INDIAN INSTITUTE OF TECHNOLOGY MADRAS

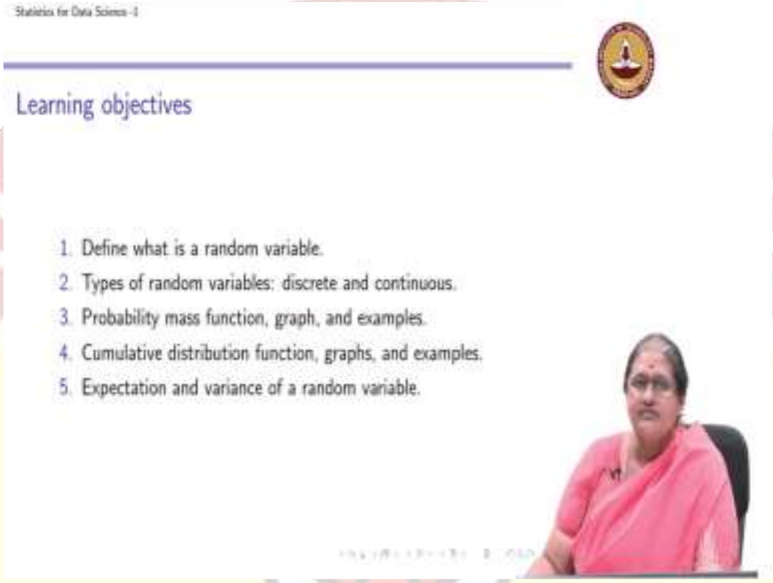
सिद्धिर्भवति कर्मजा

IIT Madras

ONLINE DEGREE

Statistics for Data Science - 1
Professor Usha Mohan
Department of Management Studies
Indian Institute of Technology, Madras
Discrete Random Variable

(Refer Slide Time: 0:16)



Statistics for Data Science -1

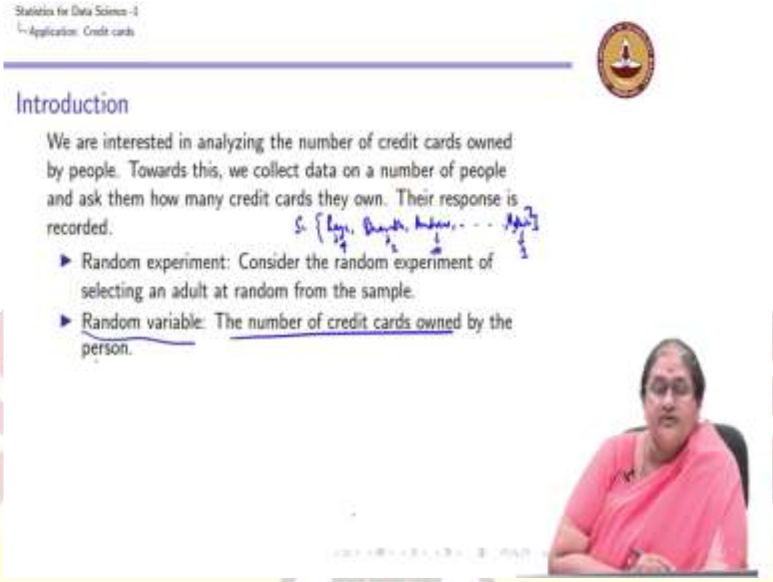
Learning objectives

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, in this lecture, we will just have a brief review of what we have done so far. So, we have already defined what is a random variable, we looked at discrete and continuous random variables, how these arise naturally in a lot of applications and then afterwards we introduced what is called a probability mass function that is when we want to know about the distribution of a random variable, we represent the distribution of a discrete random variable with what we know or what we call the probability mass function.

We also saw that the probability mass function, the graph of a probability mass function gives us an idea of the distribution of the random variable. We also introduce the notion of a cumulative distribution function which also helps us to understand how a random variable is distributed on in a complete range. So, now we are going to introduce an extremely important concept of a random variable which is called the expectation of a random variable.

(Refer Slide Time: 1:35)



Statistics for Data Science -1
└ Application: Credit cards

Introduction

We are interested in analyzing the number of credit cards owned by people. Towards this, we collect data on a number of people and ask them how many credit cards they own. Their response is recorded.

$S: \{\underset{4}{\text{Raja}}, \underset{0}{\text{Bharath}}, \underset{1}{\text{Andrew}}, \ldots \underset{3}{\text{Maithali}}\}$

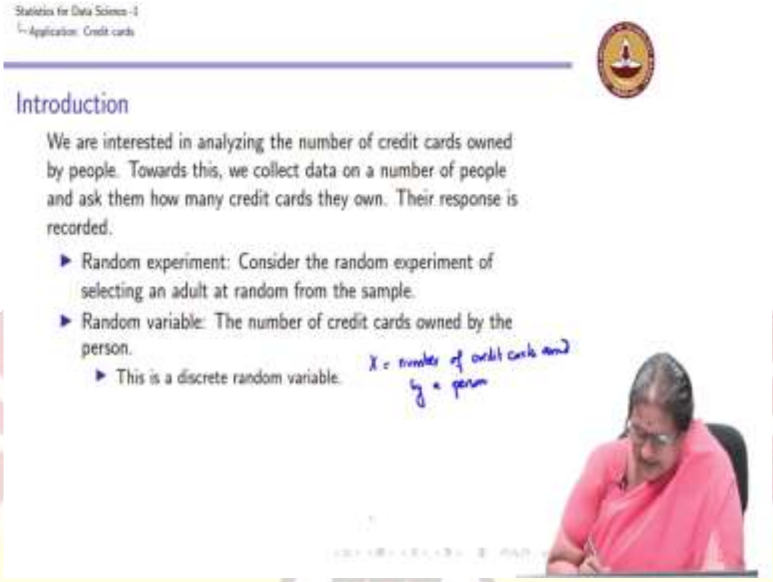
- Random experiment: Consider the random experiment of selecting an adult at random from the sample.
- Random variable: The number of credit cards owned by the person.

So, we first go and look at an application of the concepts you have learned so far. Suppose we are interested in analyzing the number of credit cards owned by people or adult population in a particular geographical area towards this we collect data on a number of people and the data we have collected on them is we ask them or what we record is the number of credit cards that they own.

So, if I am looking at it as a random experiment where I am choosing a random or a person from this collection of people, my random experiment is to select an adult at random. So, it could be 1, I am selecting an adult I just note down the person's name, I am selecting another adult, so if I am selecting 50 adults my sample space is going to be each 1 of the adults whom I have sampled from this population.

Now, associated with each adult, so I have a sample space for example my sample space could be Raja, it could be Bharat it, could be Andrew and so forth, all these people whom I am sampling it could be some then after it could be Maithali, I could just write down all these people associated with each outcome is the number of credit cards owned by the person. So, Raja could respond by saying that he owns 4 credit cards, this person could say no, this says 0, this says 1, so associated with each outcome I am defining the random variable which is the number of credit cards owned by the person.

(Refer Slide Time: 3:34)



Statistics for Data Science -1
└ Application: Credit cards

Introduction

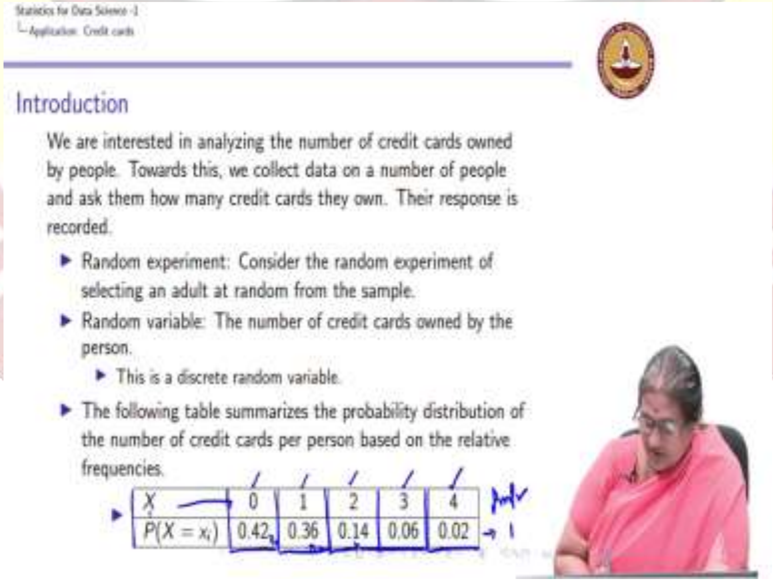
We are interested in analyzing the number of credit cards owned by people. Towards this, we collect data on a number of people and ask them how many credit cards they own. Their response is recorded.

- Random experiment: Consider the random experiment of selecting an adult at random from the sample.
- Random variable: The number of credit cards owned by the person.
  - This is a discrete random variable.

X = number of credit cards owned by a person

So, I can see that when I am looking at number of credit cards, this number is a discrete random variable. So, let me denote that by x, so x is the number of credit cards owned by a person. Now, this takes discrete values because I cannot own 1 and a half credit cards, I can own it in discrete values and the values I can see that this takes are the following.

(Refer Slide Time: 4:06)



Statistics for Data Science -1
└ Application: Credit cards

Introduction

We are interested in analyzing the number of credit cards owned by people. Towards this, we collect data on a number of people and ask them how many credit cards they own. Their response is recorded.

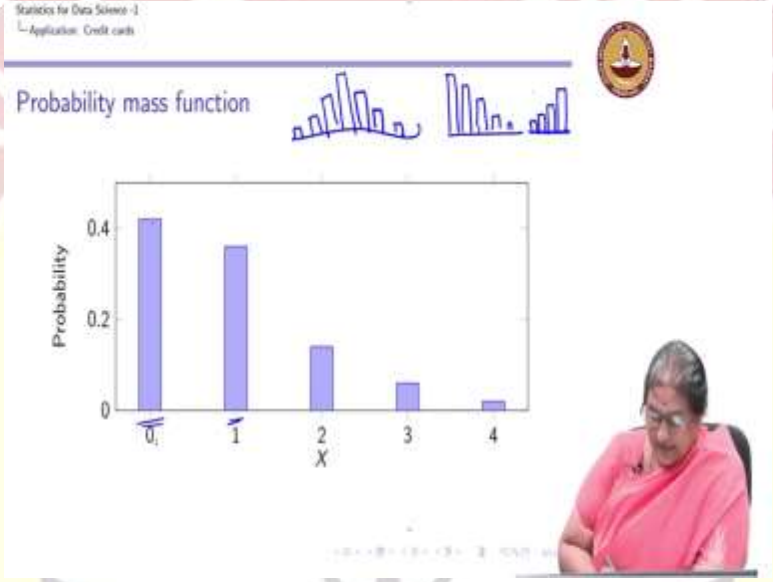
- Random experiment: Consider the random experiment of selecting an adult at random from the sample.
- Random variable: The number of credit cards owned by the person.
  - This is a discrete random variable.
- The following table summarizes the probability distribution of the number of credit cards per person based on the relative frequencies.

X	0	1	2	3	4
$P(X = x_i)$	0.42	0.36	0.14	0.06	0.02

→ 1

I can own 0, a person has either owns no credit card or 1 credit card or 2 credit cards or 3 credit cards or 4 credit cards, so that is the value that this random variable x takes which is 0, 1, 2, 3 and 4. Now, this random variable takes these values with a particular probability I can see that it is 0.42 with that now $P(X=0)=0.42$, $P(X=1)=0.36$, it takes the value $P(X=2)=0.14$, $P(X=3)=0.06, P(X=4)=0.02$, I can verify that this is a probability mass function, I add all of them and I can see that this is 0.50 + 0.50 which I get is equal to 1, hence this is a probability mass function, further everything is greater or equal to 0. So, hence I have a probability mass function which is given to me.

(Refer Slide Time: 5:25)



Statistics for Data Science -1

└ Application: Credit cards

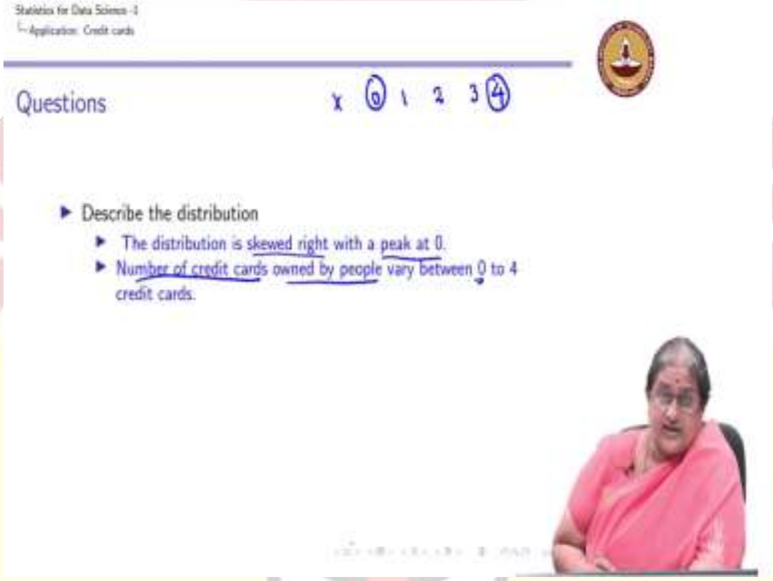
Probability mass function

Now, let us plot the probability mass function, the minute I plot this probability mass function, this gives us the distribution of the random variable, so you can see the graph of the probability mass function, the probability mass function, the distribution, remember we said the probability mass function tells us about how this random variable is distributed, we can see that the 0 has a peak when I say 0 has a peak, this says that I can expect that most of the people do not own any credit card.

And then afterwards I see that many people own 1 credit card, whereas very few people own 3 and 4 credit cards respectively. So, you can see that this probability, this table that is the probability mass function gives us an idea of the distribution. So, as an application let us look at describing the distribution. Again if you look at this distribution, you can see that it is speed at 0 and then followed by 1, then followed by 2, then followed by all the other values.

So, you can see that there is a skew to this distribution. Again recall, I say a distribution is symmetric if it is of this kind, we have seen a symmetric distribution, it is skewed, it could be either skewed this way or it could be skewed in this way and we can see that this particular example exhibits a skewness where I have a peak at the lower most value and it is tapering down.

(Refer Slide Time: 7:21)



Statistics for Data Science -1
└ Application: Credit cards

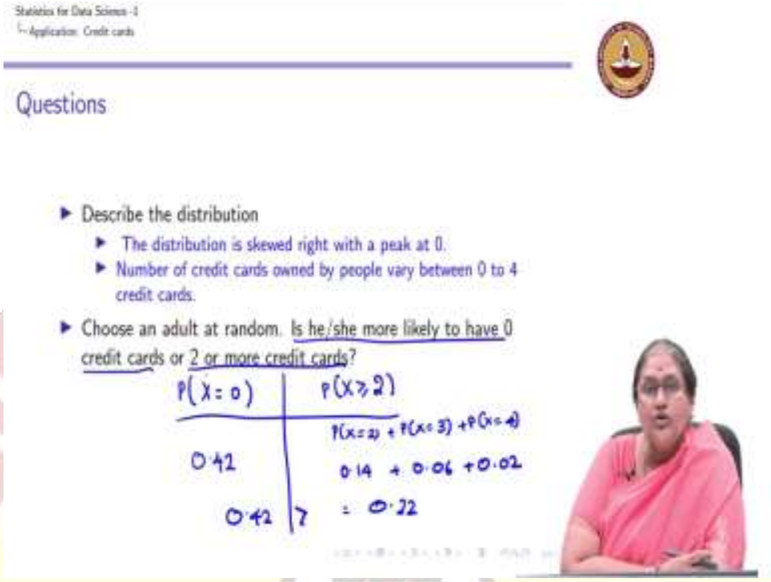
Questions

x 0 1 2 3 4

- Describe the distribution
    - The distribution is skewed right with a peak at 0.
    - Number of credit cards owned by people vary between 0 to 4 credit cards.

So, I can see that if I am asked to describe the distribution, I can say that it is skewed right with a peak at 0 and the values of x are 0, 1, 2, 3 and 4 which I can further articulate by saying that the number of credit cards owned by people vary between 0 and 4. So, now let us use this information to answer a few questions, because finally data analysis is about extracting information from the given data.

(Refer Slide Time: 8:02)

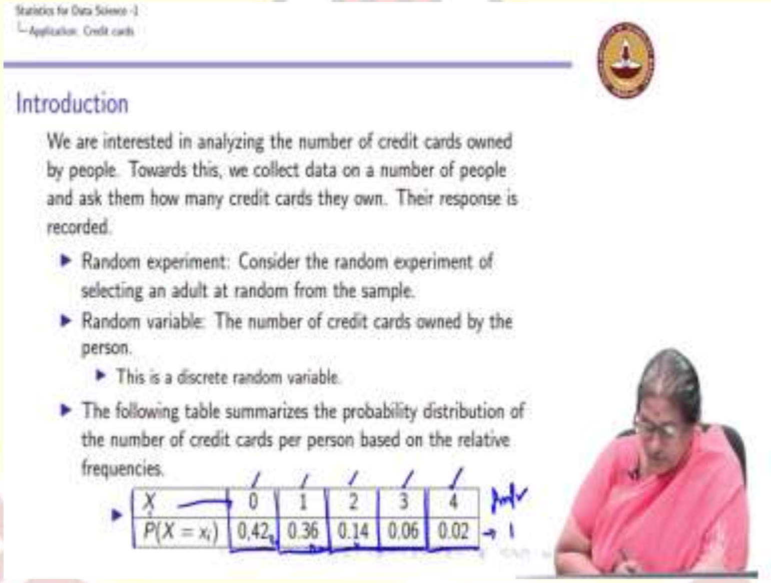


Statistics for Data Science -1
└ Application: Credit cards

Questions

- Describe the distribution
  - The distribution is skewed right with a peak at 0.
  - Number of credit cards owned by people vary between 0 to 4 credit cards.
- Choose an adult at random. Is he/she more likely to have 0 credit cards or 2 or more credit cards?

$P(X=0)$	$P(X\geq 2)$
	$P(X=2)+P(X=3)+P(X=4)$
0.42	$0.14+0.06+0.02$
0.42	$>$ $=0.22$



Statistics for Data Science -1
└ Application: Credit cards

Introduction

We are interested in analyzing the number of credit cards owned by people. Towards this, we collect data on a number of people and ask them how many credit cards they own. Their response is recorded.

- Random experiment: Consider the random experiment of selecting an adult at random from the sample.
- Random variable: The number of credit cards owned by the person.
  - This is a discrete random variable.
- The following table summarizes the probability distribution of the number of credit cards per person based on the relative frequencies.
-

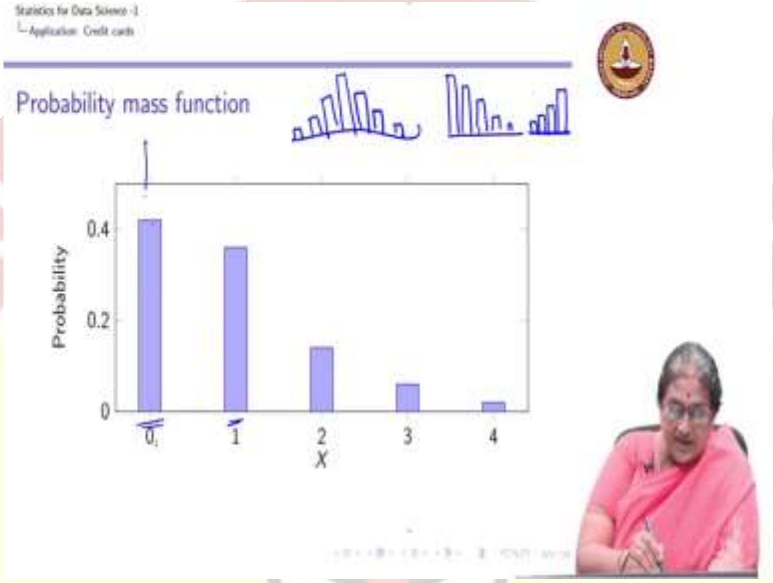
X	0	1	2	3	4	pmf
$P(X=x_i)$	0.42	0.36	0.14	0.06	0.02	→ 1

So, how do I use this information? I choose an adult at random, is he or she more likely to have no credit cards or 2 or more credit cards? Now, how do I answer this question? I know x is the total number of credit cards. So, to answer this question that likely to have no credit cards, I am looking at $P(X=0)$, 2 or more credit cards I am looking at $P(X\geq 2)$.

Now, what is $P(X=0)$? We can go back to our distribution and see that the $P(X=0)=0.42$, this comes from my probability mass function. Now, $P(X\geq 2)=P(X=2)+P(X=3)+P(X=4)$.

$P(X = 2)$ again we go back to our distribution, it is 0.14, this is 0.06 and 0.02, so I can see that $P(X \geq 2) = 0.14 + 0.06 + 0.02 = 0.22.$ I know 0.42 is strictly greater than 0.22, so I can say that an adult is more likely to have no credit cards than own 2 or more credit cards.

(Refer Slide Time: 10:02)

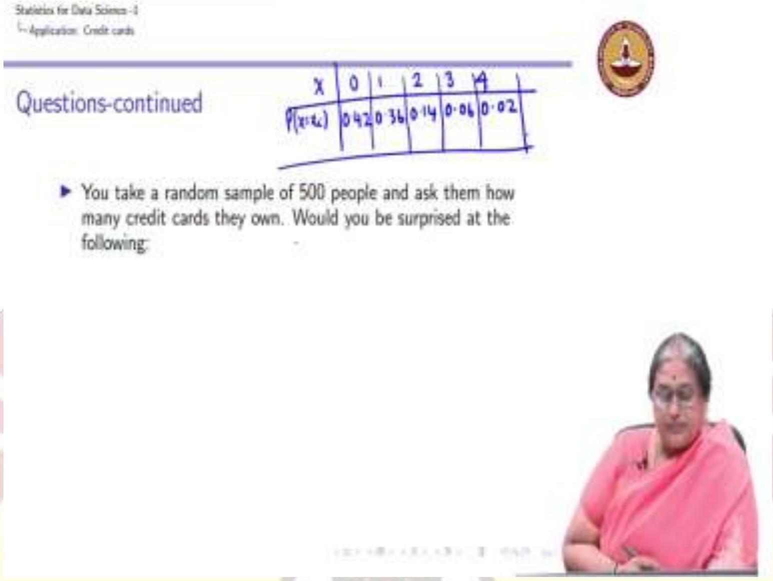


Statistics for Data Science -1
└Application: Credit cards

Probability mass function

So, that is very clear from the distribution table itself, because the peak is at x = 0.

(Refer Slide Time: 10:14)

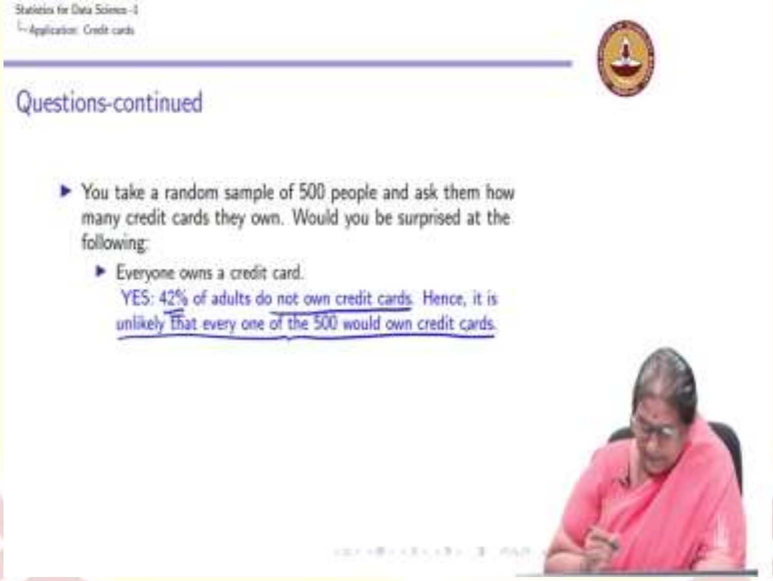


Statistics for Data Science -1
└ Application: Credit cards

Questions-continued

x	0	1	2	3	4
$P(X=x_i)$	0.42	0.36	0.14	0.06	0.02

- You take a random sample of 500 people and ask them how many credit cards they own. Would you be surprised at the following:



Statistics for Data Science -1
└ Application: Credit cards

Questions-continued

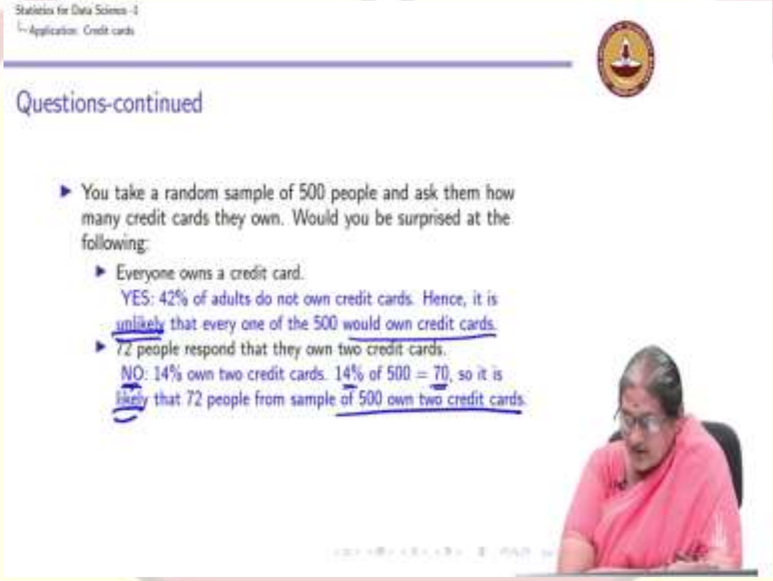
- You take a random sample of 500 people and ask them how many credit cards they own. Would you be surprised at the following:
  - Everyone owns a credit card.
    YES: 42% of adults do not own credit cards. Hence, it is unlikely that every one of the 500 would own credit cards.

Now, let us look at another question which we are trying to answer from here. The next question we are trying to answer is you take a random sample of 500 people and you ask them how many credit cards they own. So, let me rewrite the distribution x takes the value 0, 1, 2, 3, 4, the probabilities are 0.42, 0.36, 0.14, 0.06 and 0.02. So, this adds up to 1, this is my $P(X = x_i)$, this is the probability distribution I have.

So, now would you be surprised at the following? What is the first thing everyone owns a credit card. So, if I take 500 people, my distribution says $P(X = 0) = 0.42$, in other words the way I can interpret this is in the sample of 500 people that the chance that people have no credit cards is pretty high.

So, if everyone owns a credit card then it would be unlikely because 42 percent of the adults do not own credit card according to my distribution. Hence, it is very unlikely that every 1 of the 500 would own credit card. So, this if somebody is claiming that everyone owns a credit card, I would look at it with suspicion because my distribution says that 42 percent do not own a credit card.

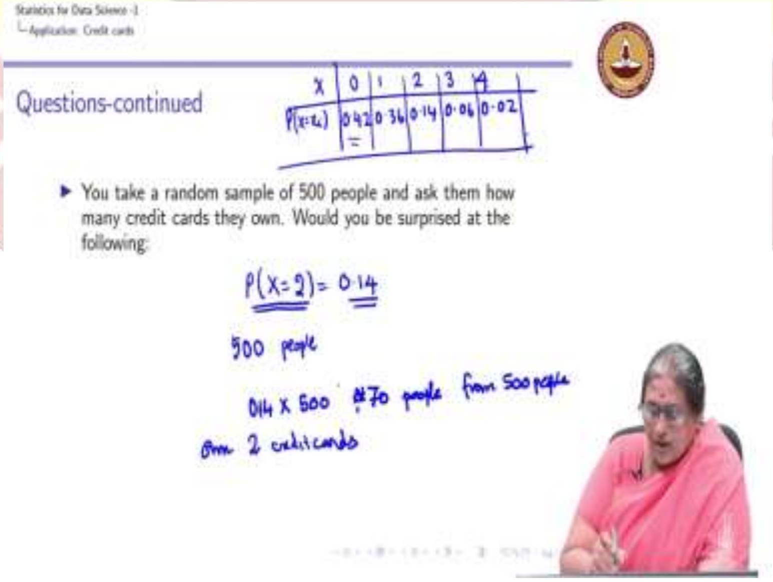
(Refer Slide Time: 12:06)



Statistics for Data Science -1
└ Application: Credit cards

Questions-continued

- You take a random sample of 500 people and ask them how many credit cards they own. Would you be surprised at the following:
  - Everyone owns a credit card.
    YES: 42% of adults do not own credit cards. Hence, it is unlikely that every one of the 500 would own credit cards.
  - 72 people respond that they own two credit cards.
    NO: 14% own two credit cards. 14% of 500 = 70, so it is likely that 72 people from sample of 500 own two credit cards.



Statistics for Data Science -1
└ Application: Credit cards

Questions-continued

x	0	1	2	3	4
$P(X=x_i)$	0.42	0.36	0.14	0.06	0.02

- You take a random sample of 500 people and ask them how many credit cards they own. Would you be surprised at the following:

$P(X=2) = 0.14$

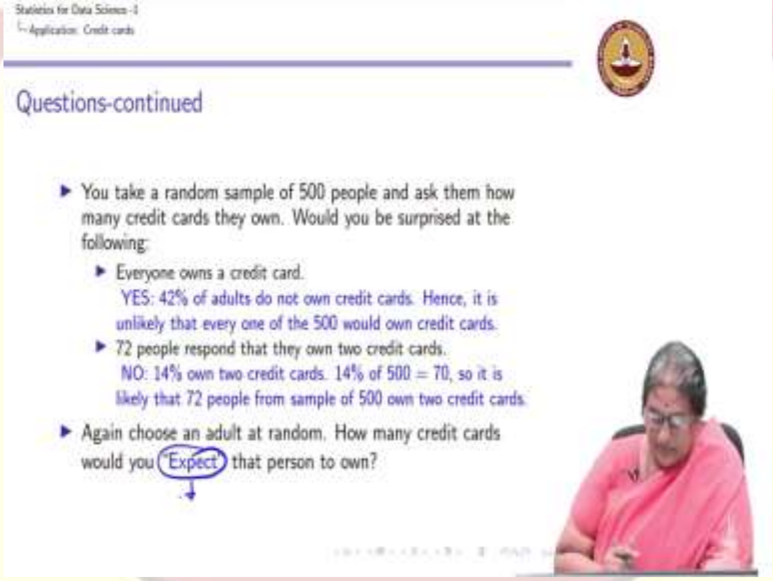
500 people

0.14 × 500 ≈ 70 people from 500 people own 2 creditcards

The second question is 72 people respond that they have 2 credit cards, so let us go back I can see that the $P(X = 2) = 0.14.$ I have 500 people, so if the chance of a person having 2 credit cards is 0.14, among 500 people I can expect 0.14 into 500 which is equal to 70 people from 500 people own 2 credit cards, that is what my probability distribution says.

The statement says that 72 respond so I would not be surprised because 14 percent of 500 is 70, so it is likely, the first case it was unlikely that everyone would not own a, would own credit card, whereas here it is very likely that 72 people from a sample of 500 own 2 credit cards.

(Refer Slide Time: 13:26)



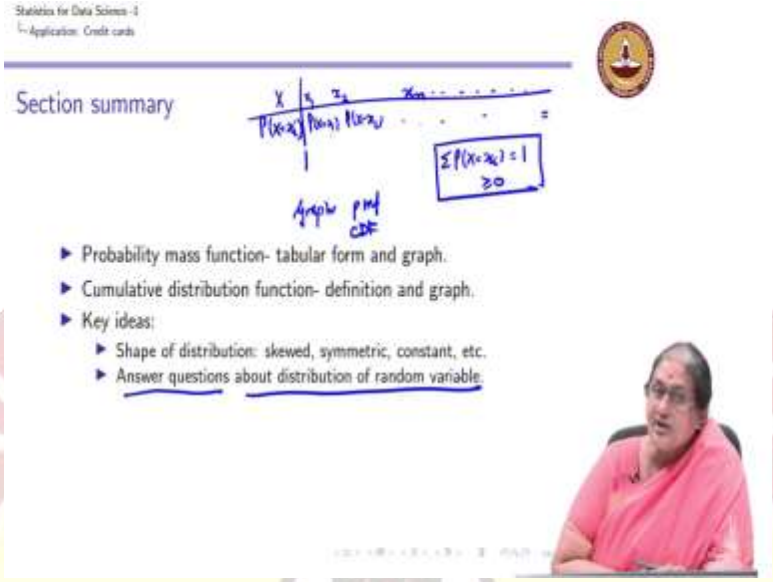
Statistics for Data Science -1
└Application: Credit cards

Questions-continued

- You take a random sample of 500 people and ask them how many credit cards they own. Would you be surprised at the following:
  - Everyone owns a credit card.
    YES: 42% of adults do not own credit cards. Hence, it is unlikely that every one of the 500 would own credit cards.
  - 72 people respond that they own two credit cards.
    NO: 14% own two credit cards. 14% of 500 = 70, so it is likely that 72 people from sample of 500 own two credit cards.
- Again choose an adult at random. How many credit cards would you Expect that person to own?

So, the next thing which we are going to see is, if I choose a adult at random, how many credit cards would I expect this person to have. So, if you see I am qualifying this word “expect”, we need to understand what we are meaning, what do we expect from this statement.

(Refer Slide Time: 13:53)



Statistics for Data Science -1
└ Application: Credit cards

Section summary

$$\begin{array}{c|cccc} X & x_1 & x_2 & \cdots & x_n \\ \hline P(X=x_1) & P(X=x_1) & P(X=x_2) & \cdots & \end{array}$$

$$\boxed{\begin{array}{c}\Sigma P(X=x_i)=1 \\ \geq 0\end{array}}$$

- Probability mass function- tabular form and graph.
- Cumulative distribution function- definition and graph.
- Key ideas:
  - Shape of distribution: skewed, symmetric, constant, etc.
  - Answer questions about distribution of random variable.

So, now we are going to introduce what is an expectation of a random variable. So, at this point of time you should know what is a random variable, we are looking at a discrete random variable, a discrete random variable can take finite number of countable values or infinite number of countable values with their respect to probabilities, $P(X = x_1), P(X = x_2)$ and so forth, we have defined the properties of a probability mass function, all the probabilities, the sum of the probability should add up to 1 and they should be greater or equal to 0, these were the 2 key properties.

And then we talked about the graph of a probability mass function, we introduce the notion of a cumulative distribution function and we looked at how do we answer questions about the distribution of a random variable through an example based on credit cards.