# PROBABILITY THEORY FOR DATA SCIENCE

**Prof. Ishapathik Das**

**Department of Mathematics and Statistics**

**Indian Institute of Technology Tirupati**

**Week - 03**

**Lecture - 11**

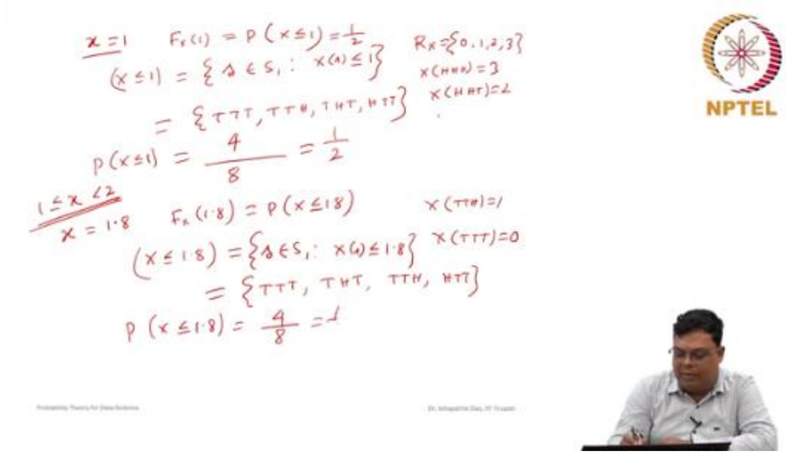
## Properties of the Cumulative Distribution Function and Discrete Random Variable

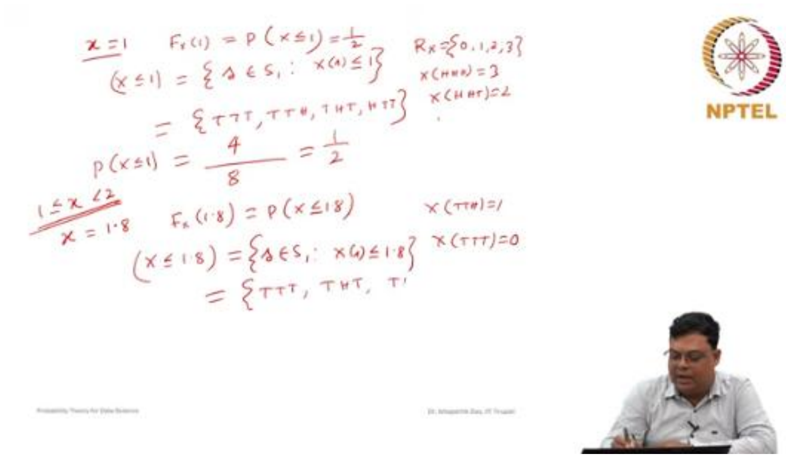
Now, suppose we take x = 1. Until now, we have been considering any x in specific intervals. When defining a function on the domain R, we must define the function for all values from $-\infty$ to $+\infty$. Here, you can see we first considered $x \in (-\infty, 0)$. This is because of a certain pattern in this function. For values between $-\infty$ and 0, this function is actually 0.

Whenever $x \in [0, 1)$ ($x \geq 0$ and $x < 1$), the value of this function remains constant at 1/8. Now, we will consider values of $x \geq 1$. Specifically, when x = 1, we want to find f(1) or F_X(1), which represents the probability that $X \leq 1$. To find this, we must first identify what it means for $X \leq 1$. This corresponds to all $s \in S_1$ such that $X(s) \leq 1$.

Remember, the range of this random variable X is {0, 1, 2, 3}, representing the number of heads obtained: for example, X(head, head, head) = 3, X(head, head, tail) = 2, X(tail, tail, head) = 1, and X(tail, tail, tail) = 0. Since we are looking at values less than or equal to 1, the random variable can only take the values 0 and 1. The value 0 corresponds to TTT, and the value 1 corresponds to cases like TTH, THT, and HTT. These four outcomes satisfy this condition. Out of 8 possible equally likely outcomes, 4 satisfy this condition, making the probability of $X \leq 1$ equal to 4/8, or 1/2.

If x is any value greater than 1, like 1.8 or 1.5, we would look for F_X(1.8), which represents the probability that $X \leq 1.8$. We then find all $s \in S_1$ satisfying the condition $X(s) \leq 1.8$. Since X can only take integer values {0, 1, 2, 3}, only 0 and 1 satisfy this condition. So, for values 0 and 1, TTT corresponds to 0, and TTH, THT, and HTT correspond to 1. So, these are the four values, which is why the probability of $X \leq 1.8$ is nothing but four out of eight, or 4/8, or 1/2.




This is constant in this interval, so F_X(x) = 1/2 whenever x ∈ [1, 2). That is what we found here. Similarly, you can show that whenever x ∈ [2, 3), and you take any value in this interval, say x = 2.3, then we want to find F_X(2.3). This is nothing but the probability that X ≤ 2.3. So, now we need to find what this set is. So, X ≤ 2.3 means this is all s ∈ $S_1$ such that X(s) ≤ 2.3.

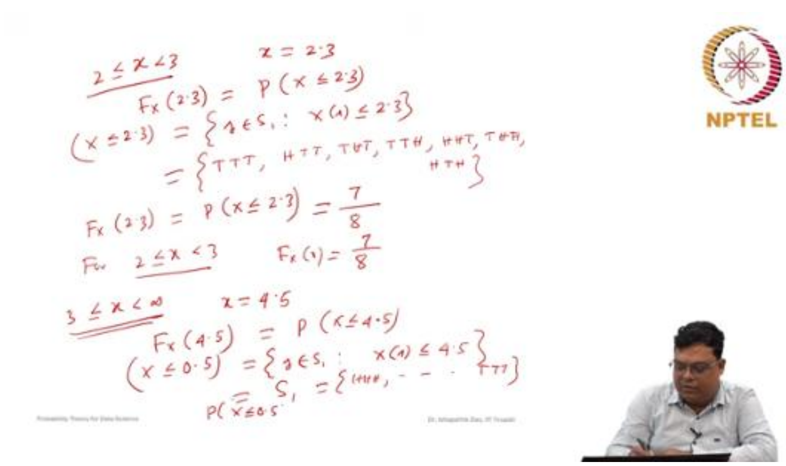
What values satisfy this? Since the random variable X takes values in {0, 1, 2, 3}, the values 0, 1, and 2 all belong to this set. So, 0 means no heads (TTT), 1 means one head (HTT, THT, or TTH), and 2 means two heads. Basically, with each additional observation, we are adding to this set, which is why it is called the cumulative distribution function. So, cumulatively, we are adding sequentially.

Now, we have to consider the two heads as well because it satisfies that X = 2 and X ≤ 2.3. Two heads will be represented as head-head-tail, tail-head-head, and head-tail-head. So, then, we have a total of 1, 2, 3, 4, 5, 6, 7 points that are equally likely. Hence,

$F_X(2.3)$ is the probability that $X \leq 2.3$. Since there are 7 points, and each point is equally likely, the probability is 7/8.

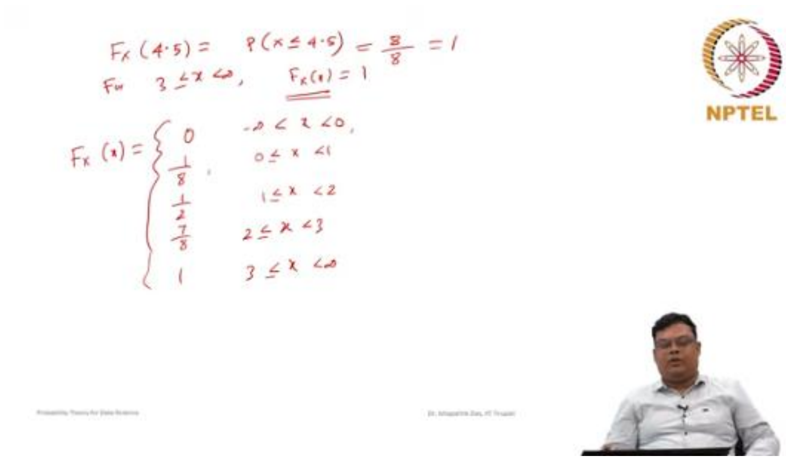
For any $x \in [2, 3)$, $F_X(x) = 7/8$ because, for any point taken here, it results in 7 out of 8. Now, we also need to consider the values 0, 1, 2, and 3. For $x \in [3, \infty)$, any value in this interval will follow the same pattern. So, let's consider an example where $x = 4$ or 4.5, or any number within this range. Then, $F_X(4.5)$ will represent the probability that $X \leq 4.5$.

First, we identify the event where $X \leq 4.5$, which includes all $s \in S_1$ such that $X(s) \leq 4.5$. Now, since X can take values only in the set {0, 1, 2, 3}, it means all s actually satisfy this relationship, covering the entire set $S_1$, as it contains all points, including head-head-head where $X = 3$. So, with this inclusion, all points are present. Therefore, the probability that $X \leq 4.5$ is 1. So, let's write this on the next page. So, $F_X(4.5)$—that means the probability that $X \leq 4.5$.




Since it includes all points, it's actually a certain event and is equal to 1. So, for $X \geq 3$ and $X < \infty$, $F_X(x) = 1$. Now, we can summarize the function. Basically, you can see that $F_X$ starts from 0. So, $F_X(x) = 0$ whenever $x < 0$, going from $-\infty$ up to 0.

Then, whenever $x \in [0, 1)$, $F_X(x) = 1/8$. When $X \in [1, 2)$, this value is 1/2. Next, when $X \in [2, 3)$, the value is 7/8. Finally, for $X \geq 3$ up to $\infty$, $F_X(x) = 1$. So, then we found the cumulative distribution function associated with this random variable.

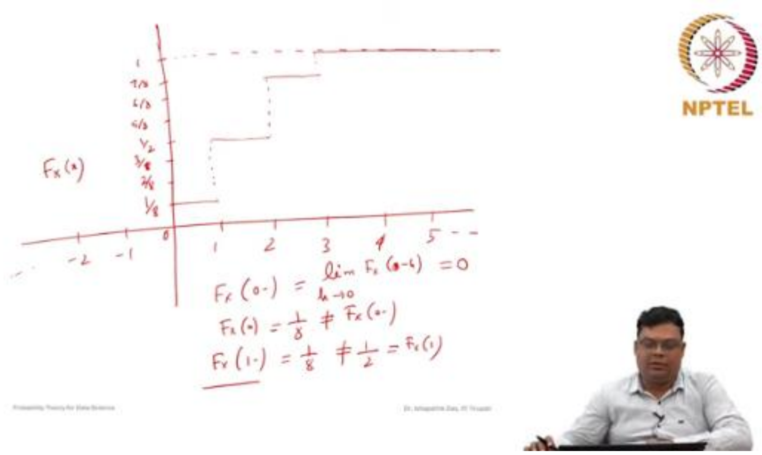



Now, we just want to represent this function graphically. Suppose this is 0, then 1, 2, 3, 4, 5, and so on, and also -1, -2, continuing like this. Now, I'll mark points like 1/4, 2/4, which is 1/2, then 3/4, and finally 4/4. Actually, let me divide it into eighths. So, 1/8, 2/8, 3/8, 4/8 (which is 1/2), 5/8, 6/8, 7/8, and finally 1.

Now, we want to show $F_X(x)$. Up to $x = 0$, this function is 0. At $x = 0$, it is 1/8. At $x = 1$, it becomes 1/2 and stays there until $x = 2$, where it reaches 7/8. Then, at $x = 3$, it becomes 1 and stays at 1.

So, this function has left discontinuities. For instance, if we take $F_X(0^-)$ as the limit of $F_X(x - h)$ as $h \to 0$, it is 0. But at $x = 0$, $F_X(0) = 1/8$, so it's not equal to $F_X(0^-)$. Similarly, $F_X(1^-) = 1/8$, but $F_X(1) = 1/2$. So, at every point where there's a probability, it has some left discontinuity, but it's always right-continuous.

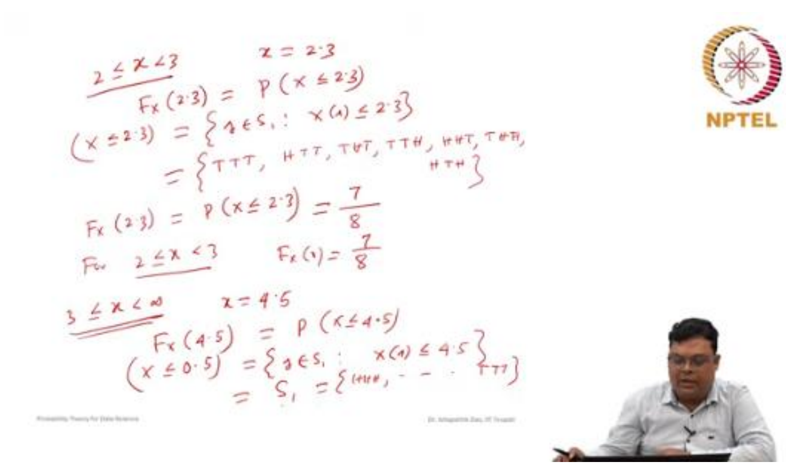
Now, let's discuss some properties of the cumulative distribution function. I hope you have understood it. Since it represents a probability, it will always fall between 0 and 1, i.e., $0 \leq F_X(x) \leq 1$. We can also show that it is a non-decreasing function. It may be constant at certain points, but it will not decrease. To prove that it is non-decreasing, consider two real numbers, $x_1$ and $x_2$, where $x_1 < x_2$.




This means F_X($x_1$) ≤ F_X($x_2$). What is F_X($x_1$)? It is the probability that $x \leq x_1$. Similarly, F_X($x_2$) is the probability that $x \leq x_2$. Now, let's look at the events corresponding to these probabilities.

The event $x \leq x_1$ consists of all $s \in S$ such that $X(s) \leq x_1$. For the event $x \leq x_2$, it includes all $s \in S$ such that $X(s) \leq x_2$. Since $x_1 < x_2$, all s satisfying $X(s) \leq x_1$ will also satisfy $X(s) \leq x_2$. So, the set defined by $x \leq x_1$ is a subset of the set defined by $x \leq x_2$. We have already established that if $A \subseteq B$, then $P(A) \leq P(B)$.

Therefore, we conclude that $P(x \leq x_1) \leq P(x \leq x_2)$. This implies that F_X($x_1$) ≤ F_X($x_2$). So, this shows that the cumulative distribution function is a non-decreasing function, and this example illustrates that property as well. Now, as $x \to \infty$, F_X(x) = 1. We won't prove this rigorously; instead, we'll provide some intuition on how it can be understood.

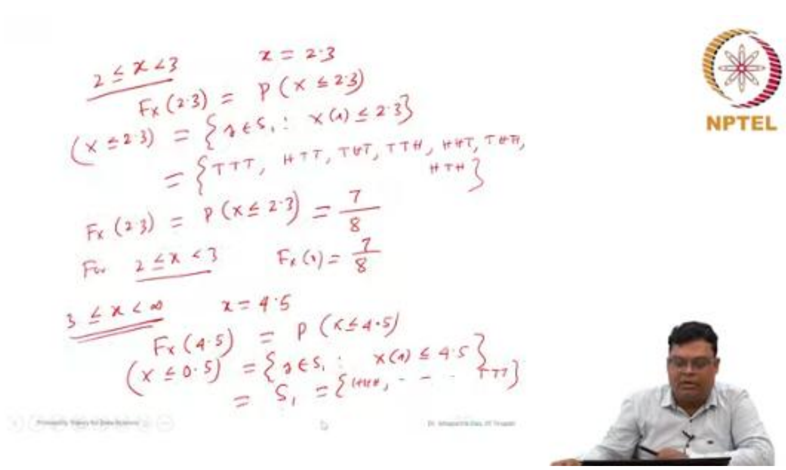


Basically, X is a random variable; it's a measurable function from the sample space S to the real numbers R, taking all values in R. What is F_X(x)? F_X(x) is the probability that X ≤ x. We can denote this as the set of all s ∈ S such that X(s) ≤ x. Now, consider the limit as X → ∞.

If we look at this set, the limit as X → ∞ will include all s ∈ S where X(s) ≤ x. Since X is a function from S to R, when X becomes very large, it will satisfy all s ∈ S. So, the intuition here is that as X → ∞, the probability P(X ≤ x) → 1. This gives us the understanding that as x → ∞, F_X(x) represents the probability of the entire sample space, which equals 1. This is not a rigorous proof; it's just an intuition.

Similarly, we can consider the case as X → -∞. When X → -∞, we look at the limit of the set where X ≤ x. Because the range of X is a subset of R, as X → -∞, no s will satisfy the condition X(s) ≤ x. Thus, this set becomes the null set, leading to the conclusion that P(X ≤ x) = 0. Again, this is just an intuitive explanation, not a rigorous proof.

We've also seen from examples that the cumulative distribution function is right-continuous but may show left discontinuities. While F_X(x) is always right-continuous, meaning F_X(a+) = F_X(a) for all a ∈ R, left discontinuities can occur due to the nature of the definition of F_X(x). These are some key properties of the cumulative distribution function. So, in this case, you can see that it is always right-continuous, but there may be some left-side discontinuities. These are some properties of the cumulative distribution function.

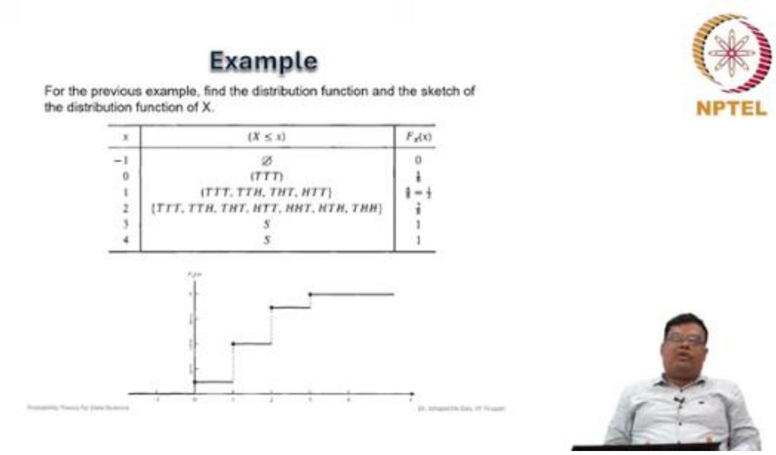


In this example, we want to find the distribution function of X. This is the same example we discussed earlier, involving the probability of the values 0, 1, 2, 3, and 4. We have already computed the probabilities at X = 0, which is 1/8. This represents the step function we have drawn in this example. We have shown the same thing here, completing the discussion of the cumulative distribution function of the random variable X.

I hope you understand it. You can go through it again or ask questions, and we can discuss further. Most of the cases we discussed involve random variables taking only integer values, like 0, 1, 2, 3, and 4. These are simple to understand, which is why we discussed them initially. However, as we mentioned before, some random experiments can have infinite sample spaces, which may even be uncountably infinite.

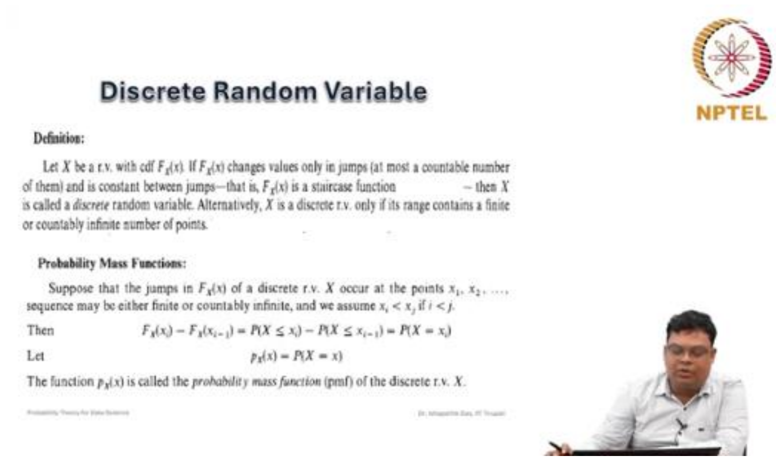
For example, if you consider the lifetime of electronic goods, the sample space would be $[0, \infty)$. In such cases, we define the random variable as continuous since we are considering a continuous range of X. We will discuss how to classify random variables based on certain characteristics, such as whether the distribution function is continuous, differentiable, or not differentiable. It may look like a staircase function or a step function. Based on the nature of the random variable, we can classify it into two categories: discrete random variables and continuous random variables.

In the next part, we'll discuss the different types of random variables: discrete random variables, continuous random variables, and how we analyze data with respect to each of these types. We'll also cover how to find probabilities related to them. So, we've learned about the distribution function and cumulative distribution function, denoted as $F_X(x)$, for a random variable associated with a random experiment. We've already discussed some examples. Looking at the distribution function here, you can see it isn't continuous—it has a left discontinuity and resembles a staircase function. Some examples will show this, while in other cases, distribution functions may be continuous.




Now, consider the range of a random variable. For a discrete random variable, it might only include values like 0, 1, 2, or 3. But let's say you're throwing a dart at a target. Since we aren't always exactly accurate, the dart lands at various points. We want to check the distance from the center to where it lands. If you measure this distance, it could be 0 if you hit the center exactly or any value up to a maximum radius, **r**.

So, the maximum distance could be **r**, and it can take any real number up to that point. Therefore, if we define the random variable as the distance from the center, it's a continuous random variable because it can be any real number in the range [0, **r**]. Based on this, we categorize random variables as either discrete or continuous. First, we'll discuss what a discrete random variable is. A discrete random variable is one where, if you look at the cumulative distribution function (CDF), the values only change in jumps.



So, in this example, that's why we discussed it first. The distribution function is constant up to 0, where its value is 0, and then at $x = 0$, the value jumps to 1/8. It stays constant again up to $x = 1$, where it jumps to 1/2, remains constant, and then at $x = 2$, it jumps to 7/8. So, it looks like a staircase. That's why we say that whenever this distribution function, $f(x)$, changes values in jumps, it has at most a countable number of these jumps.

The range of the random variable may be countable, possibly even infinite, but it would be a countable infinity. Between each jump, $f(x)$ is constant, giving it the appearance of a staircase function. Then, x is called a discrete random variable. Alternatively, X is a discrete random variable if its range contains a finite or countably infinite number of points. So, it can take an infinite number of points, but they have to be countable.

In this example, we find that this distribution function, $f(x)$, is equal to:

- 0 whenever $x < 0$,
- 1/8 whenever $0 \leq x < 1$,
- 1/2 whenever $1 \leq x < 2$,
- 7/8 whenever $2 \leq x < 3$,
- 1 whenever $x \geq 3$ up to $\infty$.

This is the distribution function. So basically, you can see that this function changes with a jump at some points. For example, if you compute $f(0-)$, it's 0, but this is not equal to $f(0)$, which is 1/8. Now, what is the difference? The difference, $f(0) - f(0-)$, is 1/8.

So similarly, we can find the difference at other points. At $x = 1$, $f(1) = 1/2$, while $f(1-) = 1/8$. So the difference, $f(1) - f(1-)$, is 3/8. This is the jump. How is this calculated? If you take the limit as $h \rightarrow 0^+$, $f(1 - h)$ approaches 1/8, which confirms this jump of 3/8. Then, at $x = 2$, we have $f(2) = 7/8$ and $f(2-) = 1/2$, making the difference $f(2) - f(2-) = 3/8$ again. Now, for $x = 3$, $f(3) = 1$ and $f(3-) = 7/8$, so the difference is 1/8.

So, when we draw it, we have these jumps: 1/8, 3/8, 3/8, and 1/8. Now we want to understand what these values represent. These jumps show the probabilities at each point where the function takes on specific values. The range is 0, 1, 2, 3.

NPTEL

Discrete Random variable | Continuous Random variable

$$F_X(x) = \begin{cases} 0 & -\infty < x < 0 \\ \frac{1}{8} & 0 \le x < 1 \\ \frac{1}{2} & 1 \le x < 2 \\ \frac{7}{8} & 2 \le x < 3 \\ 1 & 3 \le x < \infty \end{cases}$$

$$\lim_{h \to 0} F_X(1-h) = \frac{1}{8}$$

$R_X = \{$

$$F_X(0-) = 0 \neq F_X(0) = \frac{1}{8}$$

$$F_X(0) - F_X(0-) = \frac{1}{8}$$

$$F_X(1) - F_X(1-) = \frac{1}{2} - \frac{1}{8} = \frac{3}{8}$$

$$F_X(2) - F_X(2-) = \frac{7}{8} - \frac{1}{2} = \frac{3}{8}$$

$$F_X(3) - F_X(3-) = 1 - \frac{7}{8} = \frac{1}{8}$$