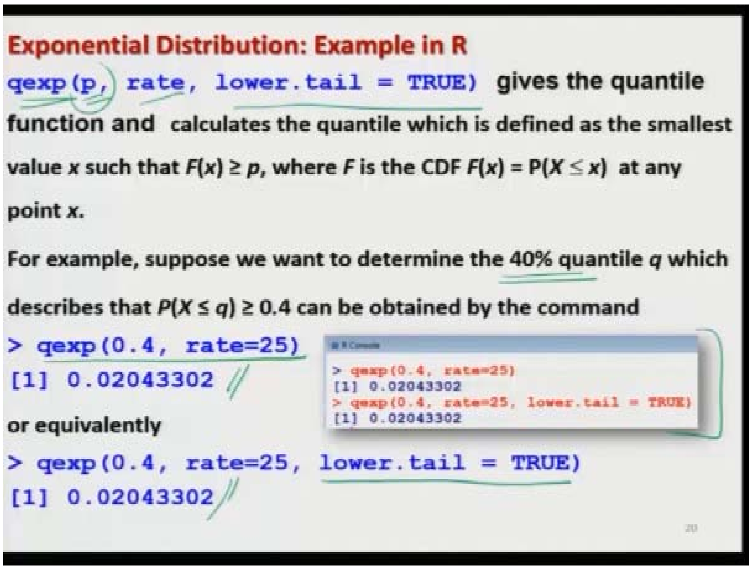

Now, let us try to consider one more example on the same data set. The question number three suppose, we want to know the interval of time such that the probability that no login occur in the interval is 0.90. So, basically, we want to know the length of the time that is the value of the random variable X such that probability that X greater than x is equal to 0.90.

Now, that is a very simple thing, you simply have to compute the probability X greater than x and put it equal to 0.90 and solve the equation. So, this probability that X greater than x is exponential of minus 25 of x equal to 0.90 and you simply have to just take the natural log on both the sides to solve this equation.

So, this will become here minus 25 x is equal to natural log of 0.90 whose value is minus 0.1054 that you can find from the table or from the R software whatever you want and if you simply try to solve it x will come out to be this value which is equal into 0.25 minutes.

So, the average time, this is the value of here X, the average time until the next log on is 1/25 that is close to 2.4 minutes and the standard deviation of the time until the next log on that will be the same as $1/\lambda$ which is 0.04 hours or 2.4 minutes, remember one thing I am asking here for the standard deviation which is the variance here is $1/\lambda^2$, so standard deviation is $1/\lambda$, so do not get confused.

(Refer Slide Time: 26:22)

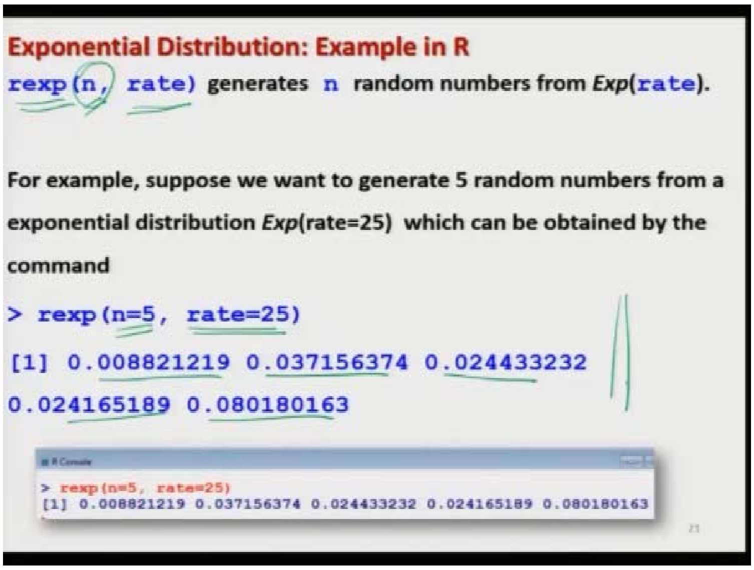



So, and if I try to do the same thing in the R software that you can solve very easily, there is nothing complicated in this thing. So, now next after considering this example I would like to show you that how can you compute the given quantile of this exponential distribution.

So, suppose, you want to determine the 40 percent quantile of this exponential distribution, so for this we have a command here qexp where you have to specify here the value of p in terms of the quantile and then you have to give here the value of rate and lower dot tail is equal to TRUE and so I can write down here qexp(0.48), rate is equal to 25 and you can get here this value.

Even if you try to use here the option lower dot tail is equal to true you will get the same value because this is the default value and this is here the screenshot. So, you can see that that computing quantiles is not different even when the observations are assumed to follow an exponential distribution.

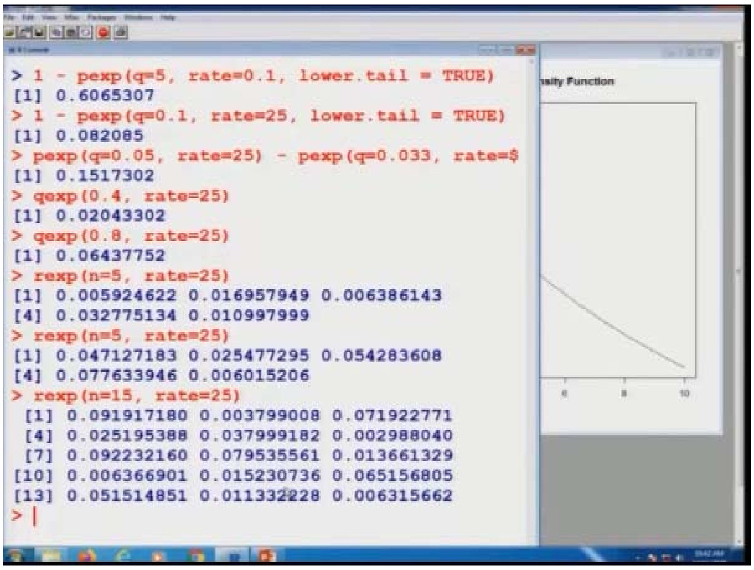
(Refer Slide Time: 27:33)




Now, similarly if you want to generate the random numbers from this distribution, you know the command here is rexp and then number of observation that you want to generate is given by n and then here the value of λ. So, now suppose, if you want to generate 5 observations, so I give here n equal to 5 from the exponential distribution with λ is equal to 25, this is here rate equal to 25 and you can see here 1, 2, 3, 4, 5 these are the 5 random numbers that are going to be generated.

Sure, now I would like to show you all these results on the R console, so I will be showing the same result which I have prescribed or what I have written on the slides except this one because you know that these are the random numbers and if you try to generate the random numbers you are not going to get the same thing.

(Refer Slide Time: 28: 25)

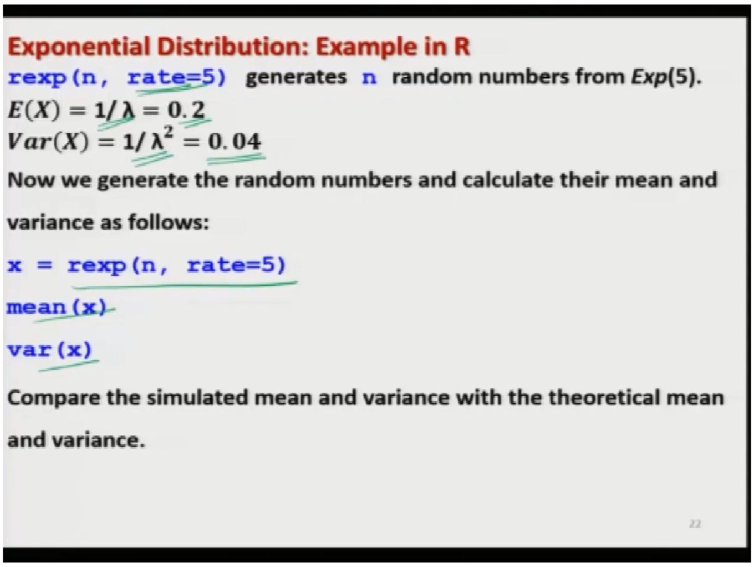



So, let us come back to our means earlier slides and try to just get confidence that these things can be found. So, if you try to consider the first example where we were trying to find out the value of 1 minus F(5), so that can be obtained here like this, let me try to clear the screen, like here you can see here 0.6065, that is the same thing and similarly if you want to consider the example two where you are trying to find out different types of values, so you can see here this value is coming out to be 0.08 and if you try to verify it on the R console this will coming out to be the same value.

And similarly, if you want to compute this probability that X is lying between 0.033 and 0.05 this also comes out to be the same what we have reported in the slides 0.15 and then if you want to compute the $40^{th}$ quantile with the rate equal to 25 then you can just generate it here like this, this command, and if you want to suppose here find out the $80^{th}$ quantile, so I can write down here instead of 0.4, 0.8 and this will come out to be 0.06, that is not difficult at all.

And if you want to generate here the random numbers, suppose I want to generate the random numbers from exponential distribution with λ is equal to 25 these are the 5 numbers which can be generated and if you try to repeat it you can see here that these numbers are not going to be repeated because they are random and similarly if you want to increase the number of observation that you want to generate, you can simply write down here n equal to 15 to generate 15 observations you can see here they are like this. So, you can see here it is not a very difficult thing to operate the exponential distribution on the R software and you can compute all sorts of probabilities.

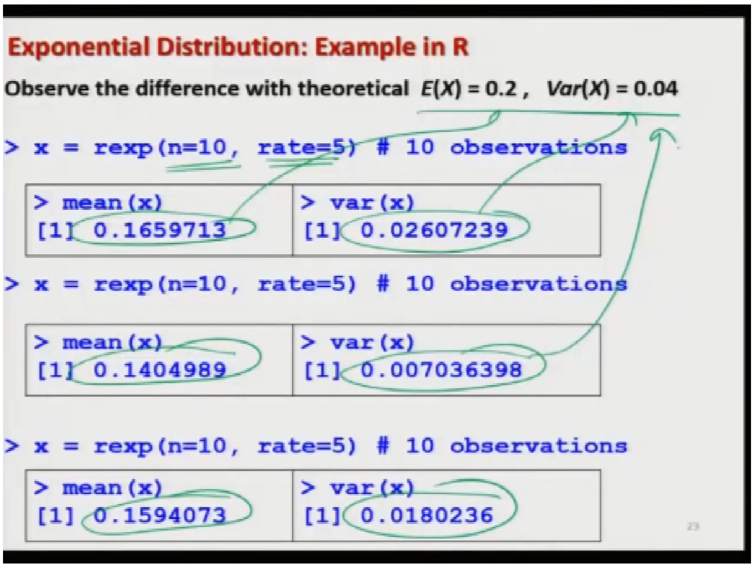
(Refer Slide Time: 30:34)




So, now let me try to give an idea that how the theoretical mean and theoretical variance are going to behave on the basis of sample size or the number of observation. So, what we try to do here that we suppose I try to consider here an exponential distribution with rate is equal to 5, so the mean value comes out to be here $1/\lambda$ 0.2 and variance will come out to be $1/\lambda^2$ which is 0.04, well you can ask me why I have not considered here $\lambda$ is equal to 25 because 1/25 was becoming a very small quantity, so it was difficult to show you here, but if you wish you can do it without any problem.

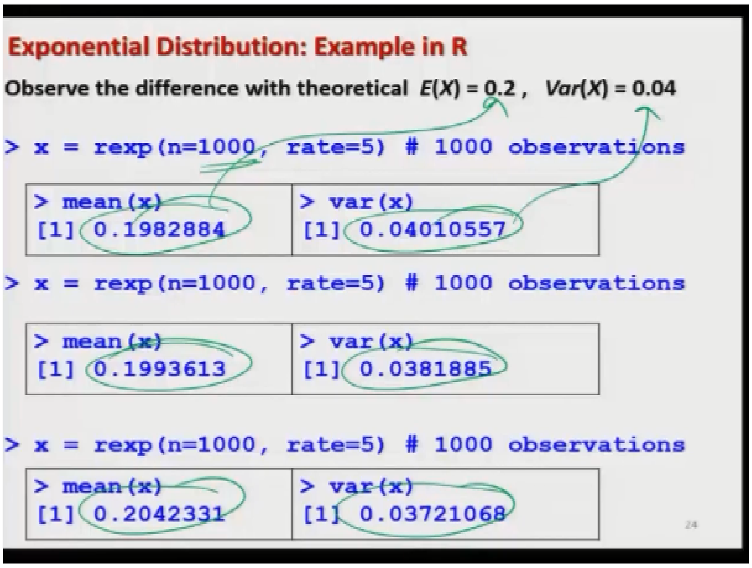
So, what we try to do here, that we try to generate here a number of observations using the command rexp and then we try to find out its mean and variance. The mean and variance of the observation that are generated and we try to see how they are getting close to the value of theoretical mean and theoretical variance.

(Refer Slide Time: 31:37)



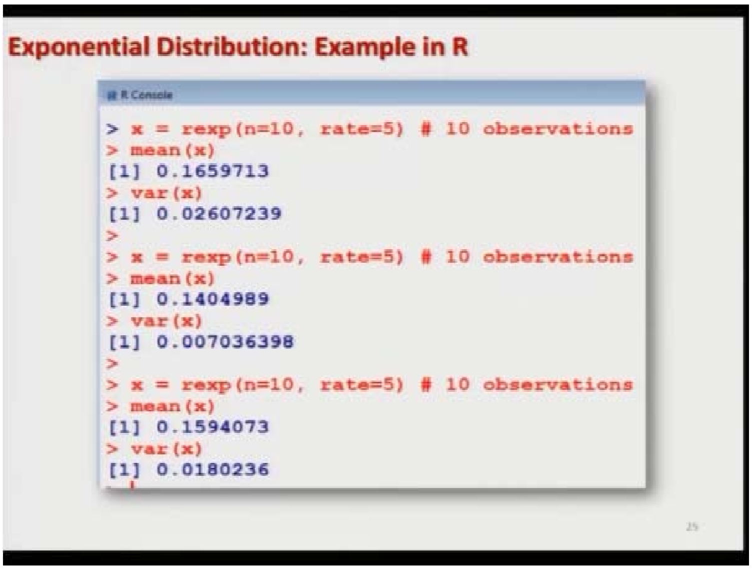

So that theoretical mean and theoretical variance here are 0.2 and 0.04 and first we try to generate here only it can observations with λ is equal to 5 and you can see here that the value of mean is coming out to be 0.16 and variance is coming out to be 0.02 and you can compare them with these two values which are the theoretical values and similarly, if you try to repeat the experiment in the next sample the value of mean is coming out to be 0.14 then 0.15 and you can see that the true value is 0.2, so they are quite away and the value of variance is coming out to be 0.007 and 0.01 which are also quite away from the true value that is 0.04.

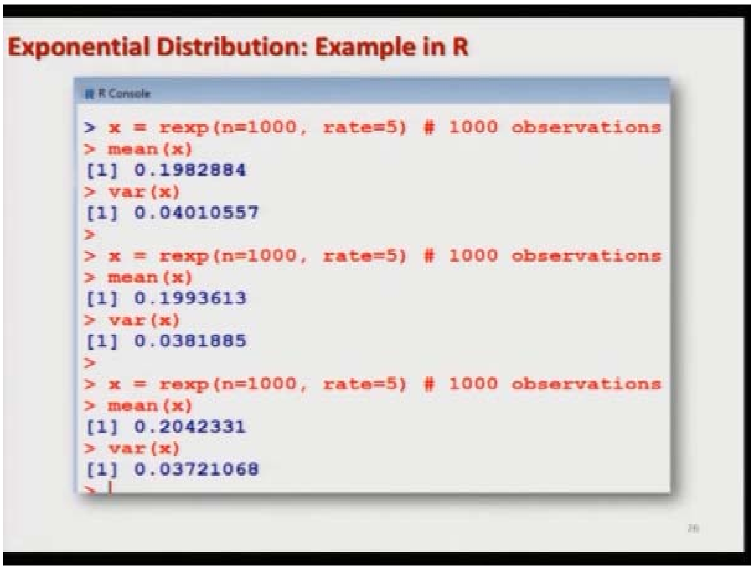
(Refer Slide Time: 32:27)




Now, in case if you try to increase the sample size and suppose you draw 1000 observation, then in this case the mean is coming out to be something like 0.19, 0.19 and 0.20, so they are very close to the true value that you can see and the value of the variance is coming out to be 0.04, 0.038 and 0.037 they are also very close to 0.04. So, you can see here that depending on the number of observations these values are getting closer to the theoretical mean and theoretical variance.

(Refer Slide Time: 33:01)






And these are the screenshots of the simulated value that I just shown you. So, now we come to an end to this lecture and with this lecture I will also stop discussing the different type of probability density function. Well as I said earlier there is a long list of probability mass function, probability density function and I am not saying at all that if I am not considering them here, they are not important.

Distributions like hypergeometric distribution, negative binomial distribution, and gamma distribution, beta distribution etc. that is a long list but surely if I try to consider all of them here you will agree with me that we will have no time to do all other important topics, so that is the reason I am not considering here and there is no question that they are not important but definitely I will request you now you have done sufficient number of probability mass function, probability density function and if you try to take up any book and try to look into this probability mass and the probability density functions I am promising you, you will not have any problem that is my promise.

Now, if you try to see what you have to do, you simply have to look into the function, you have to understand under what type of condition it is going to be helpful, what are the parameters, how to know the value of parameters and then how to compute different types of probabilities, how to generate the random numbers, how to compute the quantiles, how to compute the CDF etcetera.

And in all the probability mass functions and probability density function I have not given you the algebra, because if I start giving you algebra that will consume lot of time but on the other hand it does not mean that it is difficult, that is not difficult at all once again I am promising you, the only thing is this you have to pick up any book and then you have to just follow it and try to see that how do those expressions have been obtained, yes, surely when we are trying to do statistics there are two options whether you want to learn only the application or you want to learn the background also.

So, the best option is that if you learn the background also that will strengthen your basic fundamentals, whether those topics are going to be asked in the examination or not that does not make any difference because the examination will finish after a couple of years in your life but your professional life will start and there nobody will ask what were your marks in that topic or in that subject. Your knowledge which you are gaining that is going to play an important role, that is going to help you in your entire life.

So, more important part is to gain and acquire knowledge and for that you need to look into books and definitely the knowledge that you get from the book that cannot be substituted with anything even this course, unless and until you study from the book you read from the book your level of knowledge will not expand, I am giving you here in the limited time type of training that how you have to study, how you have to think, after that you have to take it further yourself. So, you try to practice, try to take a book, try to read it, try to understand it and I am sure that you will not have any problem and I will see you in the next lecture with more topics till then good bye.