Essentials of Data Science with R Software - 1

Professor Shalabh
Department of Mathematics and Statistics
Indian Institute of Technology Kanpur

Lecture No. 23

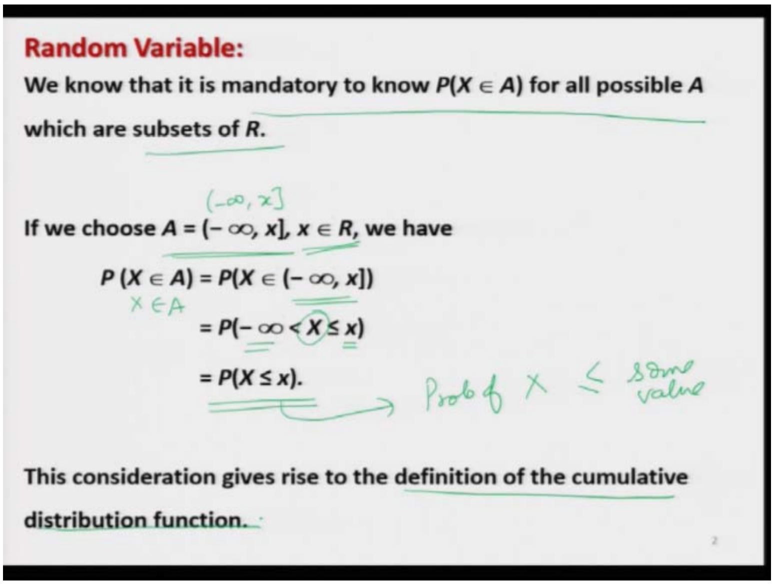
Cumulative Distribution and Probability Density Functions

Hello, friends. Welcome to the course Essential of Data Science with R Software 1, in which we are trying to understand the topics of sampling theory and statistical inference. So, now, you can recall that in the last lecture, we initiated discussion on the random variables and we had considered the concept of random variable and we have tried to associate it with some numerical values and the corresponding probabilities. And we also discussed that that the concepts related to the discrete and continuous random variable.

So, now, in this lecture I will be moving forward and I will try to discuss different types of probabilities, different types of concepts which are related to continuous and discrete random variables. One thing I would like to inform you here beforehand, that the concept for the discrete as well as continuous random variables, they are the same. But the way they are expressed mathematically that is different.

So, I will try to give you the definitions of the respective concepts in both the cases continuous and discrete, but definitely I can do it one by one only. So, in this lecture today, I am going to discuss about the continuous variable and in the next lecture, I will try to discuss about the same concept, but in a discrete variables set up. So, you have to be just careful because this lecture will be one by one. So, just in order to make it clear, I am trying to do in this particular way. So, let us begin our lecture and we try to see how we can compute different types of probabilities.

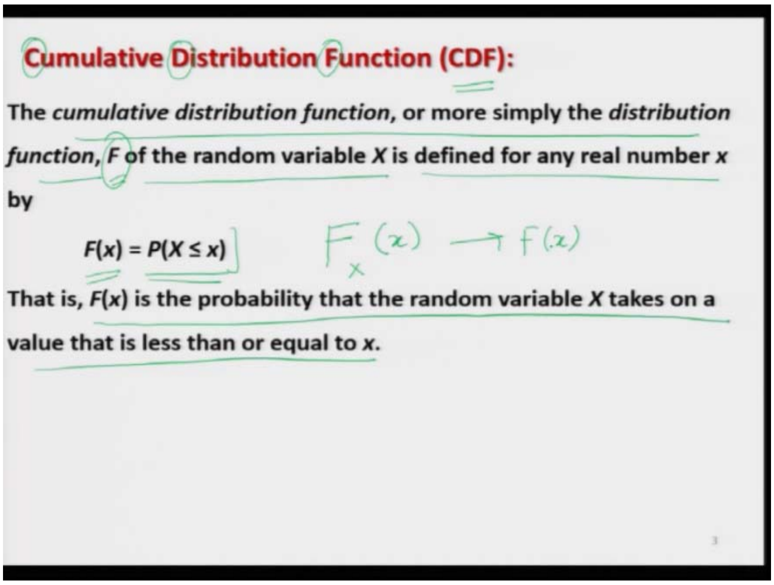
(Refer Slide Time: 02:08)



So, now, we had discussed in the last slide of the last lecture, that it is mandatory for us to know the $P(X \in A)$ for all possible A which are the subsets of real numbers R. And we had also discussed that in case if I try to take this we have

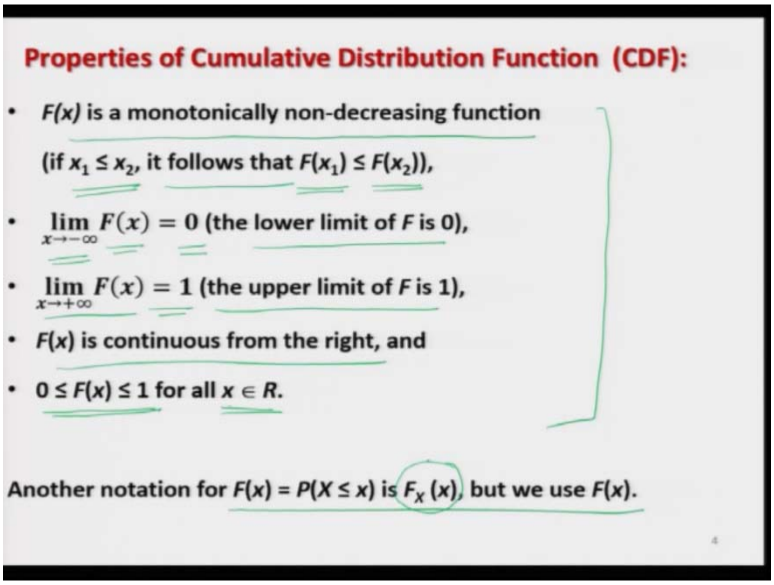
in the form of an interval like $A = (-\infty, x]$, where $x \in R$,, then we can write that the probability that random variable $X \in A$, $P(X \in A) = P(X \in (-\infty, x])$ $= P(-\infty < X \leq x)$ $= P(X \leq x)$. So, this probability this type of probability if you try to see, they are trying to indicate probability of X less than or equal to some value, like this. So, this type of probability is helping us in giving us a definition of cumulative distribution function.

(Refer Slide Time: 03:21)



So, now, let us try to formally define the cumulative distribution function which is briefly called as CDF, C is coming from cumulative, D is coming from distribution and F is coming from function. So, the cumulative distribution function or more simply the distribution function capital F of the random variable capital X is defined for any real number small x by $F(x) = P(X \leq x)$. So, capital F is the standard notation for indicating the cumulative distribution function. So, this thing can be written as capital Fx is the probability that the random variable capital X takes on a value that is less than or equal to small x. And sometimes you will see that it is also written as a capital F and the random variable comes here in the subscript and then inside the parentheses, we write the value small x. But just for the sake of simplicity, we will just consider here as say here Fx, just for the sake of simplicity, nothing more than that.

(Refer Slide Time: 04:28)



Now, this cumulative distribution function or the CDF has certain mathematical properties. And one thing which will come to your mind is that how this mathematical properties comes into existence. You see definitely you are going to compute something; you are going to compute something using some mathematical theories, mathematical functions. How to compute a probability will not be coming from the sky, but we need to develop it and for that, we take the help of some mathematical theories, mathematical concept, mathematical functions.

So, when we are trying to use those theories, concepts and function, there is some requirement. There are some conditions for the validity of those functions. So, these conditions number 1 come from there. Number 2, whenever you are trying to characterise a phenomena. So, that phenomena or process is beyond our control and we are trying to simply copy that phenomena through a mathematical function. So, the mathematical function has to be defined in such a way and certain conditions should be imposed on this such that it is resembling with the real phenomena. So, that is how these conditions will come and then you will find that in the theory of statistics we try to state many types of condition. People many times do not bother about those conditions, but unless and until those conditions are satisfied, you will not get the correct answer.

For example, if somebody has got a some health problem, unless and until the person goes to the doctor and explains the problem in the correct way, the person cannot be given a proper medicine. And that is the basic assumption that if somebody has met with an accident and fractured the bone, then the person has to go to an orthopaedic. And if somebody is facing some skin problem, he has to go to a skin specialist. So, that is our assumption.

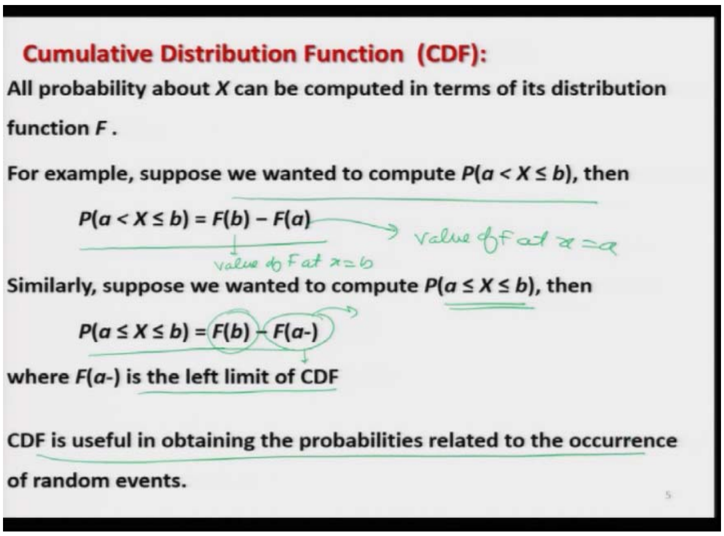
So, and we say that somebody has got a skin infection and the person says okay, I have gone to the doctor. So, it is expected and that we assume that the person has gone to a skin specialist and not to an orthopaedic surgeon. So, this is the meaning and validity or the interpretation of these type of assumptions.

So, we have here 4 possible such assumptions. First assumption is the CDF is a monotonically non decreasing function. This means what? That if I have two values $x1 \leq x2$, and if a small x1 is less than or equal to x2, it follows that CDF at x1 will also be less than or equal to CDF at x2. Second condition is l $\lim_{x \to -\infty} F(x) = 0$ That is the lower limit of F is 0.

And third condition is $\lim_{x \to +\infty} F(x) = 1$. This means the upper limit of F is 1 and Fx is continuous from the right. You know the concept of continuity from left hand continuity from right. So, and the value $0 \leq F(x) \leq 1$ for all $x \in R$, 0 and 1 are inclusive and x belongs to a set of real numbers. So, now, you will see that I will be using these properties again and again and you will see that these properties are convincing from the application point of view that whenever you are trying to characterise a phenomenon these properties will be satisfied.

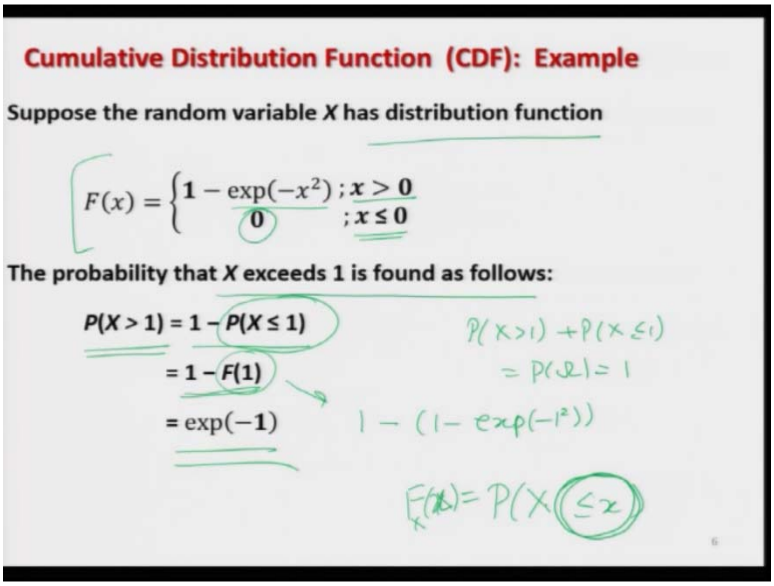
So, as I told you that that another representation of this CDF is like F in the subscript you try to write down the random variable and, in the parenthesis, you write try to write down the value, but we will use here only capital Fx just for the sake of simplicity, nothing more than that.

(Refer Slide Time: 08:18)



Now, what is the use of this CDF? All probabilities about X can be computed in terms of CDF. Once you know the distribution function Fx, you can compute all sorts of probabilities of happening or non-happening of the event. For example, suppose we want to compute the $P(a < X \leq b)$ so X is greater than a and less than or equal to b. Then in this case, this probability is going to be $F(b) - F(a)$, that means, you simply try to find out the value of F at x equal to b, and you try to find out the value of F at x equal to a and just take the difference. And that will give you the probability that capital X is lying between a and b. And similarly, if you want to compute the $P(a \leq X \leq b)$, then event this can be written as $P(a \leq X \leq b)$ will be equal to $F(b) - F(a-)$. So, b is the value of the F at x equal to b and Fa is the left limit of the CDF. So, this Fa minus is indicating the left limit of the cumulative distribution function f. So, CDF is useful in obtaining the probabilities related to occurrence of the random events. And you can see here that they can be obtained by using the concept of capital X, the random variable.

(Refer Slide Time: 09:51)



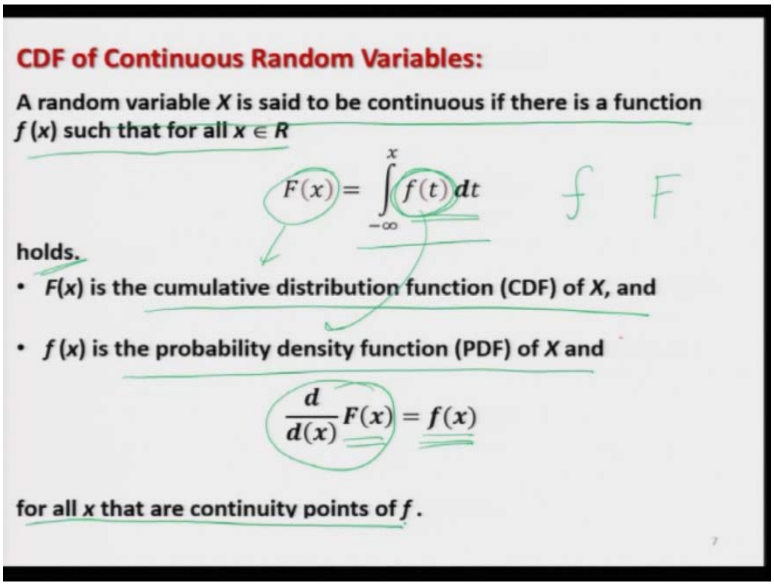
So, let me try to take here one example to explain you what I am trying to say. Suppose there is a random variable whose CDF or the distribution function is known to us. And it is given by here like this, Fx is equal to one minus exponential of minus x square, when x is greater than 0. And if x is less than equal to 0, then this capital Fx takes value 0. So, the probability that capital X exceeds 1, it has to be found. And we can write this probability probability capital X greater than 1, and that is going to be the same as 1 minus probability that X less than or equal to 1.

Because you know that $P(X > 1) = 1 - P(X \leq 1)$, this is same as probability of Ω, which is equal to here 1. So now if you try to see here, what is this value probability X less than equal to 1, this is F1. So, this is going to be here, $1 - F(1)$, and which you can write here, how?

Now how to find those F1? This will simply be be here, 1 minus 1 minus exponential of here, minus 1 squared, that is all and this will come out to be here, exponential of minus 1.
Now, here comes one more question, that here when you are trying to use the concept of CDF to find out the probabilities, it is always giving you the probabilities in terms of probability X is less than or equal to some value. But cannot you find this probability directly or do not you think that this function capital Fx as the name suggests, this has cumulative distribution function, that means it is trying to cumulate all the probabilities of the events up to the value small x.

And then it is trying to give you the probability, but if you want to find out the probability of of any particular event and you do not want up to certain event, then, what is to be done? So, for that, we can similarly define one more function which is called here as a probability density function.

(Refer Slide Time: 11:58)



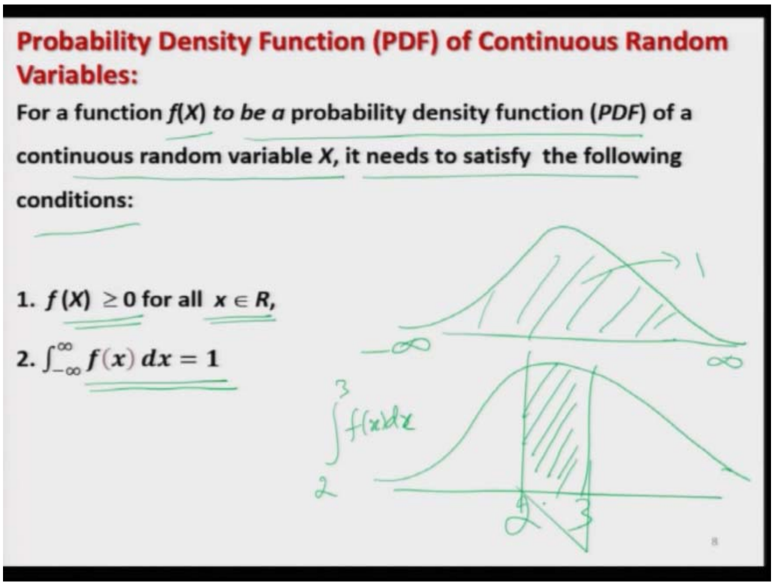
And in order to define that thing, first we try to define that under what type of conditions this random variable exists to be continuous. Although up to now, what we have understood this was a very general definition. Now we are coming to a mathematical definition. So, a random variable capital X is said to be continuous, if there is a function $F(x) = \int_{-\infty}^{x} f(t)\,dt$. You can observe here, now I have introduced here a new function a small f.

Earlier we had capital F. This integral holds and this capital Fx here is the CDF of capital X and this this is small f, this is called as probability density function of X. And the

$\frac{d}{d(x)}F(x) = f(x)$, for all x that are continuity point of f, means I now believe that you have

this much of mathematics knowledge, that what is continuity, what are the continuity point etcetera.

(Refer Slide Time: 13:25)

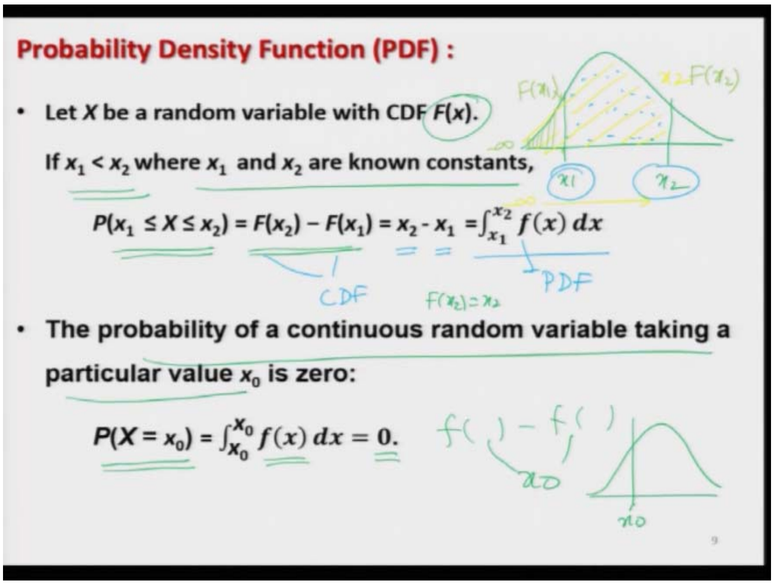


So, now I can formally define this probability density function for a continuous random variables. So, for a function small f to be a probability density function of a continuous random variable capital X, it needs to satisfy the following conditions, that $f(X) \geq 0$ for all $x \in R$ and integral over the entire sample space. So, I am assuming here my sample space is between minus infinity to plus infinity. So, $\int_{-\infty}^{\infty} f(x)\,dx = 1$.

So, now, one thing you have to keep in mind, what is the meaning of integral? Do not you think that you have been taught that integrals are indicating the area under the curve? So, what are you trying to say here, if you try to prepare here a curve or draw here a curve, where here it is something like minus infinity to infinity, you are trying to integrate and you are trying to say that the area under this curve is only 1.

And similarly, if I try to ask you here, suppose if I try to write down here 2 to 3 fx dx, what is this thing? In this curve, suppose somewhere here, I have here 2 and here I have here 3. So, it is trying to measure this area. Just keep in mind, this is going to be useful. Try to now correlate this integral with the area under the curve.

(Refer Slide Time: 14:53)



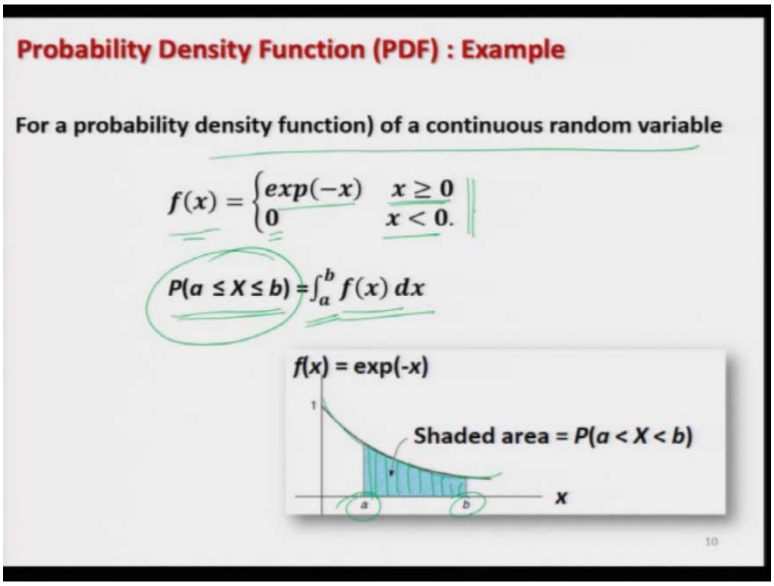
So, now I try to make here a connection between the PDF and CDF or a continuous variable here capital X, which has got a CDF capital Fx. So, if $x_1 < x_2$, x_1 and x_2, they are some known values. The probability that X is greater than or equal to x_1 and less than or equal to x_2, this is $F(x_2) – F(x_1)$. How? Means if you try to see here, what are these things? Suppose if I want to make here a small curve, and suppose this is point here, x_1 and this is point here x_2. Now I try to choose here two different types of colours. And I try to show you here.

So, this point here, which I am denoting here in yellow, well, I am taking here a light colour, so that there is no confusion. This yellow colour is here from minus infinity to this point, it is here x_2 suppose. And the area between minus infinity to here somewhere here x1 is here like this one. So, now, in case if you try to see this yellow area is going to give you here $F(x_2)$ and this point here, this is going to give you here $F(x_1)$. And if you try to subtract it here, what are you going to get here? You are getting here; this area and this area is simply here the integral of the curve between the point x_1 and x_2.

So, I can write down here $F(x_2) – F(x_1)$is equal to $x_2 – x_1$which is same as $\int_{x_1}^{x_2} f(x)\,dx$. And now you can see here this capital F, this is your your CDF and this is your here PDF. Because for the sake of simplicity and understanding, I am assuming here that means $F(x_2)$ is equal to here x_2. Yes, because for the sake of easiness to an easiness, to understand this concept.

Now, in case if I want to find out the probability of a continuous random variable taking a particular value say x0. So, now this will become here probability of X equal to say some numerical value x0 will become $\int_{x}^{x_0} f(x)\,dx$ so that will become here some value here mean something minus here something some value here and both the values are going to be dependent only on x0. So, this is going to be simply here 0. And if you try to see from the integration point, this x_0 is only here a point and the area of the point is 0. That is the mathematical concept.

(Refer Slide Time: 17:38)

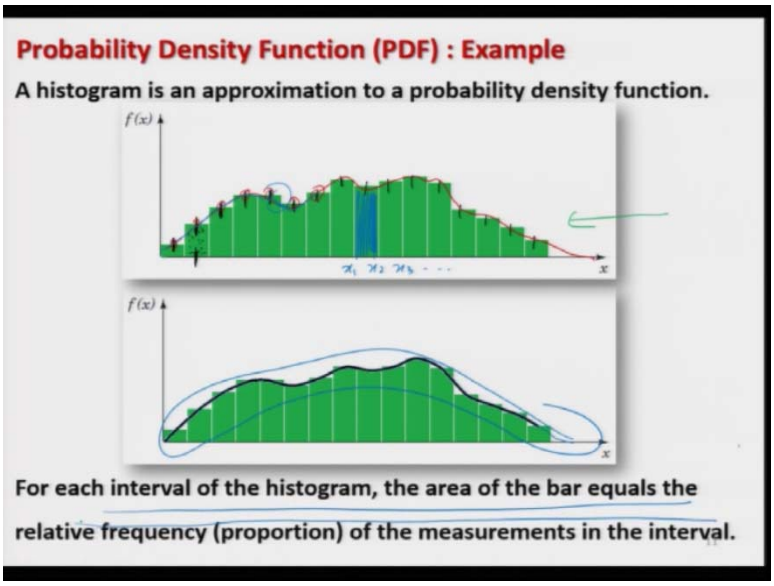


Now, can we try to take an example here and try to explain you the application of this concept. So, suppose we have a probability density function of a continuous random variable f of x, which is equal to exponential of minus x, if x is greater than or equal to 0 and 0 if x is less than 0. Well, this particular function has got a name. So, that I will try to discuss in the forthcoming lecture. But at this moment, you just try to pick this quantity here as a simple mathematical function, which is a probability density function, which is defining or which is at least satisfying the two characteristics that fx is greater than or equal to 0 for all x, and integral over minus infinity to infinity fx dx is equal to 1.

Now, we want to find out the probability that capital X is lying between a and b. So, you can see here if you try to plot this curve here fx this will look like this, as you have made it here. And if you try to take here two points here, a and b, this is the area in which you are interested. And this area under the curve can be found by simply integral over a small a to small b fx dx, that is the integral over PDF.

So, this shaded area is going to indicate the probability that X lies between a and b. So, you can see here that how convenient it is to find the probability once you know this type of function. Remember my words, once you know this type of function. So, if I can tell you this function, then finding out any type of probability will become very simple for you. So, later on, you will see we will try to introduce different types of probability density functions, which are trying to indicate different types of processes.

(Refer Slide Time: 19:30)



Now, one thing I would like to introduce you here that when you are trying to work in the data sciences, then you are dealing with the data. Observing our data on a continuous scale, it is practically not possible. That is a conceptual thing, actually, that is a concept. So, in practice, what will happen, you will have lots of data. It may be millions, billions and trillions of data. So, my objective here is just to give you an idea that, what is this probability density function and how it is obtained.

So, we can always create a histogram of all the values that we have obtained. This I believe that you know how to create the histogram that you have to make frequency intervals and then you have to obtain the frequency and those frequencies are plotted in the middle of the bars. So, suppose this is the histogram that you have obtained. Now, what you try to do here that you try to take the middle points of each of the bar. That you assume that the entire values in this interval of the bar they are concentrated in the middle of the bar.

For example, all the values which are here inside this bar, we are assuming that they are concentrated in the, at the middle of the bar. So, now, what do we have to do that, you have to just take put your pen on the histogram and you have to make a smooth curve, a smooth curve means you try to draw it here like this without putting your pen up, that means, once you start you have to move smoothly and continuously without giving a break. But it sounds very simple, but when you try to do it, many times you will stop it or something will happen so that you have to break this process.

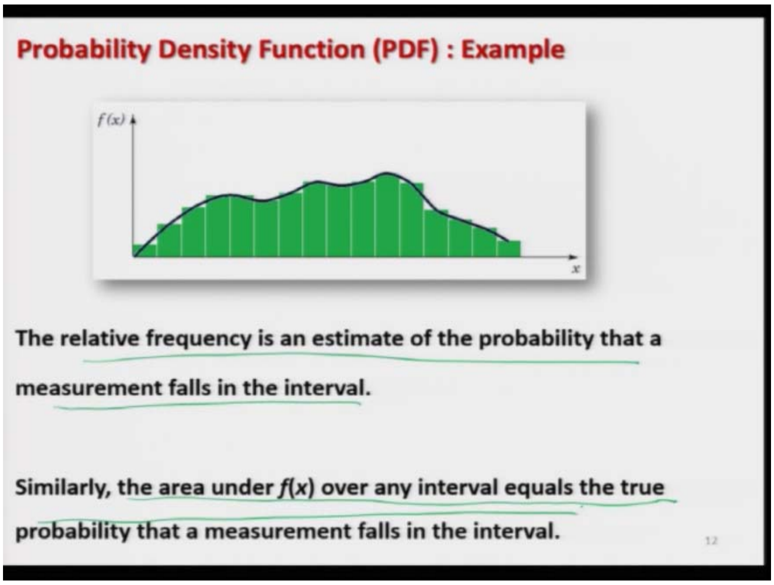
Well, that is a practical problem that we are unable to draw it but then you have been conceptually one can always draw it. And I can share you with a very small story that when I was a student, usually in the BSc part 1, we are given such type of exercise. And whenever I was trying to draw the curve there was always a break, means one and means my pen was getting stopped somewhere and then and then I was very careful in continuing exactly from the same point. But whenever I was going to my teacher the teacher was always able to find out that Shalabh, this is not a continuous curve, you have lifted the pen and I was always confused at how does he or she comes to know. But now after becoming the teacher, I can see that once you try to draw the curve without lifting the pen and after lifting the pen, there will be some change which you can observe. And this I will try to show you here. So, let us come back to our slide.

So, what I have to do here that now I will try to use the here a different colour pen say red, that I have to draw here a curve like this, which is passing through with the middle points of this bars. So, if I try to do it here, it will look like this, like this, like this. And you can see some where the points are not exactly passing to the point. So now what I try to do here, I try to use a different colour pen say blue. And if I try to make it here, and suppose I break up point here, now whatever I will do here, I will try to do here this point of break will be visible. And that that is what my professor was able to find. I was not intelligent enough at that time.

So, now the same thing I have done here, I have prepared here this curve here like this, you can see here, that is a smooth curve, which is passing through with most of the middle points. So, this curve is actually indeed indicating the probability density function. So, now I come to the moral of the story, whenever you are trying to create a histogram, this histogram is going to tell you that what is the relative frequency.

For example, if I try to say all these points, which are lying in this bar here, they are trying to give me the proportion of the points out of the total points which are lying in this bar. And these values are going to be something like x_1, x_2, x_3, etc. So, for each interval of the histogram, the area of the bar equals the relative frequency or the proportion of the measurements in that interval.

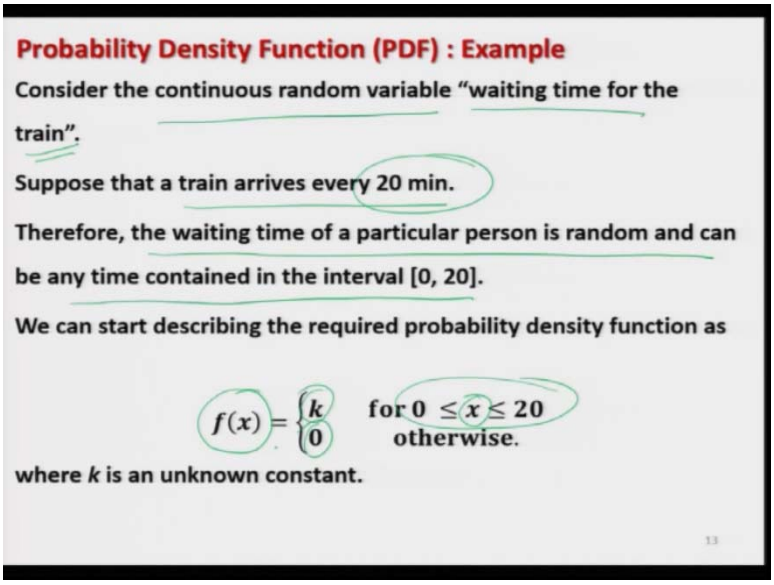
(Refer Slide Time: 24:18)



Now you can recall your definition of probability. So, can you now say that the relative frequency is an estimate of the probability that a measurement falls in the interval. So now, if you do not know any probability density function or any type of function, which is governing the probability, you simply have to observe the phenomena, you have to draw the histogram and from histogram, you can observe the phenomena you can compute different types of probabilities directly from there, and those are going to be the relative frequencies and now this relative frequencies are going to be taken as an estimate of the probability that measurement falls in that given interval.

So similarly, using the same concept, I can say here that the area under a PDF over any interval equal to the true probability that a measurement falls in the interval. So, you can see here how it is interesting and useful for us.

(Refer Slide Time: 25:12)

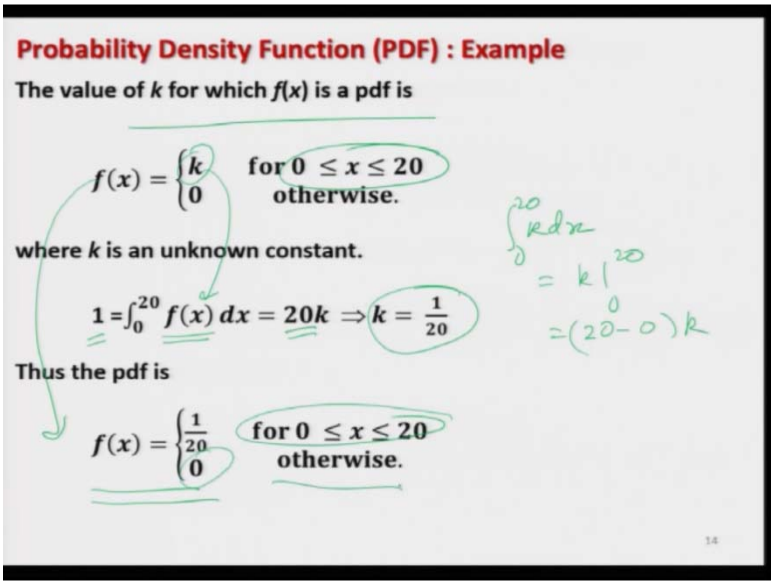


Now, I try to show you an example here. Suppose, there is a random variable, which is waiting time for the train, that means, suppose you reach to a railway station and then you are waiting for the train. Sometimes the train comes in time, sometime it is early or delayed and so on. But there is a timetable. Suppose, the train arrives every 20 minutes, you know. Now, the passenger reaches to the railway station at any given time. So, train will be arriving at a given time, but only after every twentieth minute. So, the person will have to wait.

Now, first thing is this you have to decide whether this type of waiting time will be a continuous random variable or a discrete random variable. So, obviously, this is going to be a continuous random variable. So, now, in case if you observe the phenomena that since the train is coming after every twentieth minute, so, at max the person will have to wait 20 minutes, that is all, even if the person has reached the platform when the train has just left or just living before the eyes, even then the person has to wait not beyond 20 minutes.

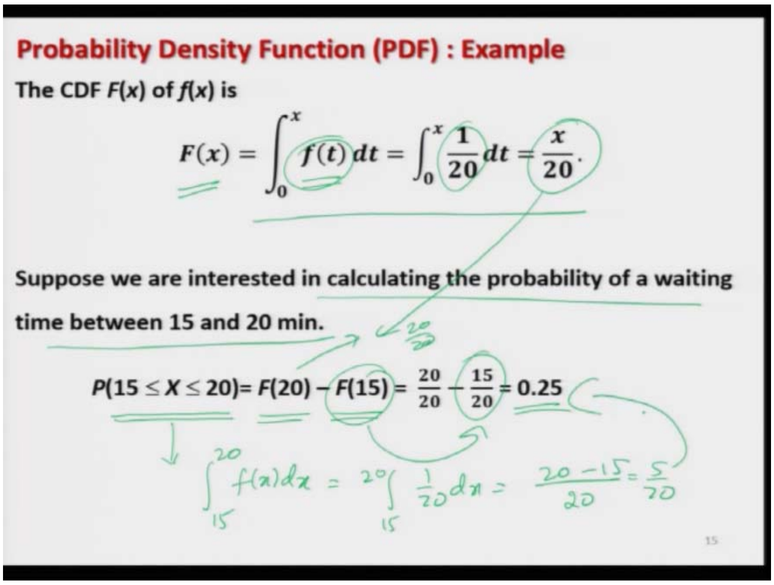
So, the waiting time for a particular person is random and can be any time which is contained in the interval, close interval 0 to 20. So, now, we have to define this process in the form of a mathematical function. So, now, the type of information which I have here that I have here a random variable X and which is defined by this small fx and x is going to take value between 0 and 20. And yeah, it will take value, suppose, it takes value here k some some unknown constant otherwise it takes value 0.

(Refer Slide Time: 27:05)



Now, how to find out this k? Now, I can use my properties of PDF and can find it out. So, first I tried to find out the value of k for which this fx is a probability density function. So, I can see here that the range of x is between 0 and 20, this function takes value 1. And beside these interval limits means all the values are 0. So, I can say here that the probability of an event over the range 0 to 20 will always be equal to 1. So, $1 = \int_0^{20} f(x)\,dx = 20k \Rightarrow k = \frac{1}{20}$. So, now, I can substitute this k and I can obtain this PDF as fx is equal to 1 upon 20 for x lying between 0 to 20 and 0 otherwise. So, you can see here that given some information you can also try to find out an appropriate probability density function, which is describing the phenomena.

(Refer Slide Time: 28:29)



Now, in case if you want to find out the cumulative distribution function from this PDF that is also easy you can write down here F(x) = $\int_0^x f(t)\,dt$. now, you have obtained as 1 by 20, where you can see here this is here now, 1 by 20. So, this you simply have to solve this integral and this will come out to be x upon 20.

Now, using this PDF or CDF, we can find out different types of probabilities. Suppose, we are interested in calculating the probability of a waiting time between 15 and 20 minutes. So, this X is going to lie between 15 and 20. So, this can be obtained using this CDF this will be F(20) – F(15). And the value of F20 is obtained from here that is 20/20 – F(15) which is obtained here 15/20. And this will come out to be here 0.25.

And the same phenomena can also be obtained by integral 15 to 20 fx dx. And we can see here this will come out to be here 20 integral here means $\int_0^x \frac{1}{20}dt = \frac{x}{20}$ that will be equal to here 20 minus 15 upon say 20. So, that will be here 5/20 which is equal to here 0.25. Same as like this. So, there are different ways here either using PDF or CD or whatever is convenient. We try to employ it to find out different types of probability.

So, now we come to an end to this lecture and I hope this lecture was interesting that now you can see we are moving towards more mathematical thing and but I am trying my best to to convince you that this mathematics is not a useless thing, but they are helping us a lot in computing different types of probabilities.

And we have discussed in the beginning itself that whenever you are trying to compute the probabilities, you are trying to describe a phenomenon. And if you try to understand the meaning of the probability distribution function, this probability distribution function or probability density function, they are trying to describe that how the entire probability is being distributed over the range of random variables.

For example, if you remember, we have done an example where we have rolled two dice and then we are trying to see that there are 36 possible points in the sample space and we are trying to observe the random variable as the sum of the numbers which are observed on the upper surface. So, now, those values will be between 2, 3, 4 up to say here 12. And we had computed the probabilities of all values. So, now, if you try to see this probability distributions are trying to help us in knowing the distribution of the probability over the entire range of the random variables. And that is why their name is probability distribution.

Now, since cumulative distribution function that is trying to give us the probability in a cumulative way, so, it is called as cumulative distribution function. The probability density function they are trying to give us the probability at in different intervals. So, they are called as probability density function as they are trying to indicate the densities of the probability over the entire range of that interval.

So, that is how you have to think. Now, I would request you to try to take some example from your book, from your assignment and try to practice them. Try to take different types of say simple function and try to see whether they are probability density function or not. If you want to know it, that means, you have to check two condition, whether the value of the function at every point is greater than 0 or equal to 0. And you have to see that we were there the integral over the entire space where the values of X are spreaded that is equal to 1.

This simple type of thing you try to practice it and I am promising you they will give you a very deep insight into the probability theory. So, you try to practice it and I will see you in the next lecture with more concepts on probability theory. Till then, goodbye.