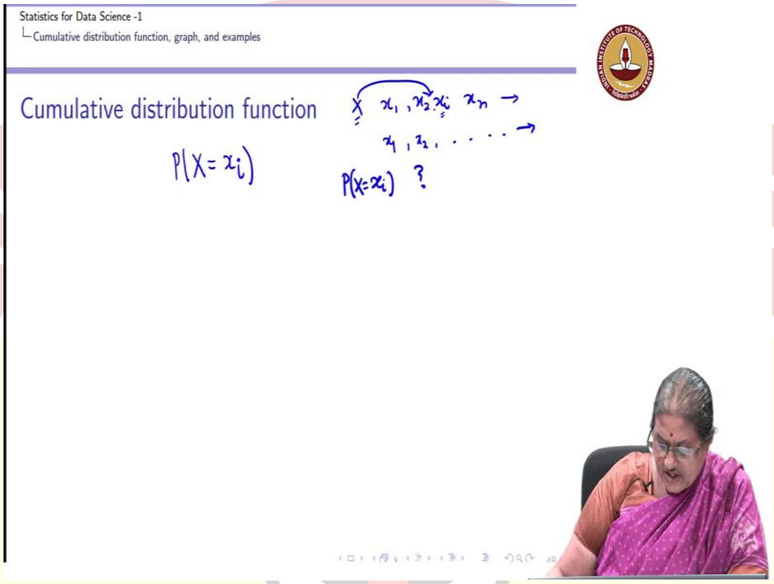
INDIAN INSTITUTE OF TECHNOLOGY MADRAS • सिद्धिर्भवति कर्मजा •

IIT Madras

ONLINE DEGREE

Statistics for Data Science-1
Professor Usha Mohan
Department of Management Studies
Indian Institute of Technology, Madras
Random variables - cumulative distribution function

(Refer Slide Time: 0:15)



Statistics for Data Science -1

└ Cumulative distribution function, graph, and examples

Cumulative distribution function

$X \quad x_1, x_2, x_i \quad x_n \rightarrow$

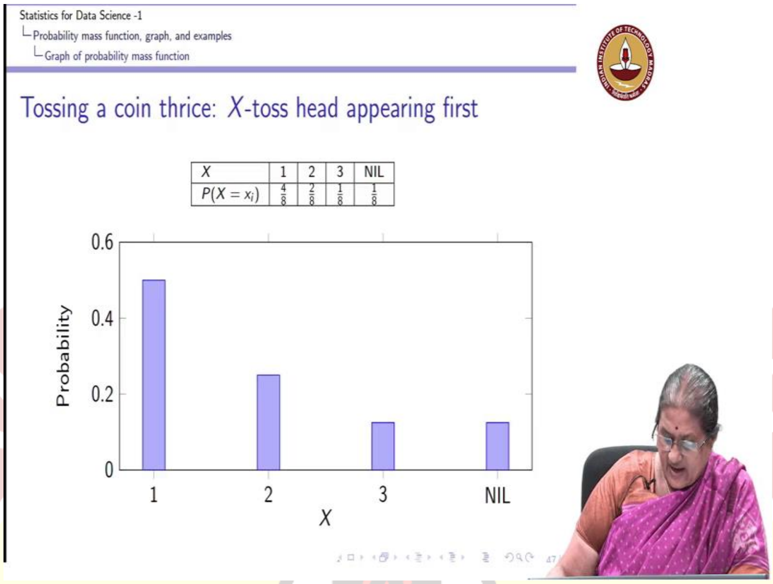
$x_1, x_2, \ldots \ldots \rightarrow$

$P(X=x_i)$?

$P(X = x_i)$

So, the next concept we are going to learn about is, what we understand by a cumulative distribution function. A probability mass function gave us a $P(X)$ taking a particular value xi. So, we assume that X takes values x1, x2, xn if they take finite number of values, it takes the values x1, x2, so forth if it takes countably infinite number of values and the probability mass function or distribution just told us what is the probability with which X takes a particular value xi. That is what a distribution function tells us. In other words, it tells us that what is the chance of this random variable taking a particular value.

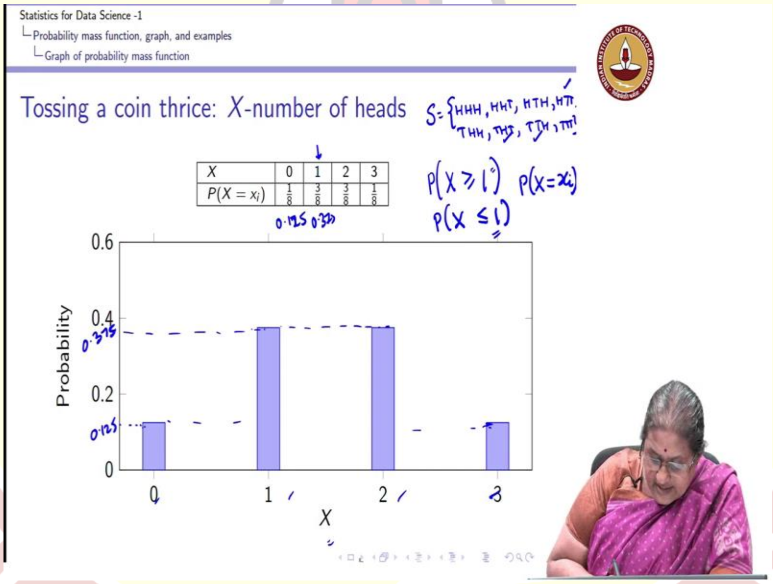
(Refer Slide Time: 1:20)



Statistics for Data Science -1
└Probability mass function, graph, and examples
└Graph of probability mass function

Tossing a coin thrice: X-toss head appearing first

X	1	2	3	NIL
$P(X = x_i)$	$\frac{4}{8}$	$\frac{2}{8}$	$\frac{1}{8}$	$\frac{1}{8}$



Statistics for Data Science -1
└Probability mass function, graph, and examples
└Graph of probability mass function

Tossing a coin thrice: X-number of heads

$S = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$

X	0	1	2	3
$P(X = x_i)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

0.125 0.375

$P(X \geq 1)$ $P(X = x_i)$

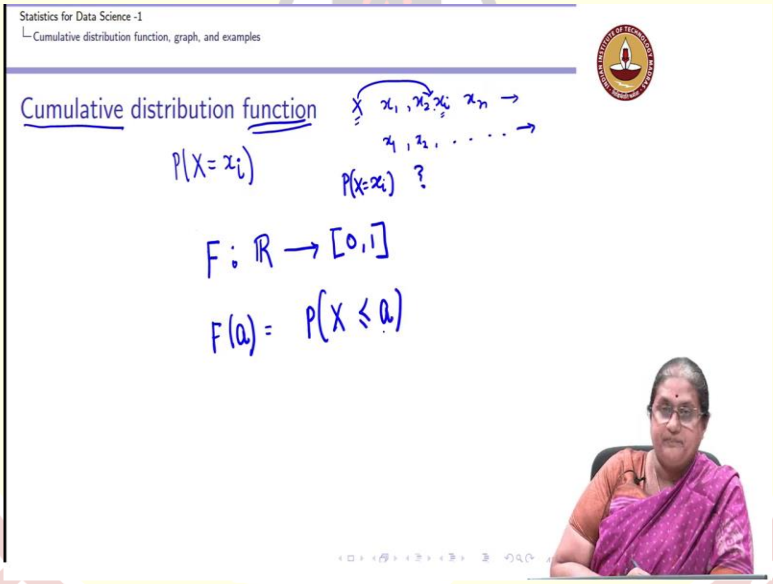
$P(X \leq 1)$

But sometimes we might be interested in knowing for example, we go back to this tossing the coin thrice and look at the solution here. So, you can see that I am counting the number of heads, so I either have, so recall my sample space here was a head, head, head; head, head, tail; head, tail, head; head, tail, tail; tail, head, head; tail, head, tail; tail, tail, head and a tail, tail, tail. Suppose I am interested in knowing what is the chance, so this probability gave me what is the chance of me having 1 head?

I can see that, that corresponds to this outcome, this outcome and this outcome and which is $\frac{3}{8}$. This is how we got a probability mass function. But now suppose I am interested in asking, what is the chance that I have got at least 1 head? So, the way I am translating this is what is the chance that $P(X \geq 1)$? The chance of me getting at least 1 head is equal to this or I could also ask, what is the chance of me getting at most 1 head in 3 tosses? So, in other way, I am asking $X \leq 1$.

So, instead of looking at what is the chance of X taking a particular value, we might be interested in knowing what is the chance of X is less than or equal to a particular value or greater than or equal to a particular value?

(Refer Slide Time: 3:11)



Statistics for Data Science -1
└ Cumulative distribution function, graph, and examples

Cumulative distribution function

$X \quad x_1, x_2, x_i \quad x_n \rightarrow$

$x_1, x_2, \ldots \rightarrow$

$P(X = x_i)$

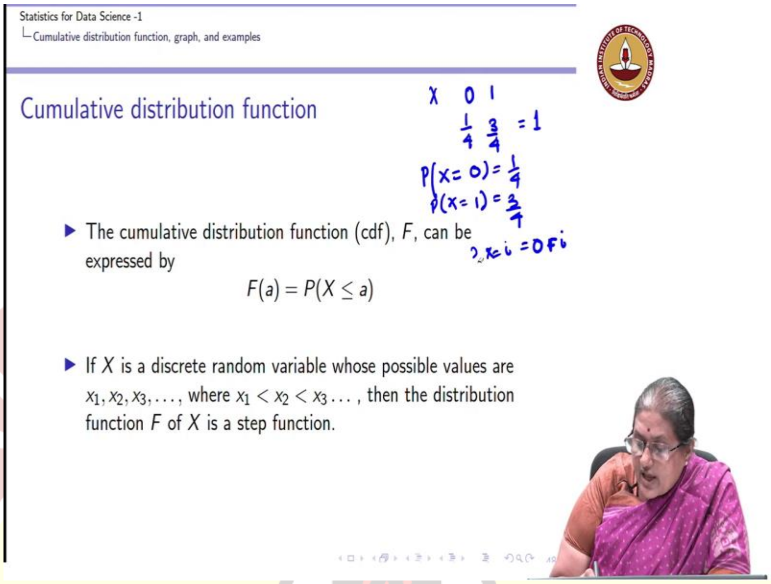
$P(X=x_i)$?

$F : \mathbb{R} \rightarrow [0,1]$

$F(a) = P(X \leq a)$

To answer this question we introduce what is known as a cumulative distribution function. As the name suggests a cumulative distribution is that function which accumulates the probabilities at different points. So, this is a function, so the minute we refer to a function, I need to understand the function is defined on what, this cumulative distribution function which is given by capital F, typically it is referred to as capital F is defined for every real value and it takes values in the closed intervals 0, 1 and how do we define it? For every real value, $F(a)$ is $P(X \leq a)$.

(Refer Slide Time: 4:16)



Statistics for Data Science -1
└ Cumulative distribution function, graph, and examples

Cumulative distribution function

$X \quad 0 \quad 1$

$\frac{1}{4} \quad \frac{3}{4} = 1$

$P(X=0) = \frac{1}{4}$

$P(X=1) = \frac{3}{4}$

$P(X=i) = 0$ fi

- The cumulative distribution function (cdf), F, can be expressed by

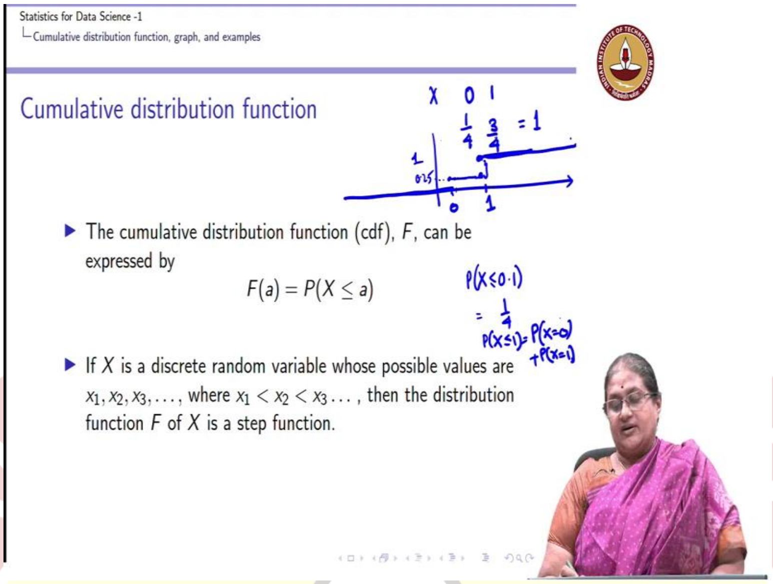
$$F(a) = P(X \leq a)$$

- If X is a discrete random variable whose possible values are $x_1, x_2, x_3, \ldots$, where $x_1 < x_2 < x_3 \ldots$, then the distribution function F of X is a step function.

I repeat, a cumulative distribution function F can be expressed as $F(a)$ is $P(X \leq a)$. So, given this $F(a)$ is defined for every value of a on my real line. So, let us look at an example. For example, if X is taking values, whose possible values are X$_1$, X$_2$, X$_3$, let us start with a very simple example.

X takes finite value 0 and 1 with probability 1 by 4 and 3 by 4. I know that the way my pmf is defined is $P(X = 0)$ is $\frac{1}{4}$, $P(X = 1)$ is $\frac{3}{4}$, it is a probability mass function because both of them add up to 1. And in addition, $P(X = 1) = 0$ for all other i. So, this is equal to 0 for everything else.

(Refer Slide Time: 5:44)



Statistics for Data Science -1

└Cumulative distribution function, graph, and examples

Cumulative distribution function

- The cumulative distribution function (cdf), F, can be expressed by

$$F(a) = P(X \leq a)$$

- If X is a discrete random variable whose possible values are $x_1, x_2, x_3, \ldots$, where $x_1 < x_2 < x_3 \ldots$, then the distribution function F of X is a step function.

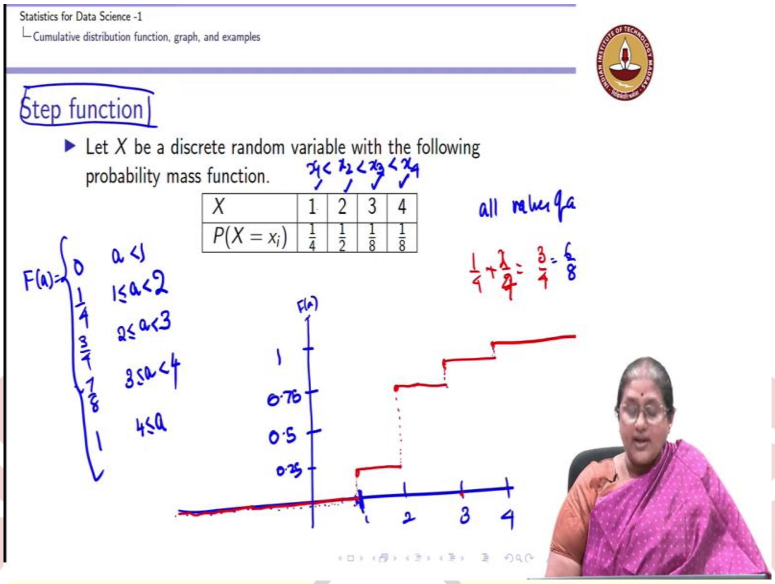
So, what is my chance so if I am going to map every point on my real line to this, then the way I can start is I look at my real line, I have a 0 here so till the point I have touch the 0, my probability is going to be 0 because I have defined $P(X)$ is equal to 0 for all points other than 0.

At the point 0, I have a probability of $\frac{1}{4}$ which is 0.25. At 0.1 so if I am looking at $P(X \leq 0.1)$, I know that this can be $P(X \leq 0) + P(X = 0.1)$, is 0, but probability X is less than or equal to 0 plus probability X is equal to 0.1, I can see that this I already know is $\frac{1}{4}$, this is 0, so this would also be $\frac{1}{4}$.

At 0.2, it will again be $\frac{1}{4}$. At 0.5, it will again be 1 $\frac{1}{4}$. At 0.6 it will again be $\frac{1}{4}$. So, it continues to be 0.4 till the time it hits 1. At X equal to 1, it takes a probability $\frac{3}{4}$. So, what happens? $P(X.1)$, is the same as $P(X \leq 0) + P(X = 1)$, , so it jumps and it takes so this, if this value is 0.25, this value is going to be 1, it is jumping at this value and you can see that it continues to take the value 1 after that. So, this is what is called a cumulative distribution function.

So, what is a distribution function? In the case of a discrete random variable which takes values X_1 which is less than X_2 which is strictly less than X_3. This is a step function. So, now let us look at another example.

(Refer Slide Time: 8:15)



Statistics for Data Science -1

└ Cumulative distribution function, graph, and examples

Step function

- Let X be a discrete random variable with the following probability mass function.

$x_1 < x_2 < x_3 < x_4$

X	1	2	3	4
$P(X = x_i)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{8}$	$\frac{1}{8}$

all values of a

$\frac{1}{4} + \frac{2}{4} = \frac{3}{4} = \frac{6}{8}$

$$F(a) = \begin{cases} 0 & a < 1 \\ \frac{1}{4} & 1 \le a < 2 \\ \frac{3}{4} & 2 \le a < 3 \\ \frac{7}{8} & 3 \le a < 4 \\ 1 & 4 \le a \end{cases}$$

For example, I have X which takes values 1, 2, 3, 4, this is my x_1 which is strictly less than x_2, which is strictly less than x_3, which is strictly less than x_4. This is the value. First step, let us check if it is a probability mass function, yes. 1 by 4, 3 by 4, 3 by 4 is nothing but 4 by, 3 by 4 is nothing but 6 by 8, 6 plus 1 7, 7 plus 1 8, 8 by 8 so it is a probability mass function, all of them are non-negative. So, it is a probability mass function.

So, now again recall, I define $F(a)$ to be $P(X \leq a)$. Now, let us start with the following. I need to define $F(a)$ for all values of a. So, for what is it, so let me start with $(a < 1)$. So, if I go back, I have 1 here, I have 2 here, I have 3 here, I have 4 here. For all a is less than 1, I know probability of x taking any of the values here is equal to 0, because x takes only discrete values 1, 2, 3, 4. So, it is going to be 0. So, I know that let me use a different colour here.

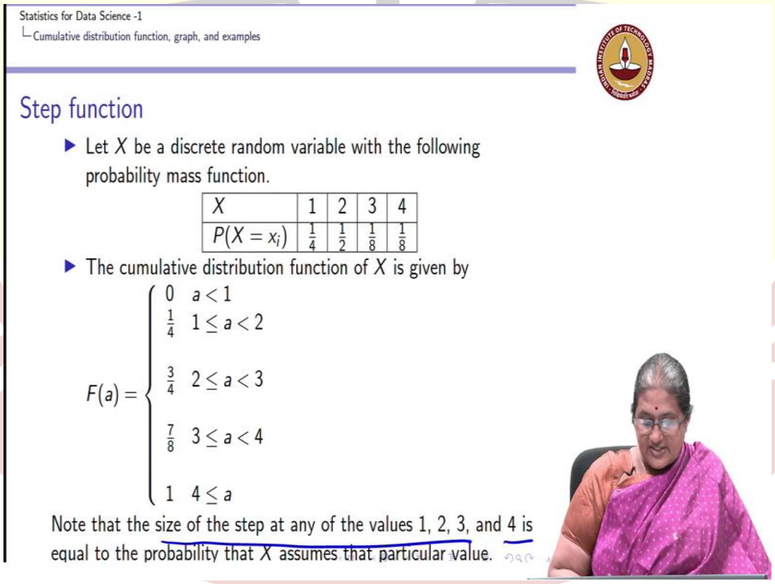
So, I know that till it, so I know that the probabilities or $F(a)$ which is given by the red line is going to be 0 till it hits 1. So, this would be 0. So, $F(a)$ is going to be 0 as long as a is less than 1. Now let me have this portion here which is $F(a)$. So, I am going to write down my $F(a)$ here, so I know $F(a)$ is 0 as long as a is less than 1. Now let us look at the interval, once it hits a, it takes the value 0.25, so let me have a 0.25 here, this is a 0.5, this is a 0.75 and this is 1. There is a 0.25 and it continues till X goes to 1. So, this is what is my value, so it takes the value 1 by 4 for a is between 1 and 2.

So, again, I will plot it using my red line. So, I know that what is it? Between 1 and 2, it is going to be, now there is a discontinuity here and that I am just going to show by dotted lines. Now when I hit 2, the probability of X less than or equal to, so it is again it goes from 1 by 4 Plus 1 by 2 which is nothing but 1 by 4 plus 1 by 2 is this is 2 by 4, so I have a 3 by 4 which is 0.75. So, I have a $P(X < 2)$, it goes here, it continues with the same probability till it hits 3.

So, I have a probability which I write as 3 by 4, for 2 is less than or equal to a is less than 3. Once it hits 3, so this 6, this is nothing but 6 by 8. So, it becomes 7 by 8 so again what happens to my graph from 3 to 4, it is 7 by 8 which is very close to this. Again there is a discontinuity here, after X equal to 4, it continues with 1. There is a discontinuity here.

So, this is how the cumulative, so what do I have here for a so I will have 7 by 8 for 3 is less than or equal to a is less than 4 and 1 for 4 is less than or equal to a and this is what I have plotted here, I know that the plot of this function is what we, it resembles a step, hence it is called a step function We can see that there are discontinuities which I have shown by the dotted line here.

(Refer Slide Time: 13:21)



Statistics for Data Science -1
└ Cumulative distribution function, graph, and examples

Step function

- Let X be a discrete random variable with the following probability mass function.

X	1	2	3	4
$P(X = x_i)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{8}$	$\frac{1}{8}$

- The cumulative distribution function of X is given by

$$F(a) = \begin{cases} 0 & a < 1 \\ \frac{1}{4} & 1 \leq a < 2 \\ \frac{3}{4} & 2 \leq a < 3 \\ \frac{7}{8} & 3 \leq a < 4 \\ 1 & 4 \leq a \end{cases}$$

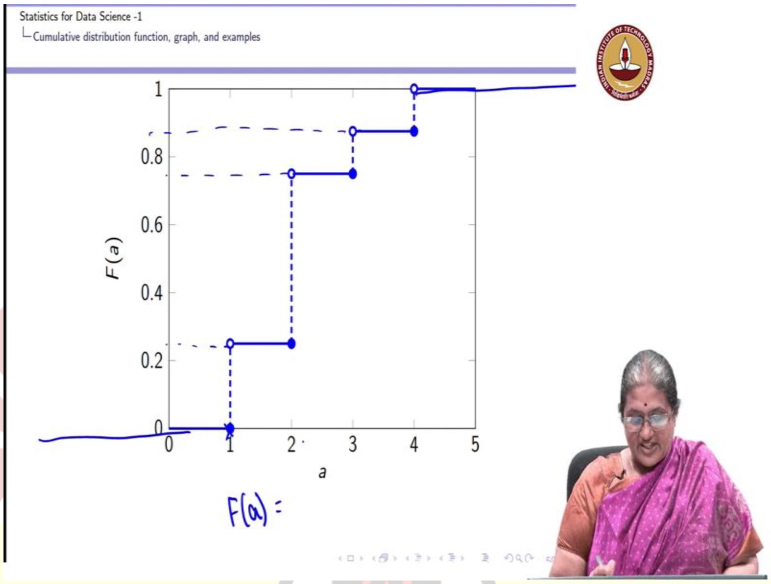
Note that the size of the step at any of the values 1, 2, 3, and 4 is equal to the probability that X assumes that particular value.

So, you can see that this is my F of a which was 1 by 4, 3 by 4, 7 by 8 and 1 another thing which you should notice is the size of the step, what is the size of a step here? The size of the step here was 0.75 to 0.25 which is 0.5 which is $P(X = 2)$, the size of the step here is 0.25 which is the $P(X = 1)$, the size of the step here is 1 by 8, the size of the step here is 1 by 8 which are the

INDIAN INSTITUTE OF TECHNOLOGY MADRAS

$P(X = 3)$ and $P(X = 4)$ and that is what we have here. The size of the step at any of the values is equal to the probability that X assumes at that particular value and what are the values that X assumes? X assumes 1, 2, 3, 4.

(Refer Slide Time: 14:21)

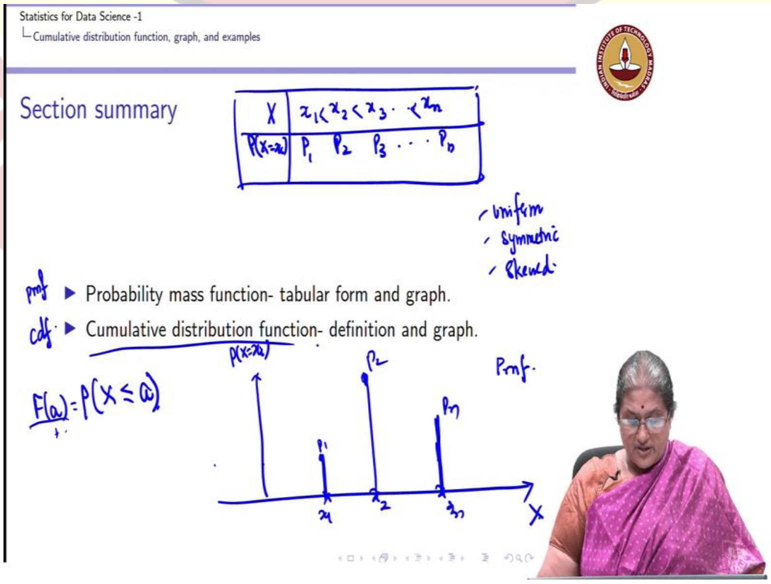


Statistics for Data Science -1

└Cumulative distribution function, graph, and examples

So, that is how you can see that this is the step function, again for X is less than or equal to 1, it was 0. This is 0.25, this is 0.75, this is 1, 3 by 6, 7 by 8 and after X equal to 4, it continues. So, this is defined for all values of my a. So, the cumulative distribution function is defined for all real values of a and the graph of a cumulative distribution function in the case of a discrete random variable which takes values X_1, X_2, X_n such that X_1 is less than or equal to X_2 is less than, less than X_2 is less than X_n, the distribution, the cumulative distribution is a step function.

(Refer Slide Time: 15:29)



Statistics for Data Science -1

└Cumulative distribution function, graph, and examples

Section summary

X	$x_1 < x_2 < x_3 \cdot < x_n$
$P(X=x_i)$	p_1 p_2 p_3 $\cdots$ p_n

- uniform
- symmetric
- skewed

pmf ▶ Probability mass function- tabular form and graph.

cdf ▶ Cumulative distribution function- definition and graph.

$F(a) = P(X \leq a)$

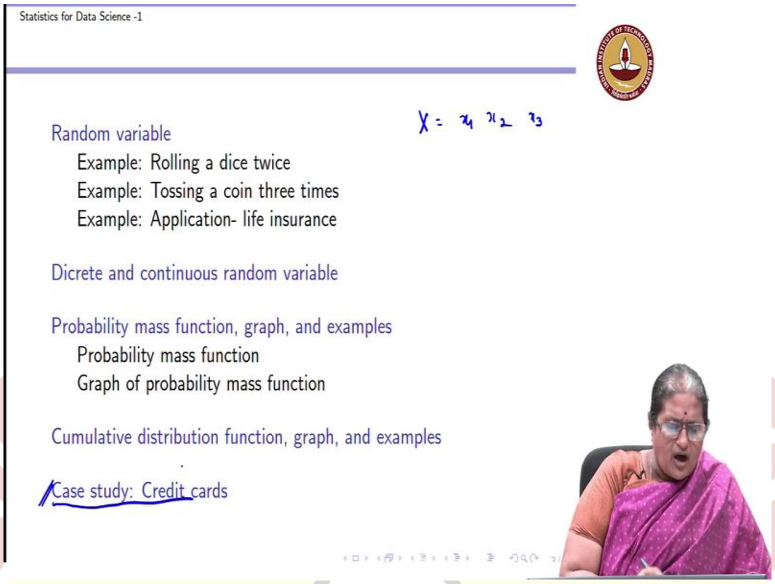
So, in summary what we have seen is, we have seen so far what is a probability mass function. So, given a random variable which takes countably finite or infinite values associated with each one of the values I define what is the probability of X taking a particular value.

I can represent it in a tabular form or I can illustrate it by the values of X on the x axis and the probability of X on the y axis. This is referred to as the graph of the pmf for example, if it take x_1, x_2, xn with probabilities p_1, p_2, and p_n, I know that this is the graph of the probability mass function.

From the graph I can describe the distribution, while describing the distribution I can see whether the distribution is uniform, whether it is symmetric, whether it is skewed. These are the things which we can see from the distribution. And then we further define what was the probably cumulative distribution function, this we define for every value or real value of a which is nothing but probability X is less than or equal to a and in the case the random variable takes x_1, x_2, x_n, with x_1 strictly less than x_2 is strictly less than x_n, then we saw that the graph of this cumulative distribution function is a step function.

Many books refer to the probability mass function as pmf and cumulative distribution function as cdf. We need these concepts to understand the bigger inferential statistics part but this at a conceptual level you need to understand what is a probability mass function and what is a cumulative distribution function.

(Refer Slide Time: 17:59)



Statistics for Data Science -1

$X = x_1 \quad x_2 \quad x_3$

Random variable

Example: Rolling a dice twice

Example: Tossing a coin three times

Example: Application- life insurance

Dicrete and continuous random variable

Probability mass function, graph, and examples

Probability mass function

Graph of probability mass function

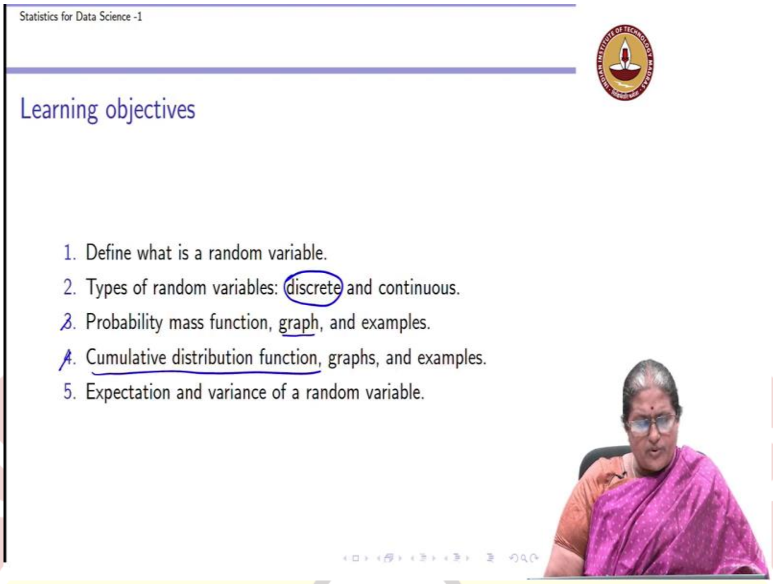
Cumulative distribution function, graph, and examples

Case study: Credit cards

The next question we are going to answer is again I said we are interested in answering or we are interested in knowing questions about typically we would be interested in answering questions about so, I know now what is a random variable, I know again what is the variable which the values this variable takes.

So, now the next thing which we are going to understand is suppose I am interested in knowing from a population I take a random sample of people and the question I ask them is how many credit cards they own? Again, there is a count of the number, so I can model the response which is the number of credit cards owned by a person as a random variable in particular I can model it as a different random variable.

(Refer Slide Time: 18:55)



Statistics for Data Science -1

Learning objectives

1. Define what is a random variable.
2. Types of random variables: discrete and continuous.
3. Probability mass function, graph, and examples.
4. Cumulative distribution function, graphs, and examples.
5. Expectation and variance of a random variable.

So, we will next look at an application of how to use the concept of both the probability mass function and the probability cumulative distribution function to answer questions from a real application which is about the number of credits cards a person uses. This is what we will be doing next.

INDIAN INSTITUTE OF TECHNOLOGY MADRAS • सिद्धिर्भवति कर्मजा •