PROBABILITY THEORY FOR DATA SCIENCE

Prof. Ishapathik Das

Department of Mathematics and Statistics

Indian Institute of Technology Tirupati

Week - 03

Lecture - 13

Continuous Random Variable and Probability Density Function

So now we will discuss continuous random variables. We have already completed the discrete random variable part and discussed some numerical examples. I hope you have understood those. You can go through some more numerical examples as well. Now we will discuss the next part, which is called continuous random variables.

A continuous random variable is defined differently than a discrete one. In the discrete case, you may remember that the distribution function is not continuous. For discrete random variables, the range of the random variable is either finite or countably infinite. In contrast, a continuous random variable does not take discrete values; instead, it takes intervals, resulting in uncountably many points. For example, consider throwing a dart at a target.

You might aim for the center, but due to imperfections, the dart may land in various places. The distance from the center can be measured along the real line, ranging from 0 to a maximum radius, r. Since this distance can be any real number within that interval, we classify this random variable as continuous rather than discrete. Now, how do we define a continuous random variable? Let X be a random variable with a cumulative distribution function $F_X(x)$.

If $F_X(x)$ is continuous and has a derivative that exists everywhere except at a finite number of points, and is piecewise continuous, then X is called a continuous random variable. To elaborate, this means that the cumulative distribution function $F_X(x)$ is defined for $x \in \mathbb{R}$. It is not only continuous; its derivative also exists and is continuous, except at a finite number of points. This is what we mean by piecewise continuous. So, what does "piecewise continuous" mean?

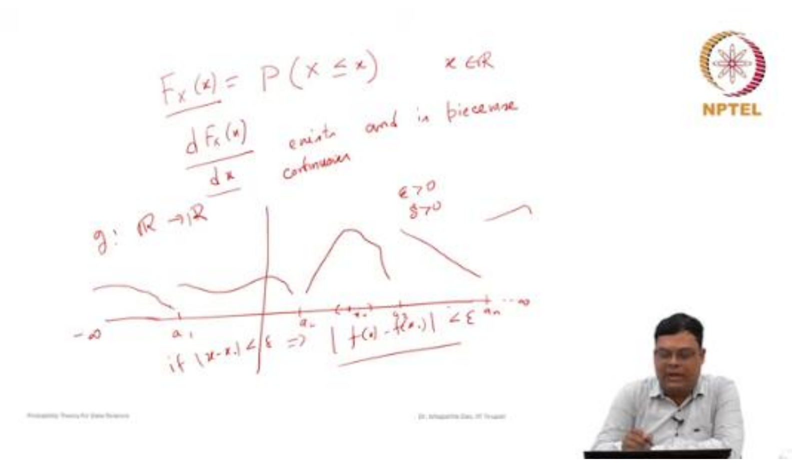
Graphically, a continuous function appears unbroken. There is a specific definition involving epsilon and delta. For any $\varepsilon > 0$, there exists a $\delta > 0$ such that the distance between $F_X(x)$ and $F_X(x_0)$ remains within the interval defined by $F_X(x_0) \pm \varepsilon$ if $|x - x_0| < \delta$. This describes continuity. However, piecewise continuous means that if we represent a function, say $g: \mathbb{R} \to \mathbb{R}$, the function may not be continuous over all real numbers.

Instead, we can break the entire interval into a finite number of intervals—say, from $-\infty$ to a, from a to a_1, from a_1 to a_2, and so on—where the function remains continuous within each of these intervals, but may be discontinuous at a finite number of points. Thus, if we find the distribution function is continuous, the derivative exists, and the derivative is piecewise continuous, then this random variable is classified as a continuous random variable.

The distribution function can be considered absolutely continuous if we go a bit further into detail. This means there is a function $f: \mathbb{R} \to \mathbb{R}$, such that the distribution function $F(x)$—the probability that $X \leq x$—can be written as:

$F_X(x) = \int$ from $-\infty$ to x $f(t)\,dt$

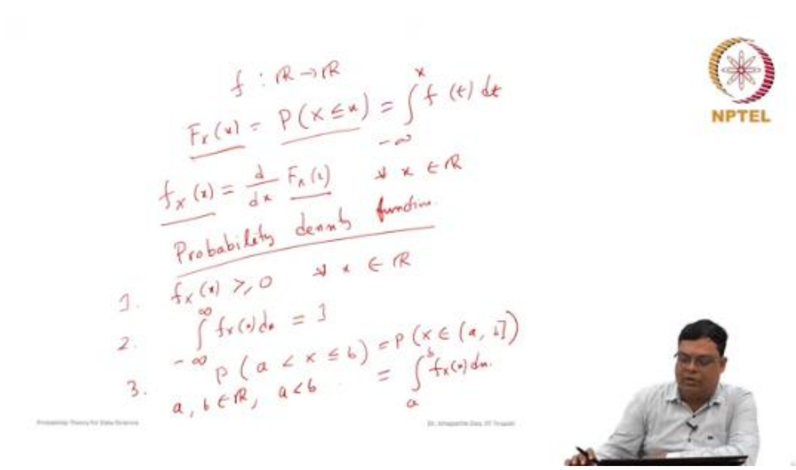
So, here $f(x)$ is the function representing the density.



This F(x) is called absolutely continuous with respect to Lebesgue measure, though we don’t need to go too far into this concept for now. For now, assume F(x) is a distribution function whose derivative exists and is piecewise continuous. Whenever this derivative exists, we define this derivative as $f(x) = dF(x)/dx$. This is simply the derivative of F(x), and since this derivative exists for a continuous random variable for all $x \in \mathbb{R}$, we call this the probability density function. The probability density function (pdf) in the continuous case can be thought of as similar to the probability mass function (pmf) in the discrete case.

In the discrete case, F(x) doesn’t have a derivative, so we define a pmf instead. But in the continuous case, since a derivative exists, we define the pdf instead. Now, consider the properties of the probability density function. Similar to the pmf, the pdf is always between 0 and 1, and equals 0 if the value does not lie within the range of the random variable. Also, like the pmf, the total over all possibilities equals 1, except in the continuous case, the “sum” is replaced by integration.

Specifically, we integrate f(x) from $-\infty$ to ∞, giving us a probability of 1 over the entire range. Lastly, for continuous random variables, we don’t define the probability at a specific point, but over an interval. For instance, the probability that X lies between a and b, where $a < b$, is found by integrating f(x) from a to b, which we denote as $P(a \leq X \leq b)$ and can write with open or closed intervals as appropriate. The graphical interpretation of the probability density function for a continuous random variable X is as follows: suppose your density function is f(x). If we draw f(x), an example will make this clearer, but let's first go over the basic properties.

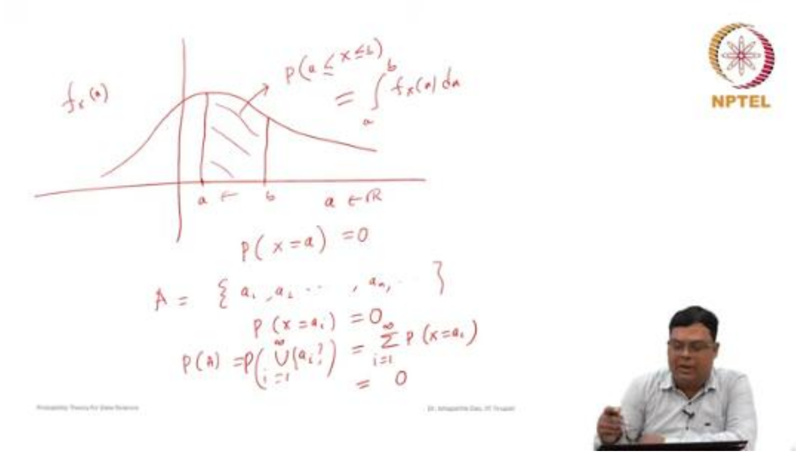


Suppose this is the density function, with values at points a and b. We want to find the probability that the random variable X takes values within this interval. This probability is simply the area under the curve, which corresponds to the integral from a to b of f(x) dx, giving the probability that X lies between a and b. For continuous random variables, it doesn't matter whether the interval is open, half-open, or closed; the probability remains the same under this curve. An important property in the continuous case is that the probability at a specific point (say, any real number like 2) is zero.

For example, the probability at point a, for any real number, is zero. This is different from discrete random variables where X takes discrete values with non-zero probabilities on those points. For continuous random variables, however, the probability at an exact point is zero. To understand why, think about $P(a \leq X \leq a)$ as the integral from a to a of f(x) dx. This is zero because it's over a single point, with no "width" to form an area. In two-dimensional space, the area under the curve for just one point is zero.

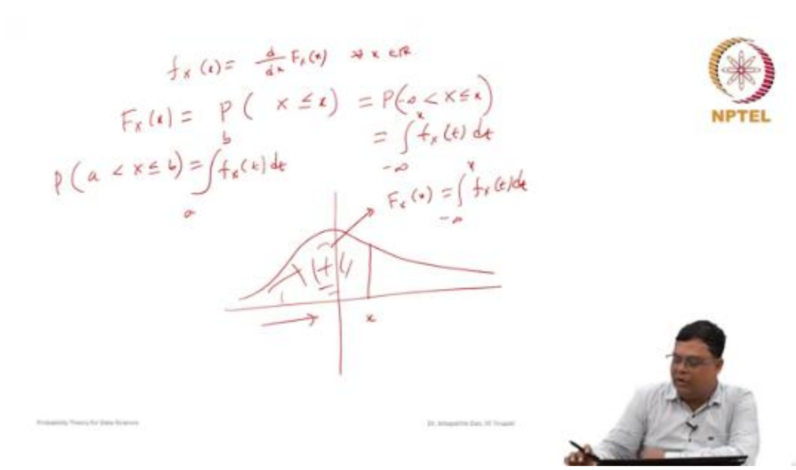
Although this isn't a formal proof, it gives an intuition of why the probability at any exact point is zero for continuous variables. Now, if someone asks if an impossible event always means a null set, the answer is no; a probability of zero doesn't always imply impossibility. For continuous random variables, a particular real number might occur, yet the probability at that exact point is still zero. Even for a countable set, the probability of each individual point remains zero. By Axiom 3, if we consider a countable set of points {a_i} with probability zero at each X = a_i, then the probability for the entire set is the sum of these zero probabilities, which converges to zero.

These are some of the key properties associated with continuous probability density functions. If you're given a cumulative distribution function, you can find the probability density function by taking its derivative. So, the probability density function is defined as the derivative, d/dx of F(x), for all $x \in \mathbb{R}$. Now, suppose you already know the probability density function. In most continuous random variable cases, it's convenient to work directly with the probability density function, and we'll usually discuss examples using it.

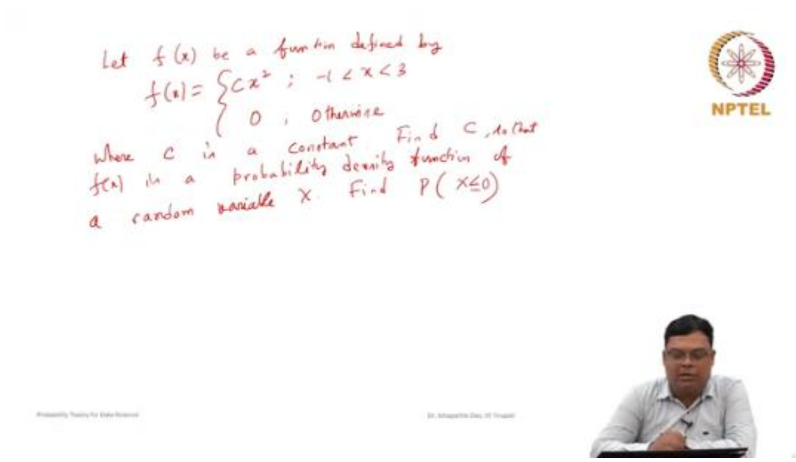


When the density function is given, you can find the cumulative distribution function. The cumulative distribution function (CDF) is the probability that $X \leq x$. So, to find $P(X \leq x)$, it's expressed as the integral from $-\infty$ to x of f(t) dt. Here, t is just a variable we're using to integrate up to x. In general, if you want the probability that X is between a and b, you would calculate it by integrating the density function from a to b. That is, $P(a \leq X \leq b) = \int[a \text{ to } b]\, f(t)\, dt$.

So, if you have the density function f(x), the cumulative distribution function F(x) can be obtained by integrating f(t) from $-\infty$ up to x. This integral gives the cumulative probability up to x, which defines the CDF. Let's go through an example of a continuous random variable or probability density function. Suppose we have a function f(x), defined as $c * x^2$ for $x \in [-1, 3]$, and f(x) = 0 otherwise. Here, c is a constant that we need to find to make f(x) a probability density function for the random variable X.



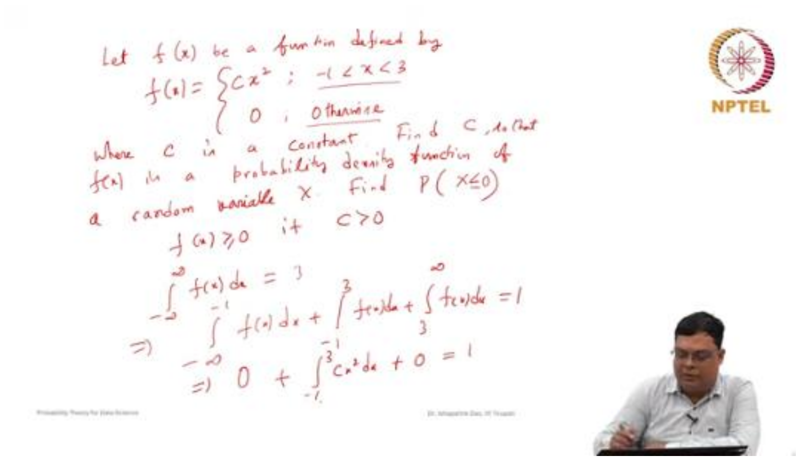
The first task is to find c so that f(x) is a valid probability density function. Once we find this value of c, we will also calculate the probability that $X \leq 0$. How do we proceed to find c, and then use it to determine this probability? Since we need to find the constant c so that f(x) is a probability density function, we will check whether it satisfies the conditions for a probability density function. These conditions include that $f(x) \geq 0$, and the integral from $-\infty$ to $+\infty$ of f(x) must equal 1.



Additionally, f(x) should be piecewise continuous, which is assumed for continuous random variables. To verify that all the properties are satisfied for any values of c, we first note that $f(x) \geq 0$ whenever $c > 0$, since x^2 is always non-negative. Next, we need to ensure that the integral of f(x) from $-\infty$ to $+\infty$ equals 1. We can see that f(x) is non-zero only in the interval from -1 to 3. Therefore, we can divide the integral into three parts: The integral from $-\infty$ to -1, which will equal 0 because $f(x) = 0$ in this interval.

The integral from -1 to 3, which will be $c \int_{-1}^{3} x^2\, dx$. The integral from 3 to ∞, which will also equal 0 since $f(x) = 0$ in this interval. This means we need to solve the following equation: $0 + \int_{-1}^{3} c\, x^2\, dx + 0 = 1$. Now we will focus on solving this integral to find the value of c.

Now we need to perform the integration, setting up the integral from -1 to 3 of $c\, x^2$. The integral of x^2 is $(x^3/3)$, giving us $c * (x^3/3)$ evaluated from -1 to 3, equal to 1. This implies $c * (1/3) [(3^3) - ((-1)^3)] = 1$, leading to $c * (1/3) [27 + 1] = 1$. Solving for c, we find $c = 3/28$. Therefore, our density function is $f(x) = (3/28)\, x^2$ for $-1 < x < 3$, and $f(x) = 0$ otherwise.



It's important to note that this function can be represented with an open interval or a half-closed interval, and the probability will remain unchanged since, for a continuous random variable, the probability at a single point is 0. Next, we find the probability that x ≤ 0, expressed as the integral from -∞ to 0 of f(x) dx. Since there is no bound to the left of 0, we can write this as the integral from -∞ to -1 of f(x) dx + the integral from -1 to 0 of f(x) dx. The first integral, from -∞ to -1, is 0 because f(x) = 0 in that interval. We focus on the second integral, from -1 to 0, which becomes $\int_{-1}^{0} (3/28)\, x^2\, dx$.

Performing the integration gives (3/28) * ($x^3/3$) evaluated from -1 to 0, simplifying to (3/28) * [0 - (-1)] = (3/28) * 1 = 3/28. Since we have a factor of 1/3 in the integration step, we conclude that the probability is 1/28. So, this is one example of a probability density function. We will discuss some more examples. Suppose we consider another example here.