IIT Madras

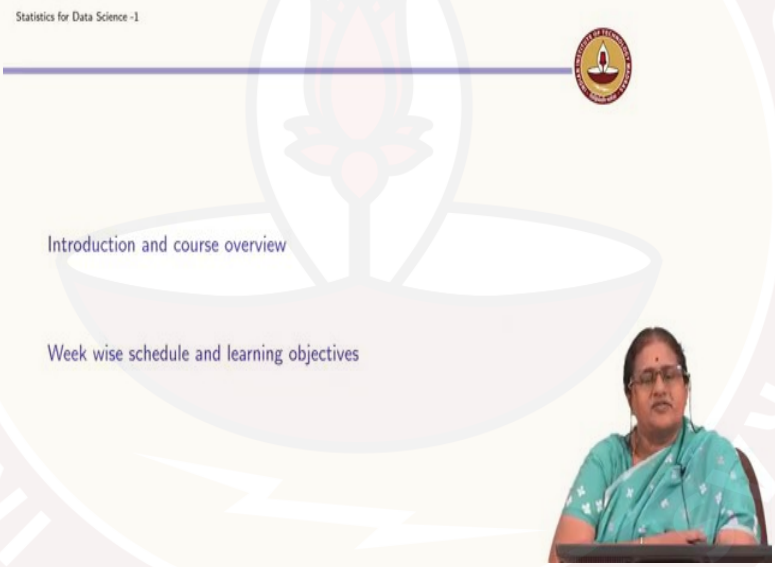
ONLINE DEGREE

Statistics for Data Science - 1
Prof. Usha Mohan
Department of Management Studies
Indian Institute of Technology, Madras

Lecture - 02
Introduction - Course Overview

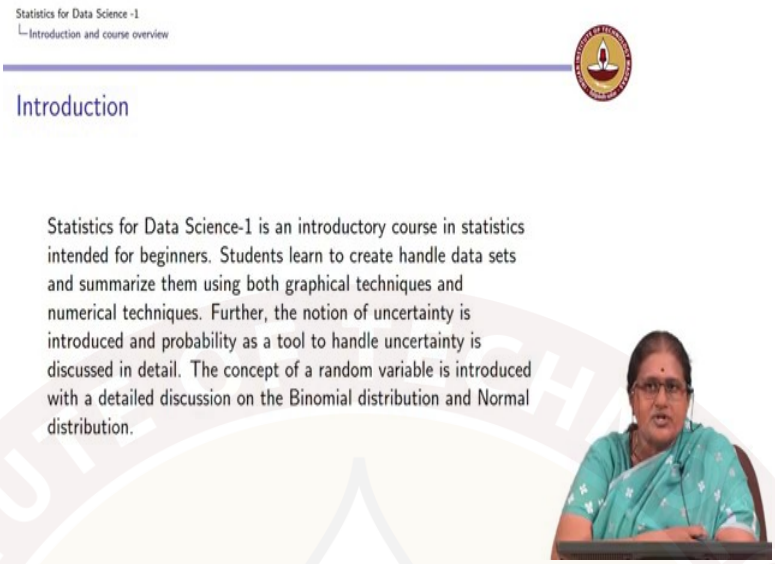
Hello. This is the Statistics for Data Science-1. This is the foundation course that you would require and this will give you the statistics. This is a preliminary course in statistics and in this lecture, we will be just going through and we will just see what the course overview is about; what are the learning objectives of this course and at the end of this course, what would you expect to learn.

(Refer Slide Time: 00:44)



In this lecture, we start with a brief introduction to this course and we also will discuss as to what are the week wise schedule and learning objectives of this course. Now, why this course? What is the main learning objective of this course?

(Refer Slide Time: 01:00)

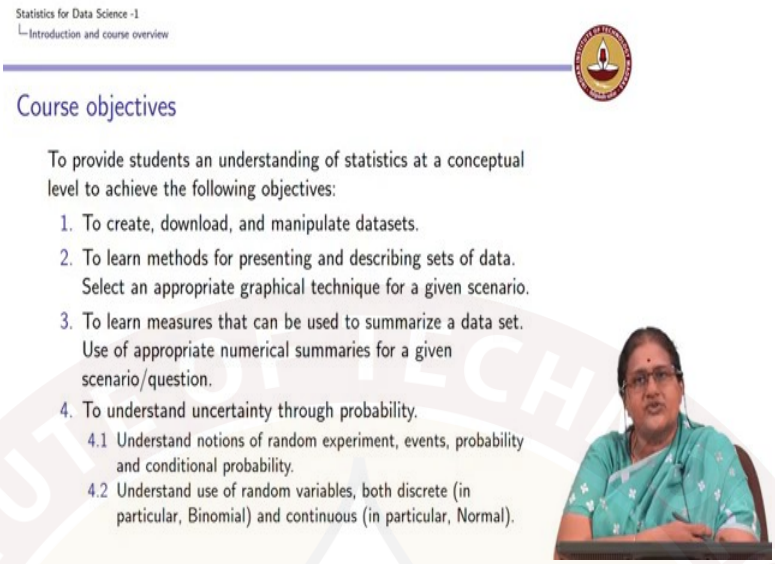


It is an introductory course. It is intended for beginners and by beginners, we mean by any person who has had tenth, class 10 level math. The, I think any person who has done math up to till; class 10 should be able to take this course comfortably. The main idea of this course is to help students learn to create data sets and summarize them and when we talk about summarizing them, we also talk about using both graphical techniques and numerical techniques.

The next important thing about this course is we are going to also introduce what we understand by the notion of uncertainty and the theory of probability not the mathematical theory. But we are going to use and discuss probability as a tool to handle this uncertainty. We will discuss about the using probability as a tool in some detail.

Finally, we introduce a notion or concept of a random variable and we focus on two important distributions; the binomial distribution and normal distribution and we will study applications about it. Throughout the course, we are going to focus only on applications and the understanding at a conceptual level. The focus is not going to be on the theory behind statistics and not behind theorems and proofs. The focus is going to be at an application level.

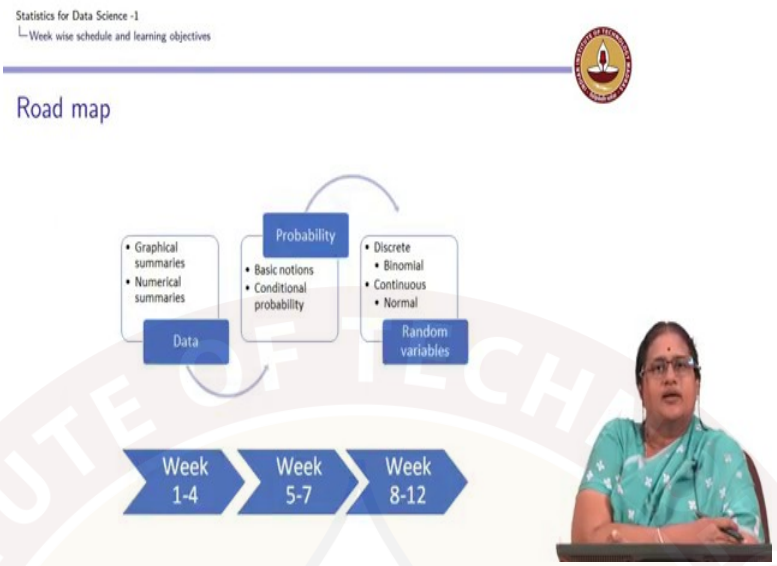
(Refer Slide Time: 02:37)



So, what are the objectives of this course? The objective of this course is to provide students with an understanding at a conceptual level. At the end of the course, students should understand how to create download and manipulate data sets. This is one of the first objectives. The student should learn methods of presenting and describing data. By presenting and describing data, we expect the student would be able to understand what is an appropriate graphical technique; given a scenario, we talk about both graphical summaries of data and numerical summaries of data and very quickly, we will realize that numerical summaries cannot be given for all kinds of data. So, we would also focus on numerical summaries given a scenario and question.

So, all through the course, the focuses also going to be how students are going to formulate questions given data. We understand uncertainty through probability and as a course of understanding uncertainty, we are going to understand notions of what is a random experiment and understand use of random variables. So, this is at the end of this course, a student should be familiar about summarizing data, any kind of data and then afterwards have a good conceptual level of understanding of the basics of probability.

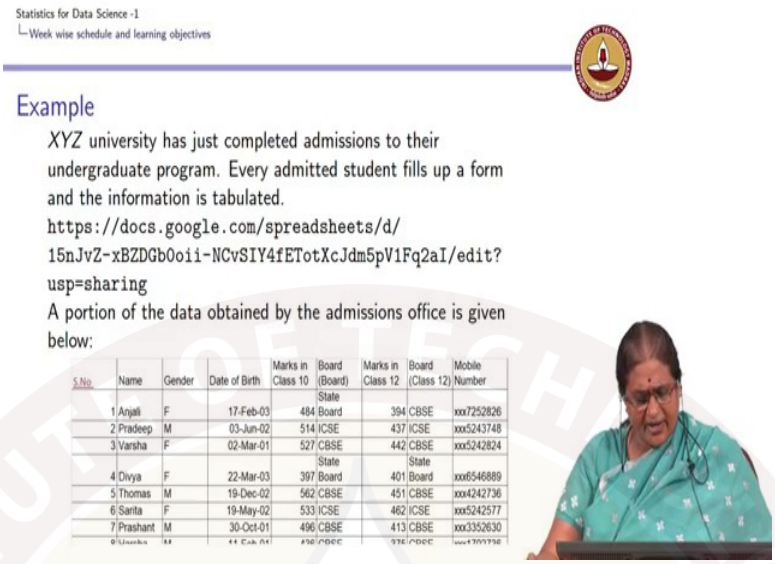
(Refer Slide Time: 04:07)



This course has been planned to span over 12 weeks and the roadmap for this course is given as in the following diagram; graph. So, if you modularize the course, we have 3 modules. In the first module, we are going to understand the basics of data; what are the types of data; how do we summarize data and we expect to achieve this in 4 weeks.

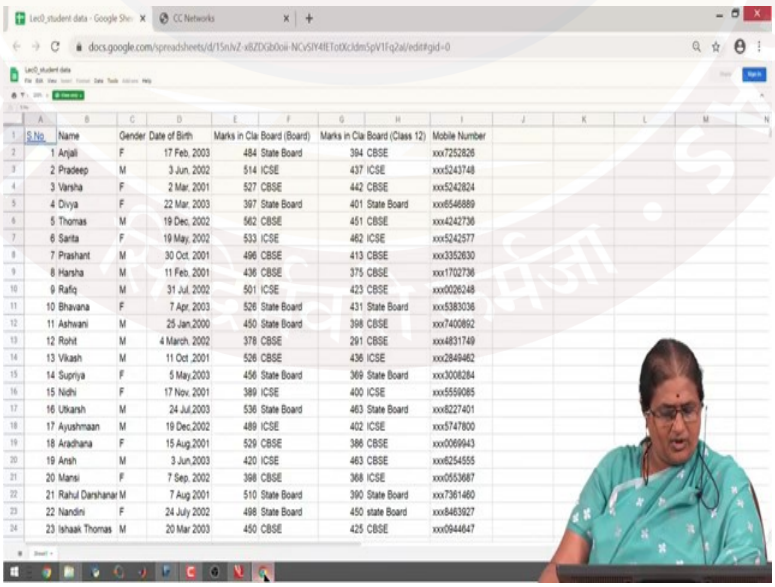
Once we know about data in this first 4 weeks, we are not dealing with any kind of uncertainty. But once we finish week 4, we move on to the concepts of probability, we learn basic notions of probability, we spend about 3 weeks to understand what are applications of probability and then, afterwards we move on to the notion of what we call random variables, wherein we talk about discrete random variables and continuous random variable with specific focus on the binomial and normal distribution.

(Refer Slide Time: 05:17)



So, what we are going to do now is going to start with what to expect from the course and give a week by week expectation, set the expectations for the course. Let us look at a very simple example, where I have a university let me call it a XYZ university which has just completed the admissions to their undergraduate program. Every admitted student has asked to fill up a form and the following information is tabulated.

(Refer Slide Time: 05:46)



So, the minute I say information is tabulated, you can see that. So, if you look at the data, this is just some hypothetical data we have created. Wherein, you can see that what the

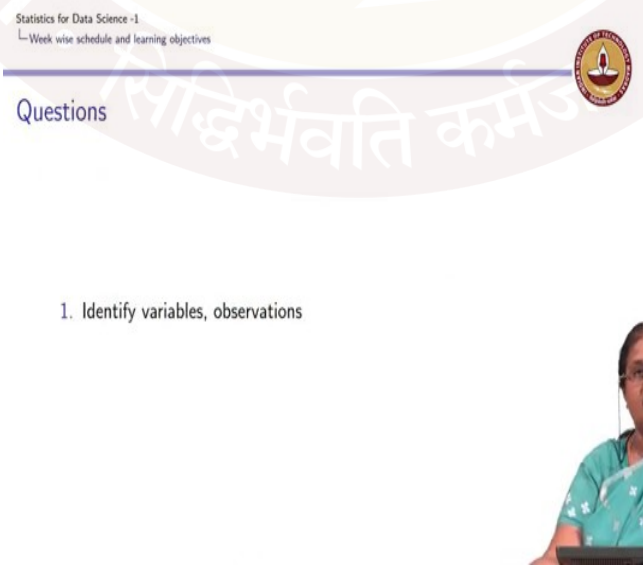
information that was captured in the application form was the name of a person, the gender, date of birth, marks obtained in class 10, the board from which the student passed or wrote the exams; for class 10, marks obtained in class 12 and their board and their mobile number.

At the first look, you see that this is any all of us are familiar with data sets of this kind or data of this kind. So, what is the information that we normally would seek from this data set? So, we what we are going to do is we are just going to start asking a few questions based on this data set. It is a very simple data set and we are trying to see that what are the type of answers that we can hope to answer during the course. So, given this data set, what we are going to do here? We are going to actually start you can see that this I just pasted a portion of the data set.

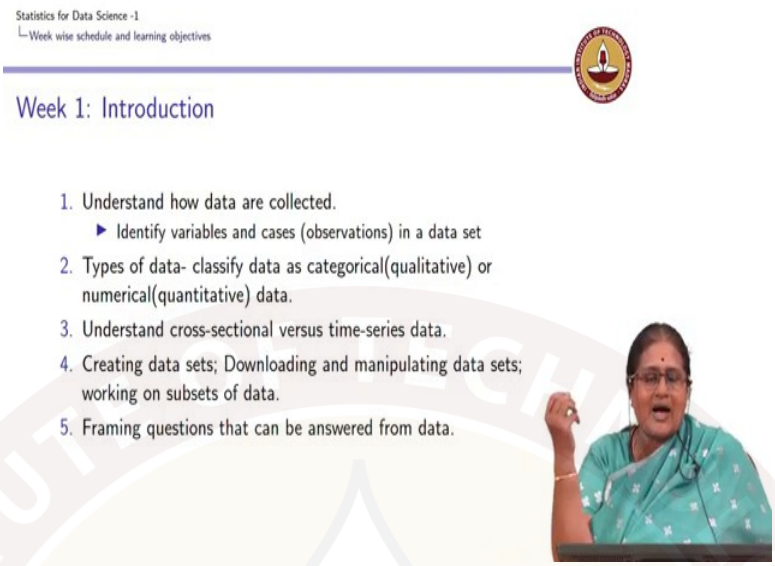
So, perhaps the first thing which we would want to see from this data set is to try and see that you already see that you have name, you have gender and you also see that when you are looking at the data set, you have some like marks in class 10 and marks in class 12 are captured as numbers. I do have mobile number also which is captured as numbers.

But when I talk about the board, I have both the state board and the CBSE, I have ICSE and I have date of birth is also given to me. So, the first thing which we notice when we see a data set of this kind is, we already know that I do not have one kind of variable. So, we need to understand what is the thing that is varying here.

(Refer Slide Time: 07:49)



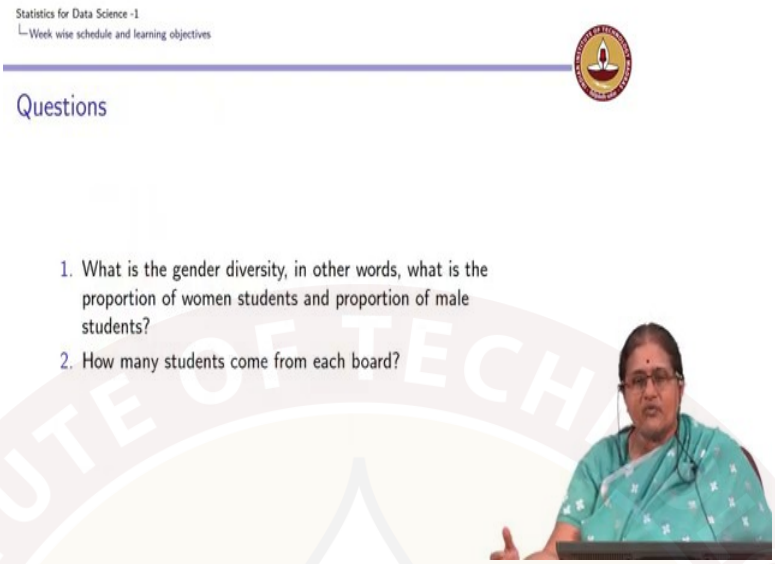
(Refer Slide Time: 07:53)



So, in week 1, we would introduce students to data sets of various kinds. So, the first objective in week 1 is to understand how data is collected. Once we understand how data is collected, we go back to understand what we mean by variable and observations. This is the first step in understanding data. So, before even jumping into doing the mechanics of statistics, we need to understand our data very well. When we need to understand our data very well, we need to understand what are the type of variables, what is the type of data we have collected and what are the different classifications that are available. The two major classifications which are available are we can classify data as quantitative and qualitative or numerical and categorical data.

So, the first thing, we are going to focus on is how do we create data sets or how do we download existing data sets. We might not be wanting to work with entire data set, we might be just wanting to work with a subset of data. So, how do we create the subsets of data? But the focus during this week 1 of our introduction would be on what are the questions that we are going to frame or what are the questions or what are the answers, we seek from data. Can we answer all these questions from data that we need to find out? But at least we want to train you in answering or in framing the right questions that you seek from data.

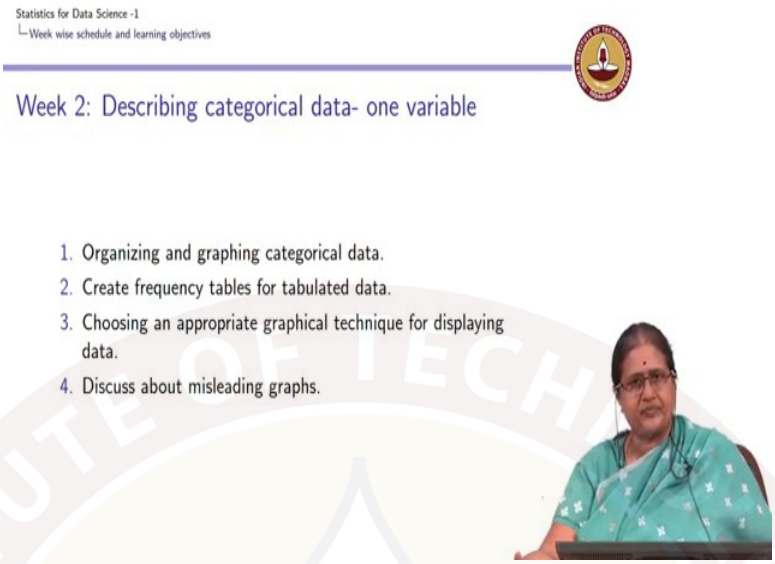
(Refer Slide Time: 09:39)



As an example, you can see that what we might want to ask from the data set that I presented to you earlier is we would want to know what is the proportion of women students to male students. This is a question which anybody would want to know or you would want to know what is the proportion of people who have come from CBSE? What is the number of people who have come from CBSE? What is the distribution of people who have come from ICSE or different boards? And if state of a student has been captured, we would like to know what is a regional representation of students?

So, there are many questions which you would want to know which are either we would like to know what is the proportion of people who are from a particular region or what is the count of people who are from a particular region. So, this is the type of questions pertaining to us, type of data which we refer to as categorical data.

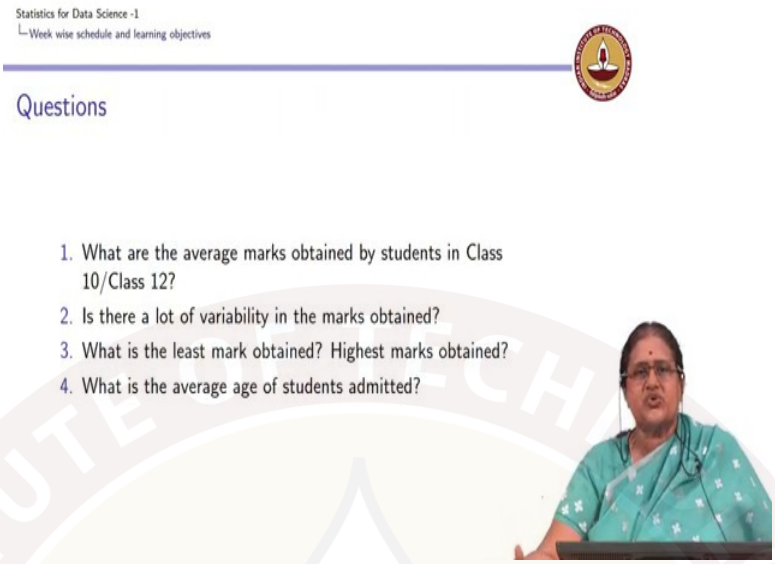
(Refer Slide Time: 10:27)



So, the first-second week, we are going to spend time in understanding what is categorical data; what are the questions we want to find out from categorical data; what are the kind of answers, we seek from categorical data. And during this time, we are going to focus on how do we organize graph categorical data by creating frequency tables and the most important thing is though there are graphical techniques available to us, we are going to focus on which is an appropriate graphical technique to answer a particular question. This is the focus which we want to give is you frame a question and identify what is the right technique that you need to answer a question.

Many a time, we find that there are a lot of graphs which could be misleading. So, we would also focus on and discuss a bit of how important it is to correctly describe or correctly convey data through graphical summaries. Also, said that when you can broadly classify data into categorical data or qualitative data and numerical data. Once we understand how to summarize categorical data using graphical summaries, we move on to see what are numerical data.

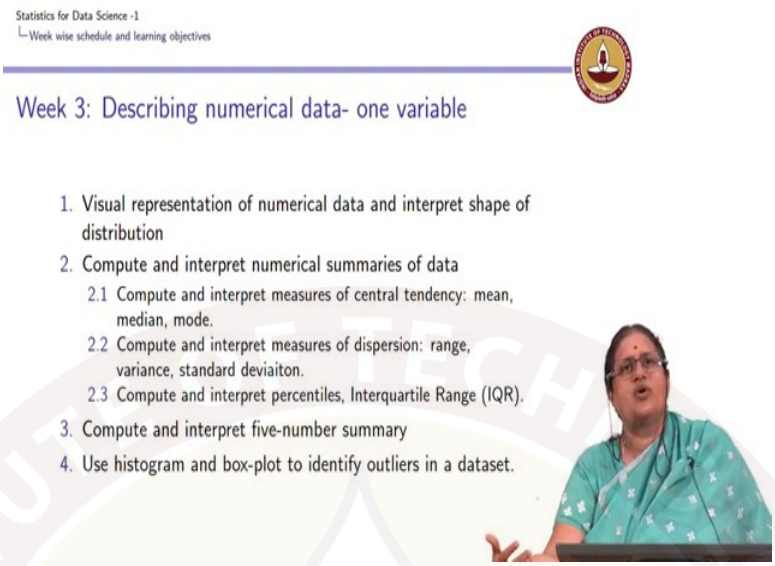
(Refer Slide Time: 11:48)



And what are the type of questions you have? The minute I say it is numerical data, we understand that we can do some arithmetic operation on it and get some mathematical summaries on it. One of the most often used summary is called the average or the mean. We are going to ask and what are the questions that, we hope to answer at this point of time. Again going back to a student data, one might be interested in knowing what was been the average marks obtained by the students in either their class 11, class 10 or class 12.

Another question people might want to know is has there a lot of variability in the marks that has been obtained or are people obtaining marks which are very close to each other. What is the least mark? What is the highest mark? What is the average age of students? So, you can see that these are very natural questions that many of us have been wanting to ask, when we are presented with a data set of that kind. In other words, the questions we seek to answer here involves some mathematical summary or a numerical summary.

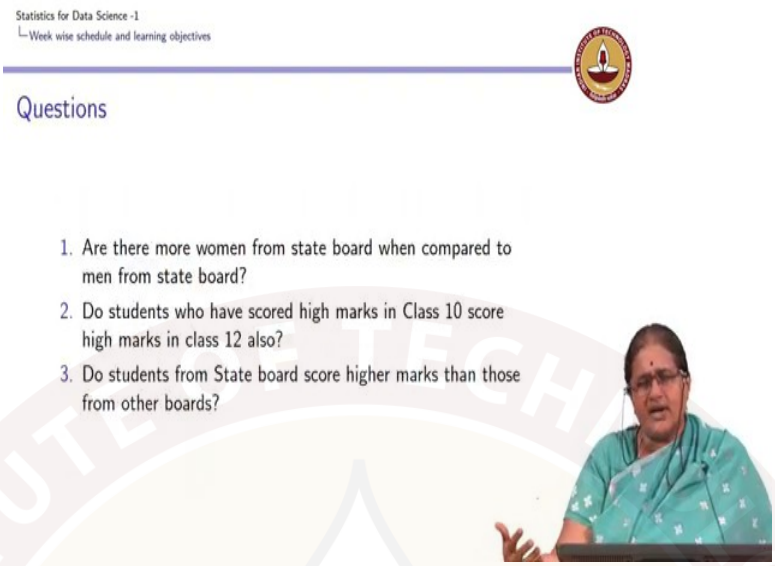
(Refer Slide Time: 12:54)



So, in week 3, we are going to go and focus on how you describe numerical data using numerical summaries. While we focus on numerical summaries at this point of time, we also discuss graphical summaries of continuous data; but the focus during this week is going to be pretty much on how you compute the well-known measures or how do you summarize this data using numerical measures. Broadly, measures of central tendency and variation.

So, we are going to focus on what are the main measures of central tendency and variation and we also will relate this to certain graphical things. Mainly, we are going to relate it to how you are going to have, how do you construct histograms and box plots. So, by the end of third week, you would know how to summarize a categorical data and numerical data; but till this point of time, we have been focused only on summarizing one variable.

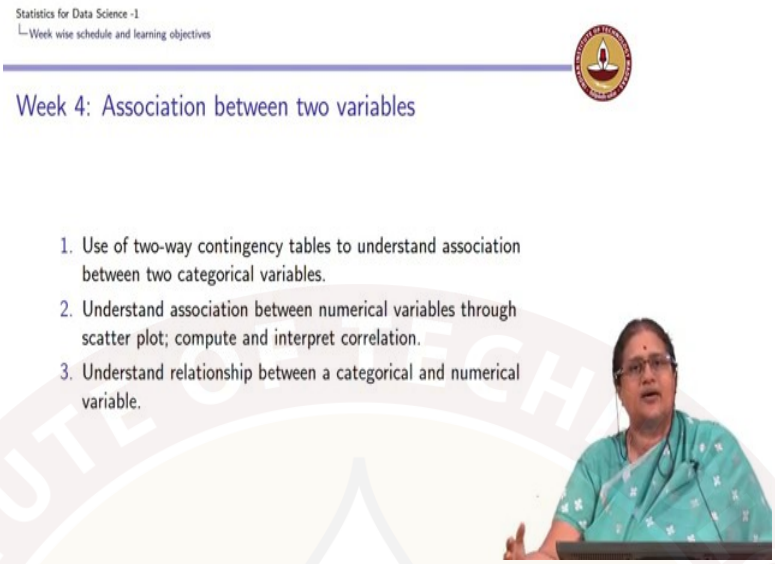
(Refer Slide Time: 14:06)



So, the next thing, what we would want to understand is given again go back to your school data set. In your school data set, you might want to ask questions like are women from state board, how do they; are there more women from state board compared to men from state board? Do people who have scored high marks in class 10? Do they also score high marks in class 12? So, we are asking questions now, where we are trying to understand whether two variables are related to each other or associated with each other.

I want to tell here a word of caution, we are not asking anything whether a causes b or b causes a. We are just interested in answering the question whether a student who obtained marks in class 10, does equally well in class 12. This is a very natural question to ask or do people from state boards score higher than marks than people from other boards. So, in a sense, we are looking at whether I can come up with measures or summaries which can actually summarize the relationship between variables. So, from moving from summarizing one variable, we move on to understand how do we capture association between variables.

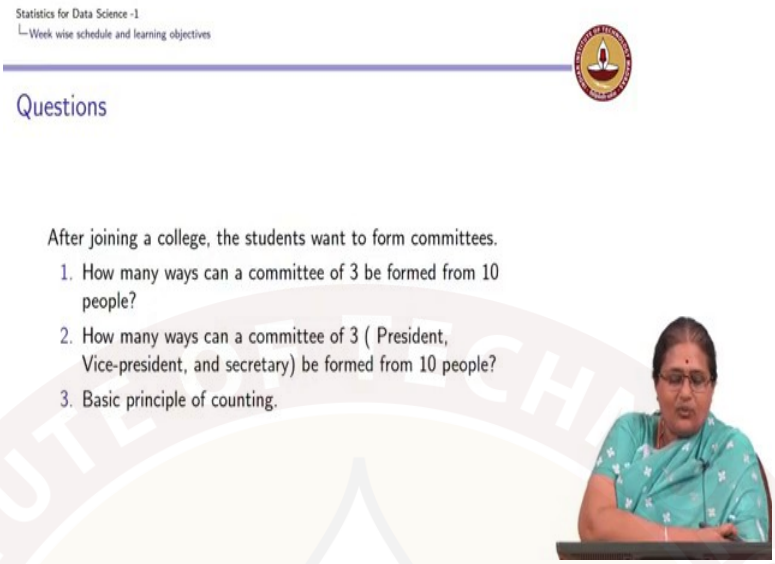
(Refer Slide Time: 15:25)



That is going to be the focus and the questions, we are going to ask in week 4 is to deal about association between two variables through what we call contingency tables, a graphical method which is called scatter plot, where we talk about numerical variables and we also understand how a categorical and numerical variable are related to each other. So, this is about the first module which is the week 1 to 4, where you would have had a reasonable and you should have a conceptual level understanding of what is a data set; what are the types of variables in my data set; how can I categorize them; how can I classify them as a quantitative, as a qualitative or as numerical or categorical and how do I summarize them.

When I talk about summaries, it is very important that we again, we ask a question is what is it I am seeking and what is the appropriate measure; be it a graphical summary or a numerical summary. I think the focus here is to be very clear about what is the question you are seeking to answer like the question you are seeking to answer, we are asking that what is the information you are seeking from the data set and how are you going to achieve it. This is the first module and at this end of the module this is where you should be. So, this is what you would be expected to know at the end of the fourth week.

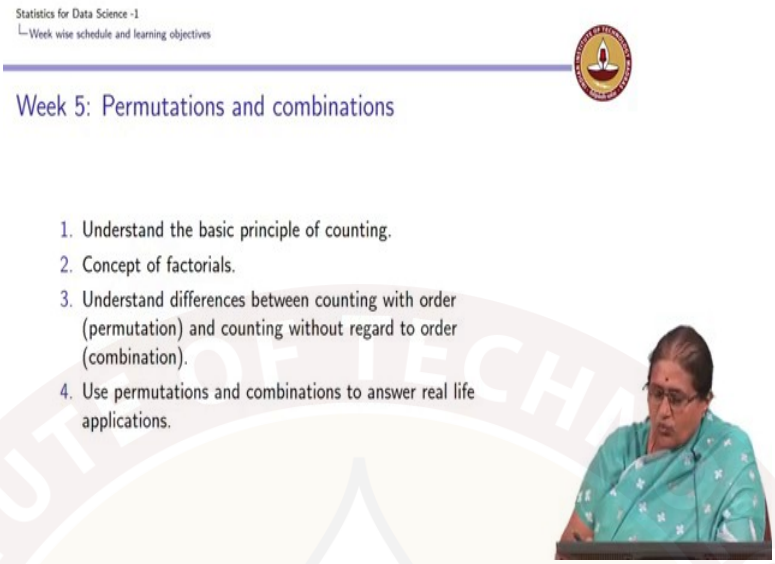
(Refer Slide Time: 17:01)



Now, the next module is the key module, where we are going to introduce the notion of probability. Now, why do we need to understand probability? Probability is extremely important because we live in uncertain times and what we want to know is we always ask questions, where there is an element of chance that is involved. Whenever there is an element of chance or whenever there is uncertainty, we need a very robust tool to handle this notion of uncertainty and probability is a good tool to handle this uncertainty.

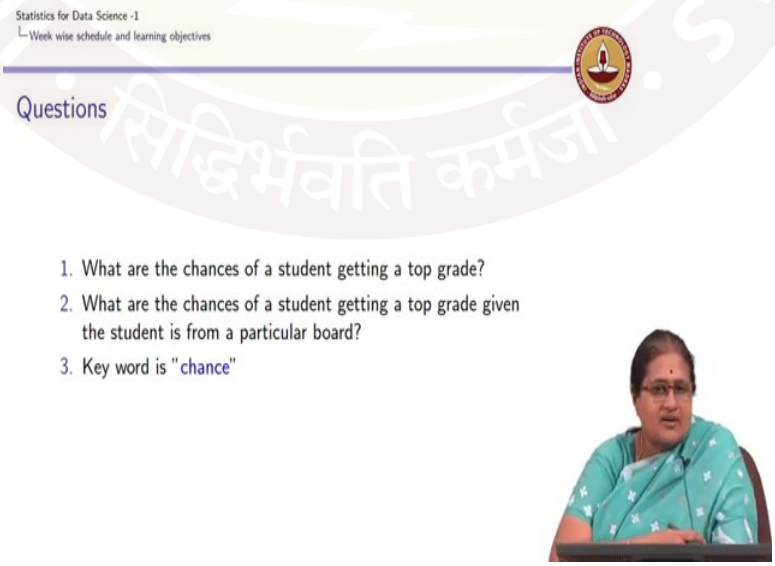
But even before we understand what is the tool of probability, we need to understand something. We need to understand the basic principle of counting. Why do we need to understand the basic principle of counting? For example, after joining a student people; students typically after joining a college would want to form committees. The most natural questions you want to ask is how many ways a committee can be formed? How many ways can a committee of 3 can be formed? The difference between the first two questions is in first question, I was just interested in knowing the number of ways a committee of 3 can be formed; whereas, in the second question, I am interested in an order; I am interested in the President, Vice-president and secretary. Now, many of us would have been already introduced to this concept in high school which is famously known as permutations and combinations.

(Refer Slide Time: 18:33)



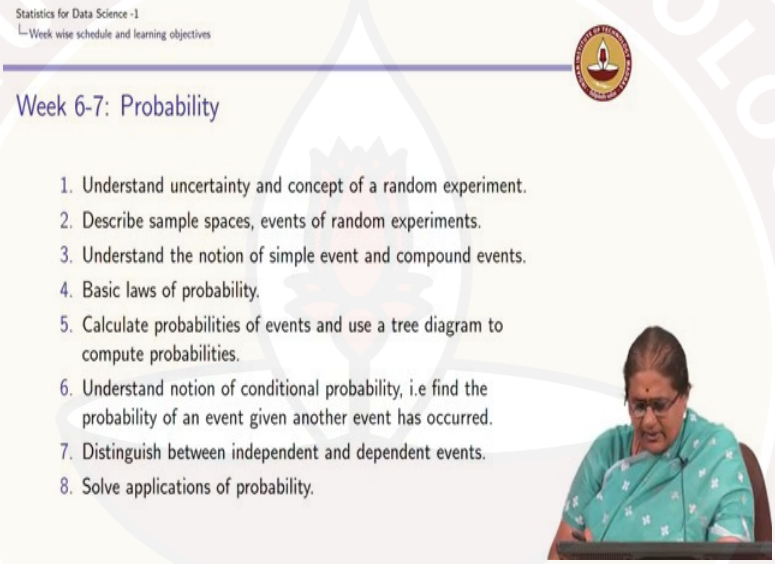
So, what we will do in week 5 is introduce a student to the basic principles of counting and we will understand how to apply the notion of permutation and combinations which is basically counting with order and counting without regard to order and the main focus of this basic principle of counting is to help a student understand how to use these permutations and combinations to answer real life applications. At the end of week 5, once you understand how this basic principle of counting is applied; then afterwards , we move on to said what is the; what are the questions we are really asking here.

(Refer Slide Time: 19:12)



So, when there is lot of uncertainty, a student has just joined college based on the marks, they have obtained in the 10 and 12th classes. So, the immediate thing, they would like to know is what is the chance of me getting a top grade? Now, the other questions you might want to ask is and from an administration, what is the chance of a student being a topper given or conditioned on the fact that they come from a particular region or they come from a they belong to a particular gender or they actually there lot of questions that you would want to ask, where you are actually asked the key word is chance; what is the chance.

(Refer Slide Time: 20:05)



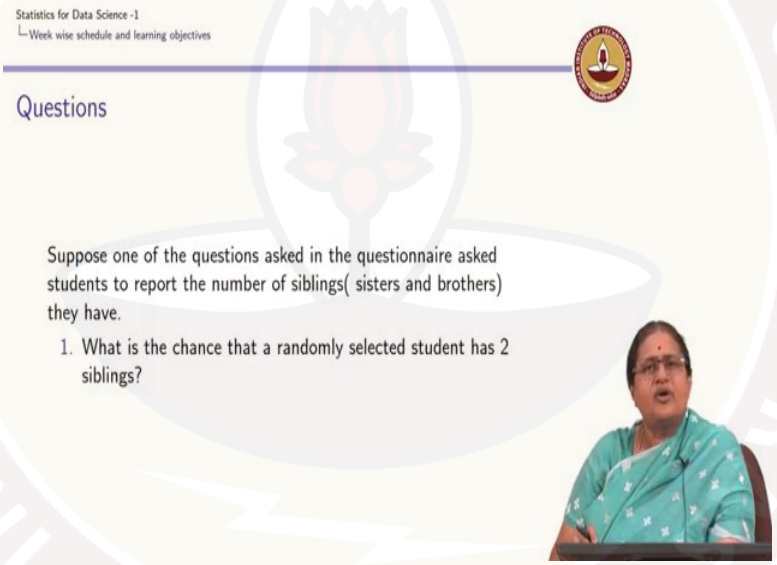
So, we are used to this notion of a chance and this notion of a chance is basically captured and what we are going to introduce to the student to in the 2 weeks that follow is the basic notion of uncertainty. All of us know, you toss a coin, you know that there would be a head or a tail. But you really do not know whether the outcome is actually going to be a head or a tail. So, there is some uncertainty associated with this. We introduce the notion of what is randomness here and throughout these 2 weeks, what we are going to focus is we are going to focus on understanding what is a simple event or compound event.

And at this point of time, you should have learnt about sets in your math-1 course. Because probability and understanding the notion of simple events and compound events would need that we start representing events as sets, representation of events or sets

requires some idea of set algebra and at this point of time, you would have already had an introduction to set operations and set algebra from your math-1 course. So, you will be applying those concepts here to develop the notions of probability.

So, at the end of week 7, you should know the concepts or what are mutually exclusive events? What do I mean by independent events, that is does the fact that I get a head in the - I am tossing a coin twice - getting a head in the first toss, does that affect my outcome of the second toss or are they independent of each other? So, these are the notions which we are going to help understand, we are going to help develop the probability framework to answer these questions. So, this is what you are expected to know at the end of the 7th week.

(Refer Slide Time: 21:58)

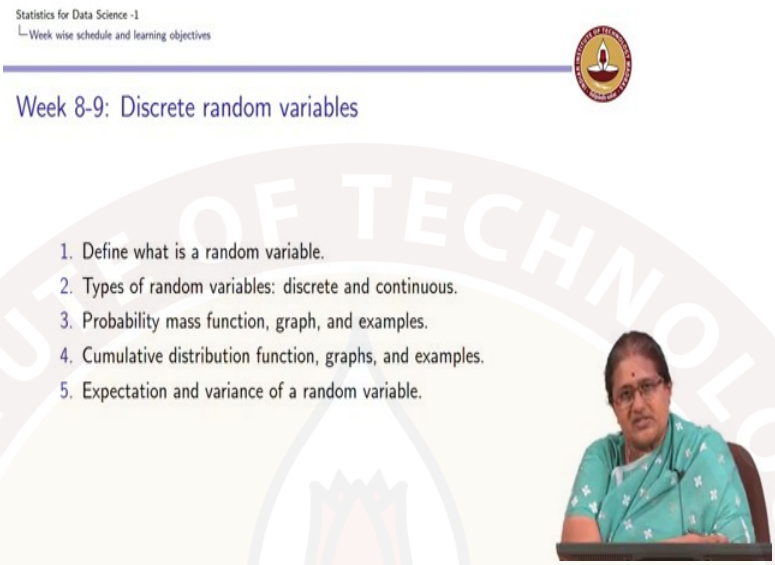


Now, till this time, we have been always focused on events. I said we are going to talk about events as a set and we also know that when you talk about a set, it need not mean that you always have only numbers or which are elements of the set. But at some point of time, we need to ask questions. For example, in the same questionnaire which we refer to - suppose a student has been also asked to record or give the number of siblings; sisters or brothers they have and you are at by chance you are selecting some student.

A question you might want to answer is what is the chance that a randomly selected student from my database has 2 siblings? I am just restricting it to 2, but it could be 1, it

could be 0. In other words, I am associating a numerical value with whatever I want to achieve and this is what I am going to do through the concept of a random variable.

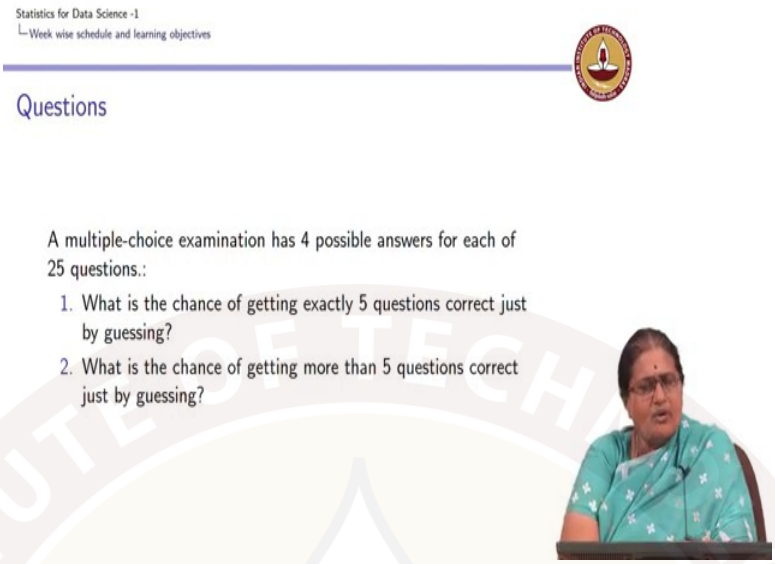
(Refer Slide Time: 23:06)



So, we introduced the notion of a random variable at this point and we start with notion of discrete random variable, where the random variable takes discrete values numerical values and at this point of time, at the end of the random variable once we introduce what is a random variable, we would like to since it takes values describing some sort of a summary with this random variable would make sense. For example, you might want to know what are the on an average how many students have siblings, 2 siblings or what is the average number of siblings people have from the university.

So, these are the questions which you would want to answer. So, we introduce notion of expectation and variance towards answering these questions. Once we introduce a student to discrete random variables, we focus on a very important distribution; we spend 1 week to try and look at the questions for example.

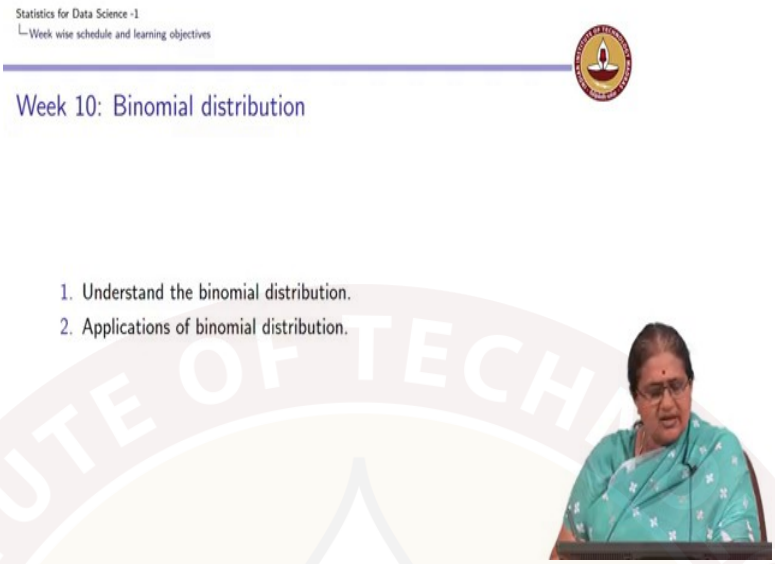
(Refer Slide Time: 24:05)



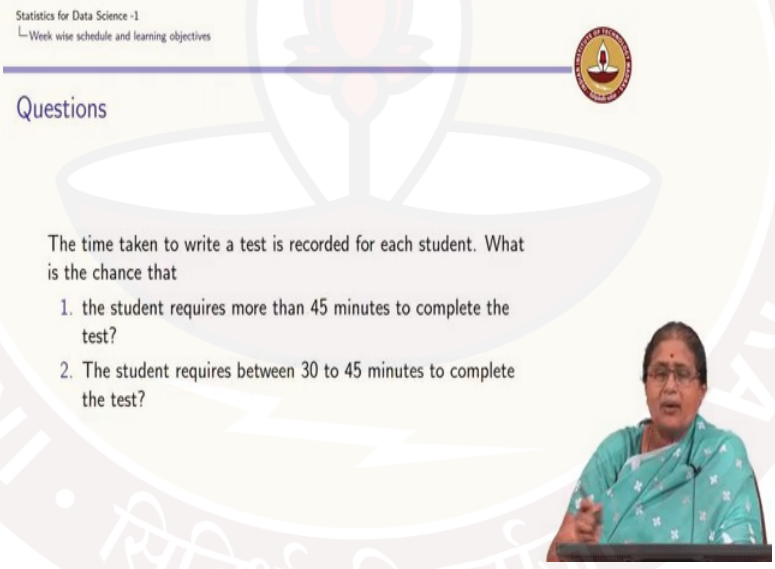
We might want to know your all students attempt and I am sure that all of us are at some point of time, we have taken examinations which has multiple choice questions. And if you are certain, if you know the answer there is no uncertainty there; but many of us guess answers in a multiple choice question. So, the natural question to ask when you are writing a multiple examination which has multiple choice questions is what is the chance of me getting questions correct just by guessing. This is a very important question. This is a very natural question for us to know.

Again, guessing is an element of chance. So, here we are focusing on getting an answer right or wrong. So, experiments of this kind constitute what is called a Bernoulli experiment and then, after wards the distribution that helps us answer questions which are very similar to the type of questions which we are posing now, from what are called Binomial distribution.

(Refer Slide Time: 25:10)



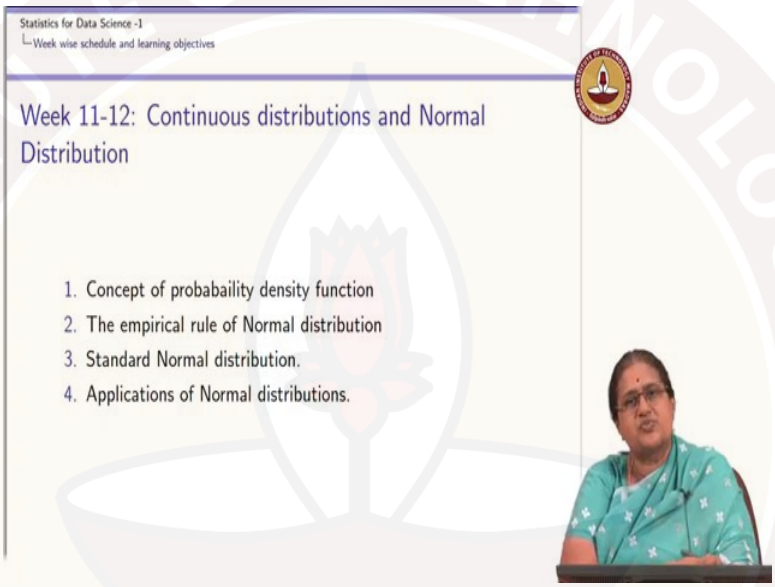
(Refer Slide Time: 25:16)



So, we are going to spend week 10 in understanding Binomial distribution and the focus again here is going to be on applications of binomial distribution. Now, till week 10, we have focused on random variables which take discrete values. But many a time, we are interested in answering questions, for example, if I am recording the time taken by a student to write an examination. The questions we are interested in knowing is what is the chance that the student requests more than 45 minutes to complete the test or what is the chance that the student would require between 30 and 45 minutes to complete the test?

So, the minute again you see that the question, we are answer asking is about the chance and the variable of interest here is time. So, we need a way to capture this variable of interest which is time and we notice that this time can take anything between 0 minute, 1 minute, one and half minutes, one and three-fourth minute. So, it is in a sense, it is a continuous variable. So, we focus on addressing this variable, these are called continuous variables; the type of questions we are going to ask on. So, first we will identify what are random variables that are continuous in nature.

(Refer Slide Time: 26:36)



And in the last week, we are going to focus on the last week 11 and 12; we are going to focus on variables that can take that are actually continuous in nature. Basically, we introduce the notion of probability density function; we are going to focus predominantly on a very very important distribution that arises again and again in our study of statistics which we refer to as the Normal distribution. We will focus on what is called the empirical rule of the normal distribution and again, the focus is going to be on applications of the normal distribution.

So, this has been the roadmap from week 1 to week 12. So, at the end of week 12, the student is expected to first differentiate between understand data, manipulate data sets, identify the types or classify the types of variables. To classify the type of variable, the student has to first know what is a variable; what is an observation and once you know that, how do I summarize these variables? To know how to summarize variables, a

student is expected again to ask questions, what are the why do I need to summarize a variable; what is the purpose to summarize a variable; what are the questions I need to answer; what is a questions I am seeking out of this data set?

What are the appropriate measures of summary; what are the appropriate summaries of the variable? Can I talk about association between variables? In the event of uncertainty, how do I handle uncertainty? We live in uncertain times, what is the basic notion of probability? What are the notion of random variables and what are the applications of these random variables? So, this is the course overview. This is what is expected. These are the learning objectives and at the end of the week 12, a student should be comfortable with the conceptual level understanding of whatever is presented so far.

Thank you.