INDIAN INSTITUTE OF TECHNOLOGY, KANPUR

NPTEL

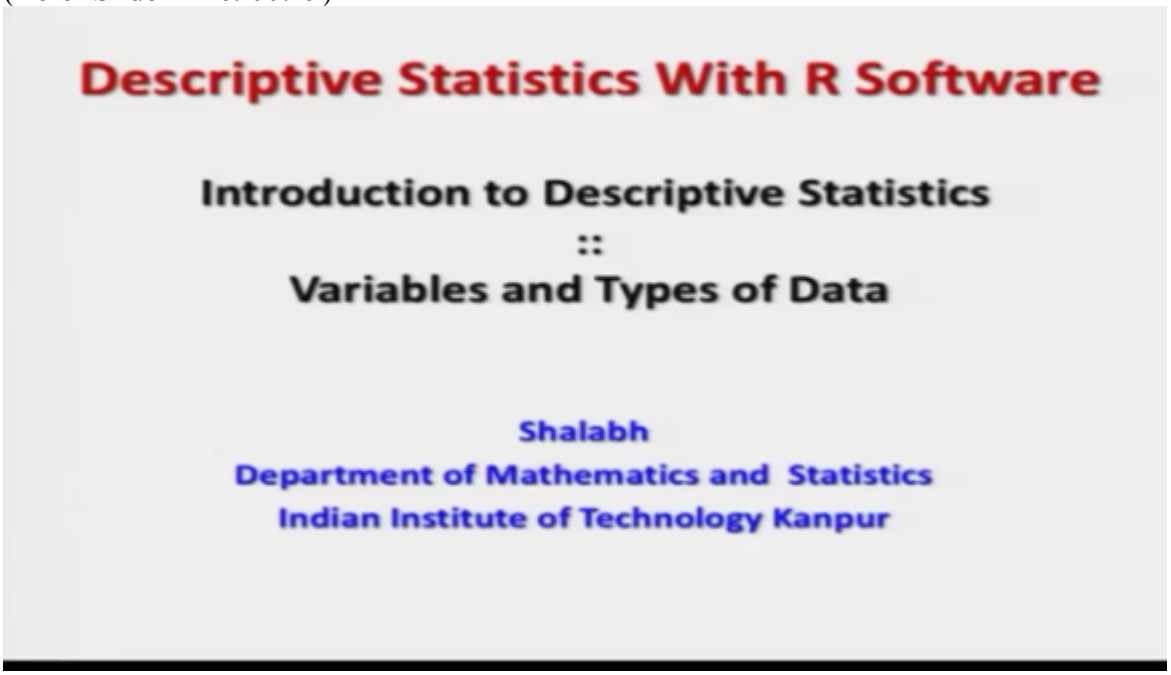
NATIONAL PROGRAMME ON TECHNOLOGY ENHANCED LEARNING

COURSE TITLE
DESCRIPTIVE STATISTICS WITH R SOFTWARE

LECTURE-07
Introduction to Descriptive Statistics – Variables and Types of Data

BY
PROF. SHALABH
DEPARTMENT OF MATHEMATICS AND STATISTICS
IIT KANPUR

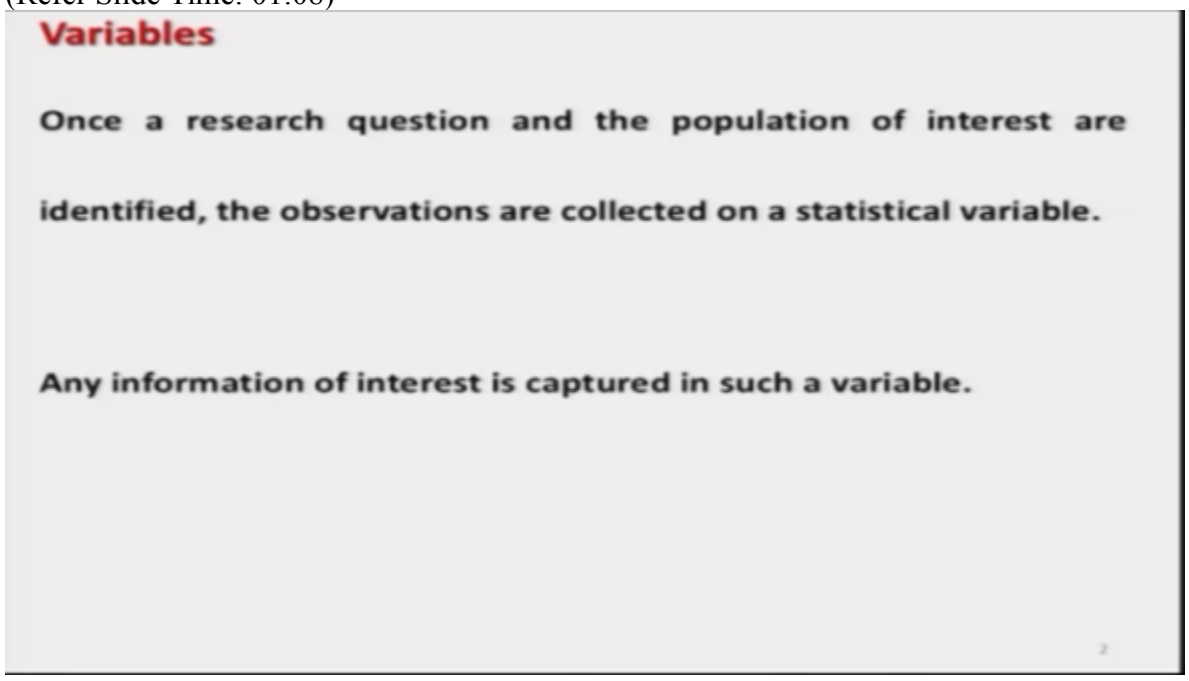
Welcome to the lecture on the course Descriptive Statistics with R Software.
(Refer Slide Time: 00:19)



You may recall that in the last lecture we had a small discussion on different aspect of descriptive statistics.

Now from this lecture I will be moving towards more mathematics, well in this lecture this is a very small amount of mathematics but my idea in this lecture is to make you understand what are the different types of terminologies and how they are represented, once we understand that what is the nomenclature and how the things are being represented that will help us in better understanding in the topics in the further lectures.

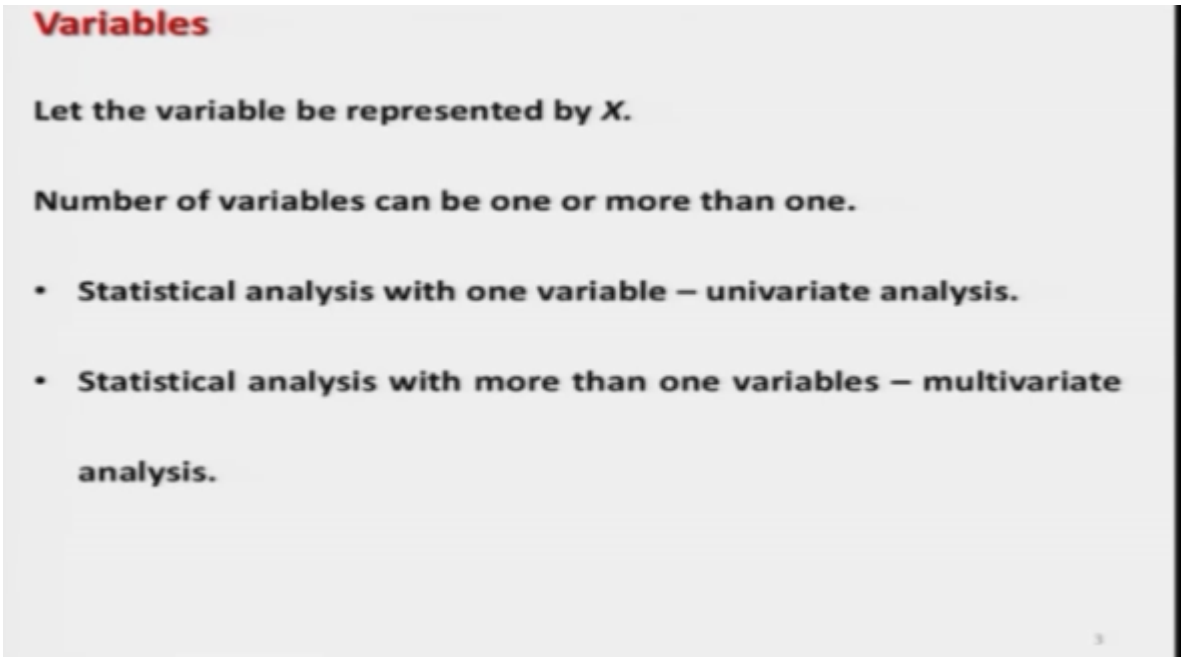
(Refer Slide Time: 01:08)



Now let us try to start our discussion with the first topic, what is a variable? Whenever we are trying to conduct an statistical analysis, before that there is a collection of data, and even before that there is always an objective and the objective is based on the research problem, or in simple words what we really want to know, once I decide this question what we really want to know based on that I tried to collect that data on the relevant variable from a population, and then I tried to collect that data on that aspect which I want to know and this aspect in simple language is a variable, so I can briefly say that once a research question is fixed and the population of interest is identified, then we try to collect the data on something, data on what? We tried to collect the data on a statistical variable, what does this mean? That whatever is my objective based on that I will try to collect the data on a relevant question which is going to be interpreted by that variable.

Before I go further I must tell you that there is a strong mathematical definition of these variables, random variables what we are going to discuss in the further slides, but here my objective is not to go to that mathematical level, my modest objective is that for our beginner how the things have to be understood, for example I will be dealing with the definitions of continuous random variable discrete, random variables, and if you come on the area of measure theory in the statistics there is a hardcore mathematical definition of these concepts, but definitely my idea here is to give you a flavor or make you understand what are these things and how they are going to be used in the collection of the data, so please keep this thing in mind once you try to understand these topics, right.

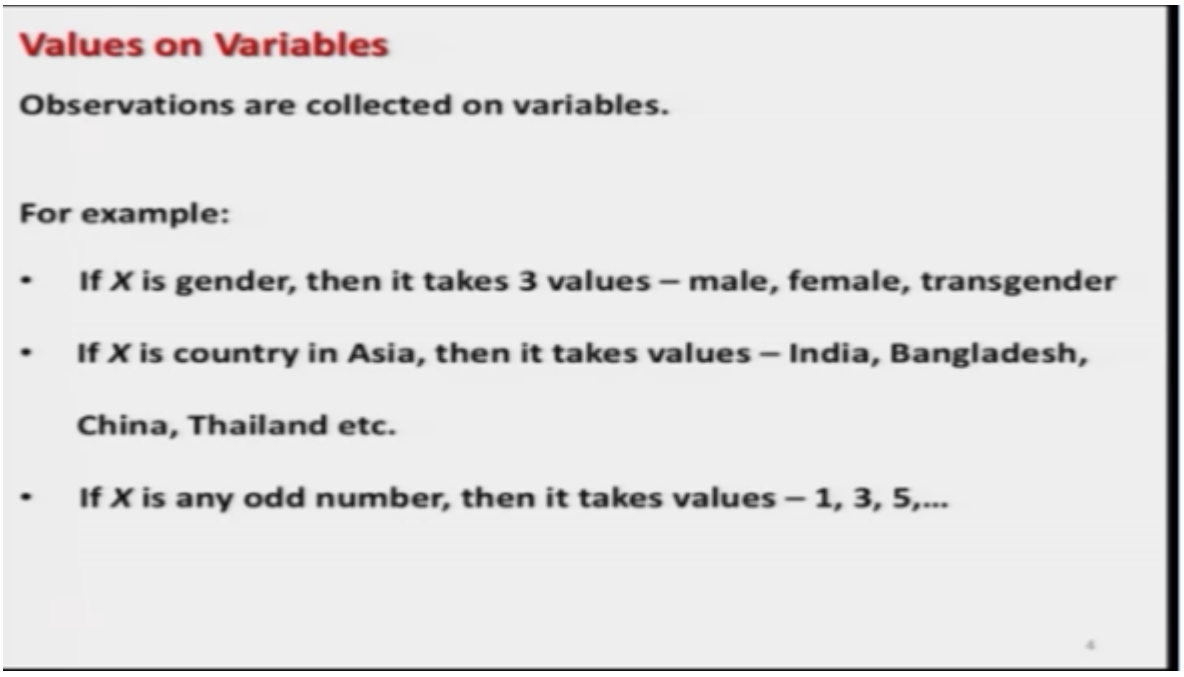
So I can see here whatever is the information in which we are interested that is captured inside that, or through a variable,
(Refer Slide Time: 00:47)



now in statistics there is a convention that these random variables are always represented by say this is your capital letter, and in our case usually I will be denoting the random variables by X, Y, Z etcetera, and when I try to type it then usually they are typed in a mathematical mode which is an sort of a italics mode, right, so this is what you have to keep in mind that whenever you try to see a capital letters usually that is indicating the variable.

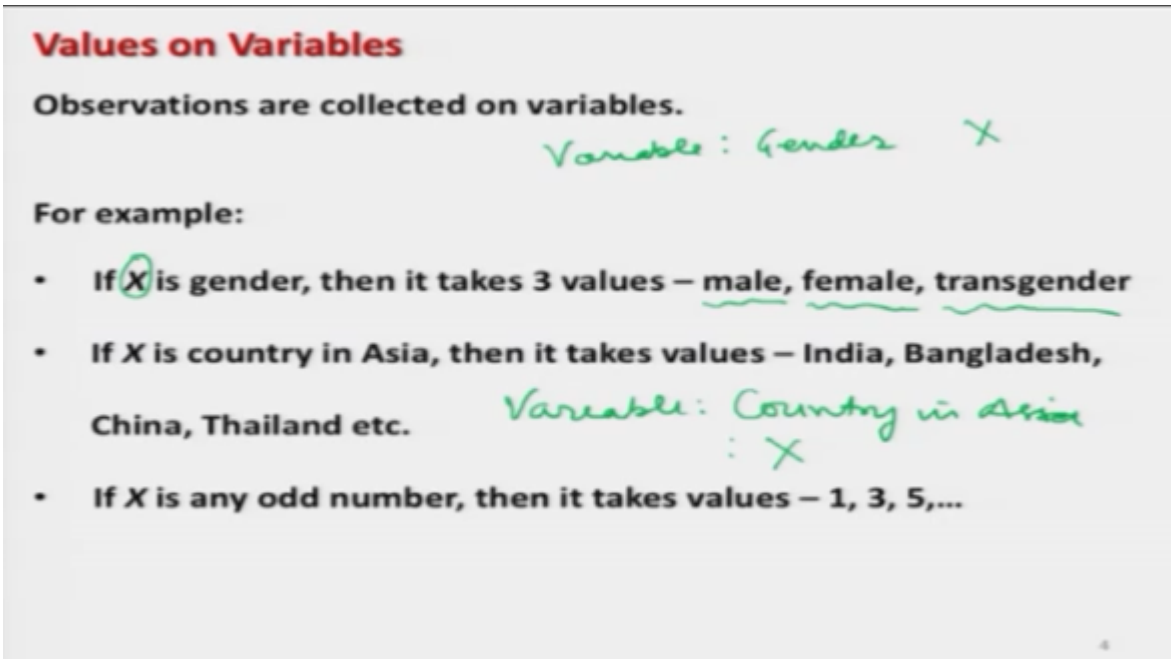
This number of variable can be 1 or they can be more than 1 also, so whenever we are trying to deal with one variable, then the statistical analysis is usually called as univariate statistical analysis or univariate analysis, and whenever we are trying to deal with more than one random variables at the same time then we call it as multivariate analysis or multivariate statistical analysis.

(Refer Slide Time: 04:58)



Now what is the role of variable? The observations are collected on the variable, now I'll try to take some examples to make you understand that how the variables are defined and how the observations are collected on them, suppose I want to know in some college that how many male students, how many female students and how many transgender students are there, so in that case my variable will be gender, gender of the student, and this I will denote as say here capital X, and I'll type it as italics like as here, and this variable will take 3 possible values, one will be here, the student can be male, the student can be female or the student can be transgender.

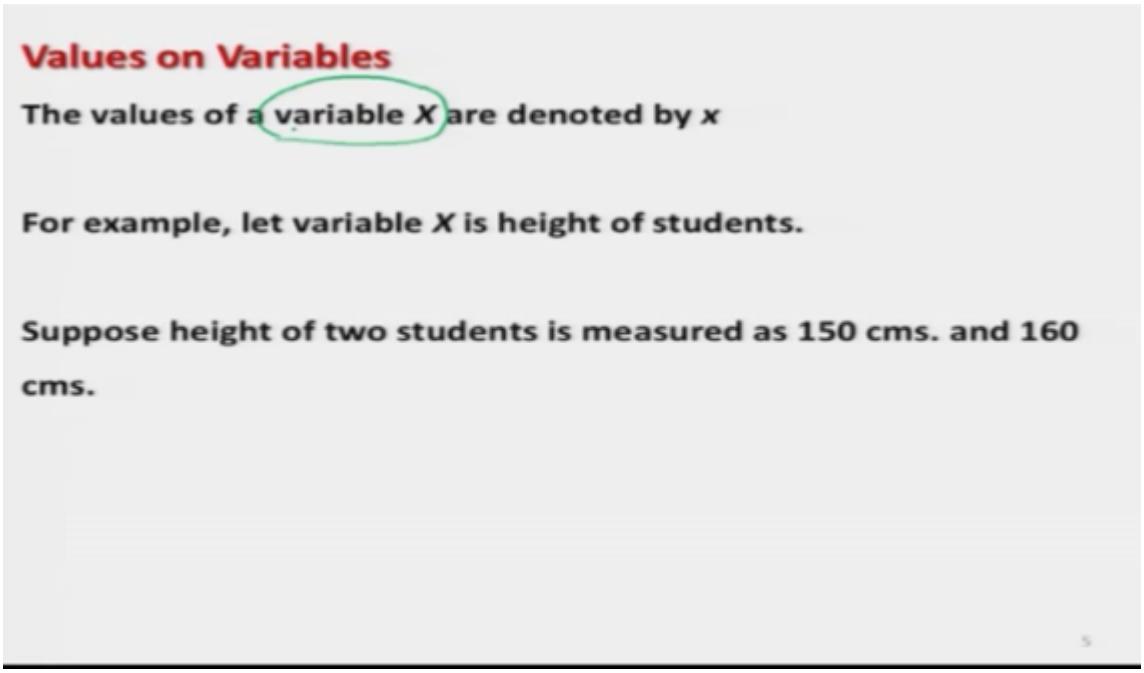
Similarly in case if I want to denote or if I want to study about some country in Asia, then I can also define that my variable here is say country in Asia, and I will denote this by here capital X, (Refer Slide Time: 06:30)



now this capital X can take different values, the countries in Asia can be India, they can be Bangladesh, this can be China, this can be Thailand, this can be Bhutan and so on, so now you can see I have done here two things, I have defined the variable and I also have given an idea that the variable can take different possible values, right.

Similarly if I try to take any other example, suppose if I denote say any odd number, so I can denote the variable by here X, and X is going to denote any odd number, and what are the different possible values we just can't take in this case? The odd values can be 1, 3, 5, 7, 9 and so on, so if you try to see through this example I have given you one more aspect that the number of values which are variable can take this can be finite and also not, so this is what you have to keep in mind when we are going into the further lectures. And in this example I have defined the random variable as X which you always have to keep in mind.

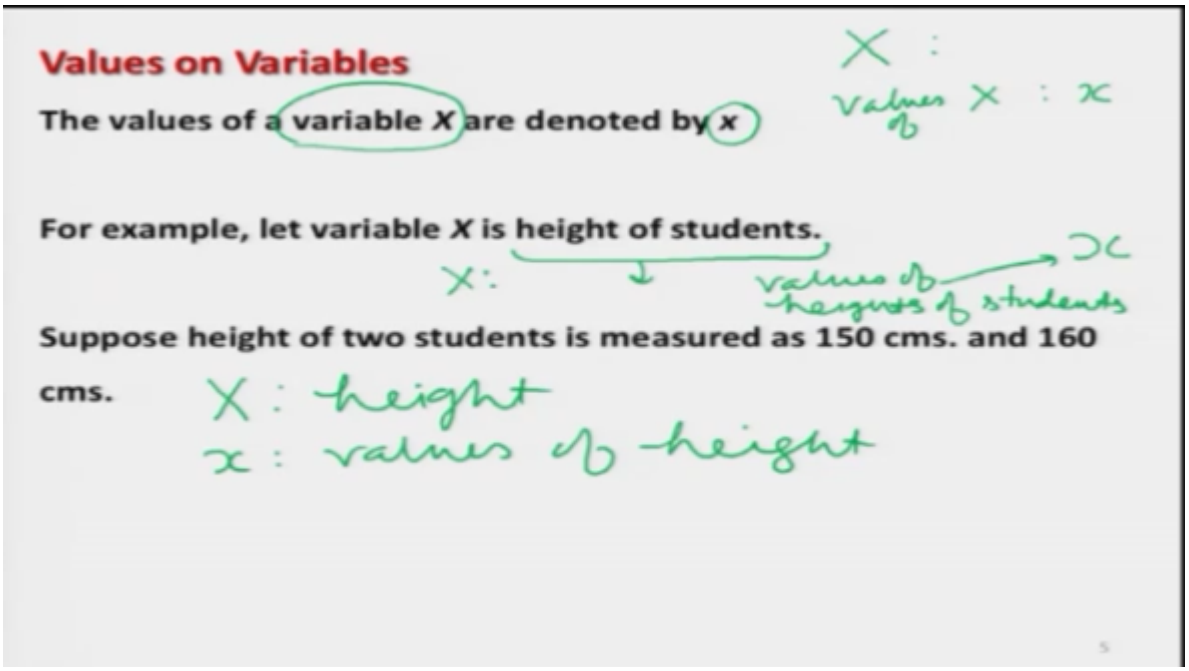
(Refer Slide Time: 07:53)



Now what will be our next step? Once I have defined the variable, then I would try to draw a representative sample or simply a sample from the population, and whatever are my sampling units I'll try to collect or I will try to record data on them, for example in case if I want to record the ages of children, in the age group of 5 to 7 then suppose I try to choose say 10 children whose ages are between 5 and 7 years from a population of that city, of that country, or that state whatever you want, and now I would try to record the age of those children, so the value of the age that will always be denoted by the corresponding small letters, so if I try to say here I have denoted the random variable by X, then the values which X is going to take they are denoted by small x like as here, so the value of the ages of those children will be denoted as small x.

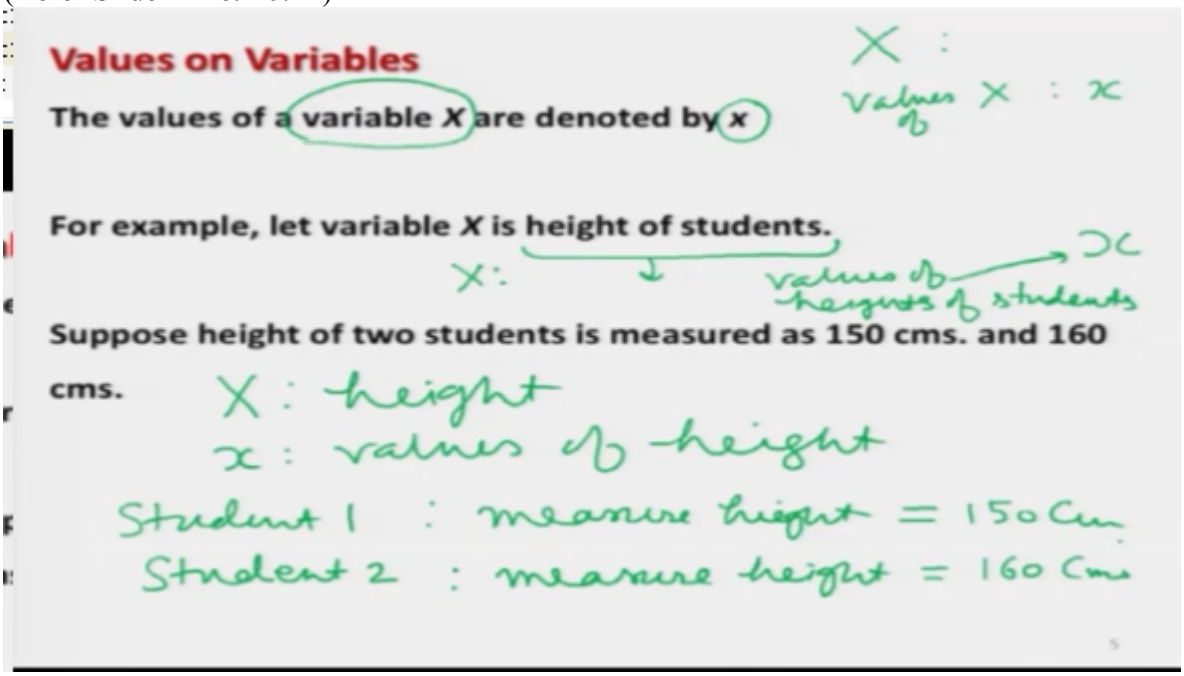
And suppose if I want to find out the average height of the students in a school, then my objective will be to collect the data on the height of some students of that school, so I would try to represent here capital X to be the height of the student, and whatever are the values of heights, heights of students this I'm going to denote by here small x, so I can now say here X is going to denote my height, small x is going to denote by values of height, and now I started collecting the data,

(Refer Slide Time: 10:04)



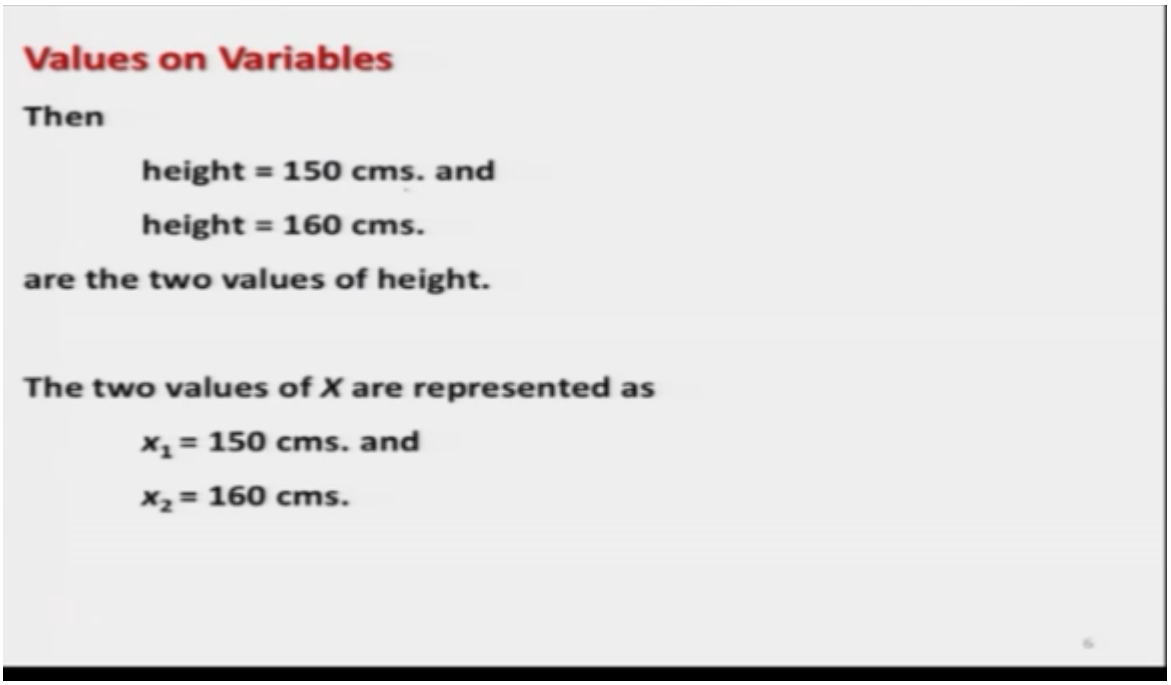
suppose I take two students, student number 1 and student number 2 and I try to measure the height, and again I try to measure the height of the second student, so suppose I find that the height of the first student is 150 centimeters and height of second student is 160 centimeters,

(Refer Slide Time: 10:44)



now how to denote this 150 and 160 that is the question which I'm going to now address here,

(Refer Slide Time: 10:54)

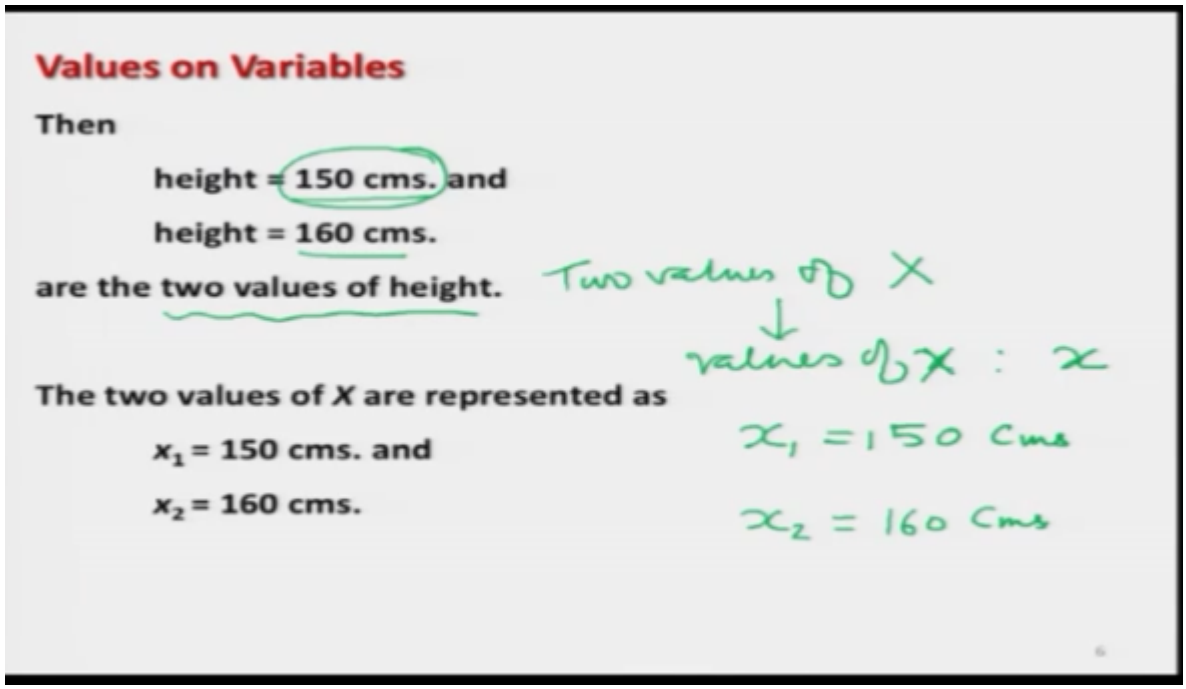


so height equal to 150 centimeter and height equal to 160 centimeters they are the two values of the variable height, these are the two values of X.

And values of X are denoted by small x, so now I can denote here that the first value which is here 150 centimeter this is here the value of random variable x and this is the first value, so I can denote it here say X1 and I can write down here X1 = 150 centimeter.

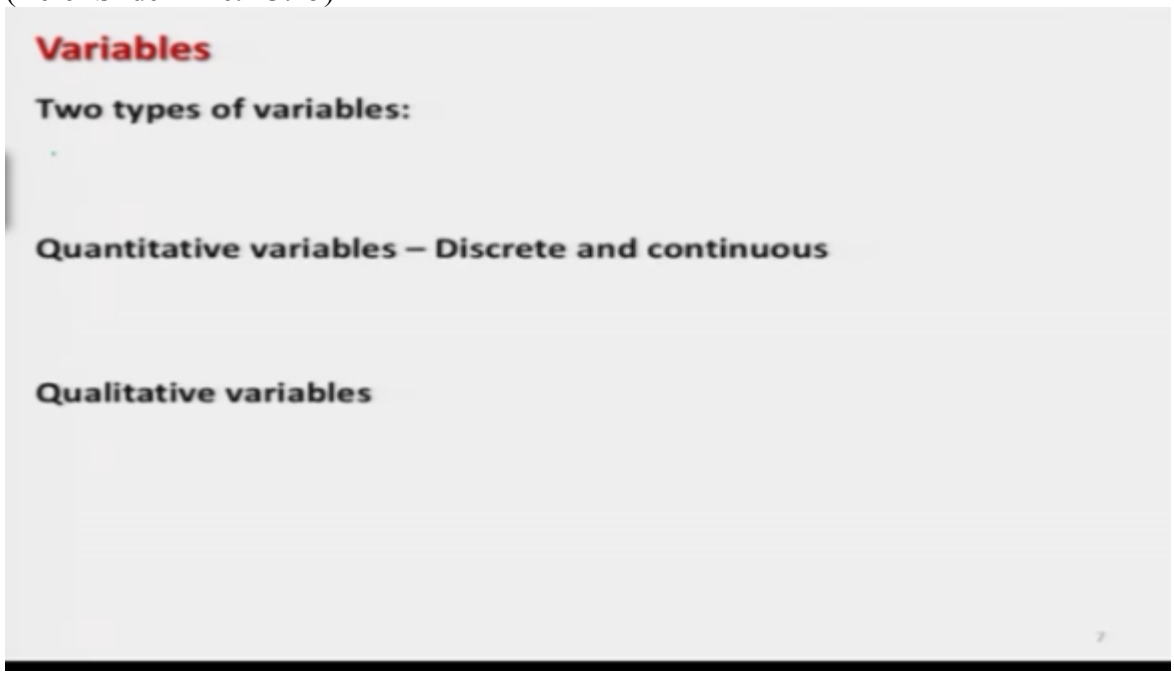
Similarly the height of the second student or the value of the height of the second student, this is X and this is my second student so I can write down here X2, and this I can write down here 160 centimeters, right,

(Refer Slide Time: 11:58)



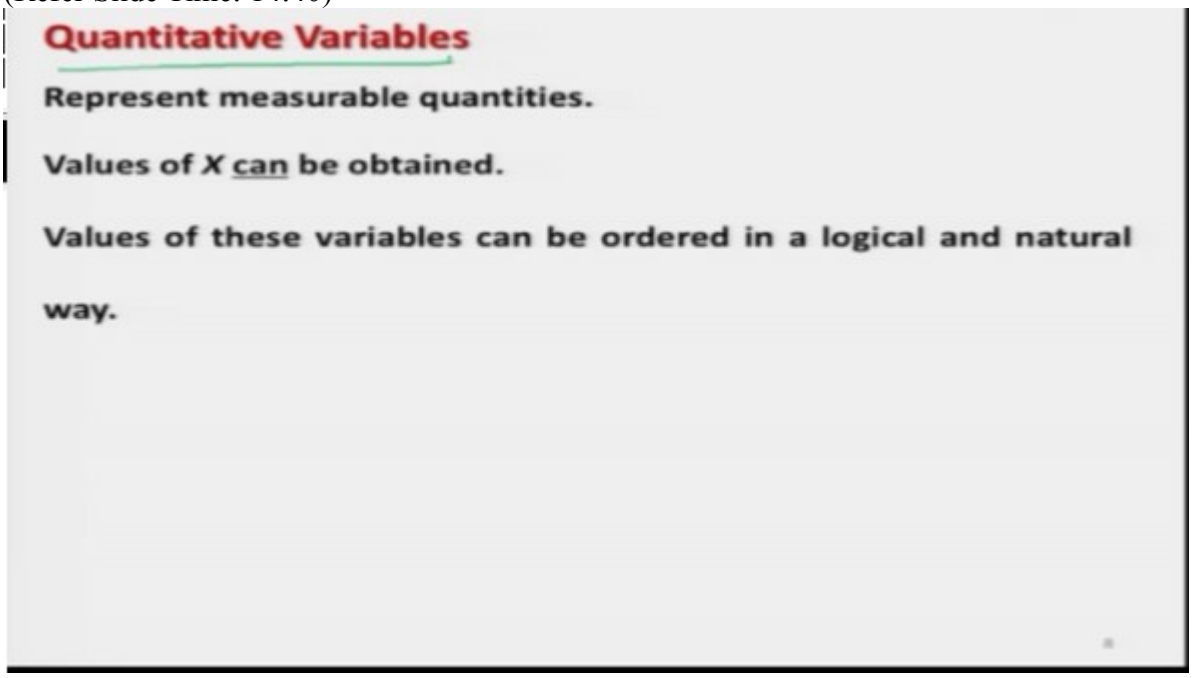
so you will see in the statistics that is a very common sentence let X1, X2, XN be our sample from some population, so once I tried to write down here let X1, X2…XN be a sample or a random sample, what does this mean? This simply means that these X1, X2, XN they are some numerical values, nothing more than that, and they are the numerical values on what? They are the numerical values of the data, data on, data that is recorded on the variable X, so if I say X is height, and suppose I have collected 20 students and I have recorded their heights, then the values of those heights that is going to be denoted by X1, X2 say … say X20, so this is the simple interpretation of this notation, and which is going to be used in all the further lectures.

(Refer Slide Time: 13:15)



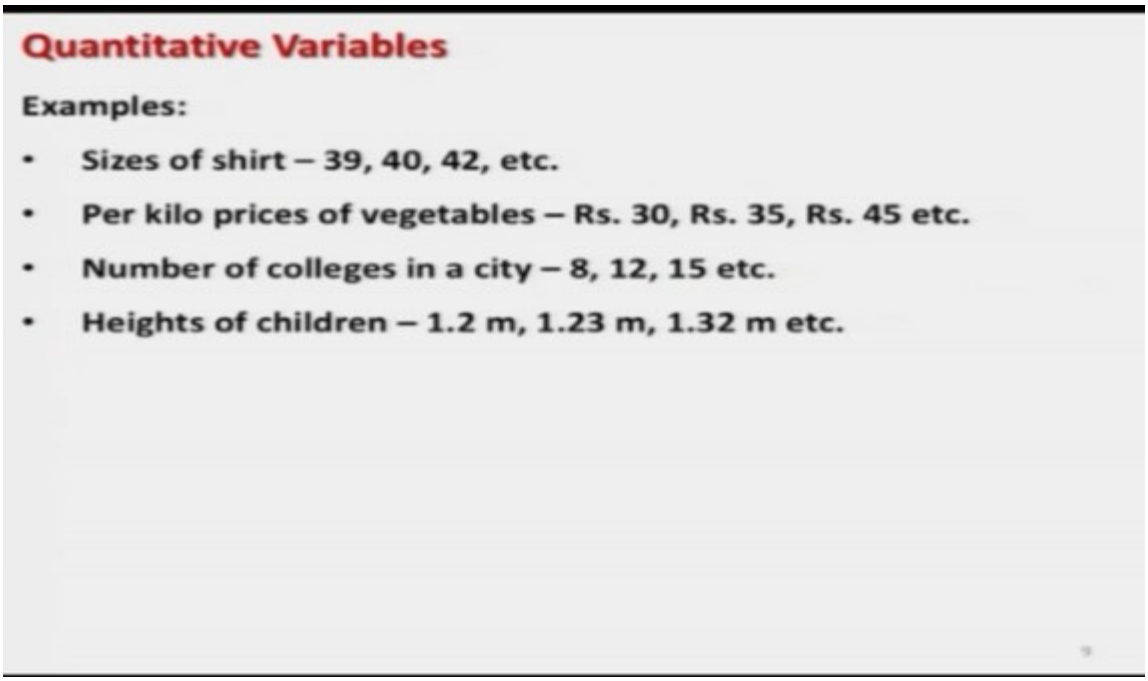
Now the next aspect which I'm going to address about the variables is that there are two types of variables, one is quantitative variables and other is qualitative variable, and under the quantitative variables we have two types of variables, one is called discrete variables, and other is called is continuous variable. Once again I would like to reiterate here that there is a strong mathematical definition of this quantitative variable, discrete variable, continuous variable, qualitative variable in a statistics, but definitely my objective is not here and I'm not going into that detail, but my simple objective is to make you understand that given a situation when you are trying to collect that data, the corresponding variable will be a discrete variable, continuous variable or qualitative variable or a quantitative variable, once you are able to judge this thing, then after that you will see that the tools for example quantitative variables, qualitative variables they are different, the statistical tool for discrete data for the continuous data they are also different, so that will help you in choosing the correct and appropriate statistical tool, okay.

(Refer Slide Time: 14:40)



So first I try to address here the aspect of quantitative variable, in very, very simple words I can say without going into the mathematical details, that quantitative variables represents some measurable quantities, measurable quantities means the values of X, the numerical values of X can be obtained and once these numerical values are obtained then they can be ordered in a logical or a natural way, possibly this is one of the most simple definition I can give you about this quantitative variable, and what does this actually mean?

(Refer Slide Time: 15:25)



Let me try to take some example to make you understand, suppose I want to buy a shirt, and I go to a shop then the shop keeper will ask me what is the size of your shirt, that can be 38, 39, 40, 41, 42, 43, 44, 45, 46 right, so if I try to take here the size of the shirt this can be 39, 40, 41, 42 and so on, what does this mean? 39 means there will be some dimension, some size of the shirt, and if I say what is the difference between 39 size shirt and say 42 size shirt, we understand that the size of the shirt with number 42 is going to be larger than the size of the shirt containing the number 39, so you can see here that this 39 and 42 they are representing some numerical values and they can be ordered, that means I can always say that the size 42 shirt is larger than the size 39 shirt.

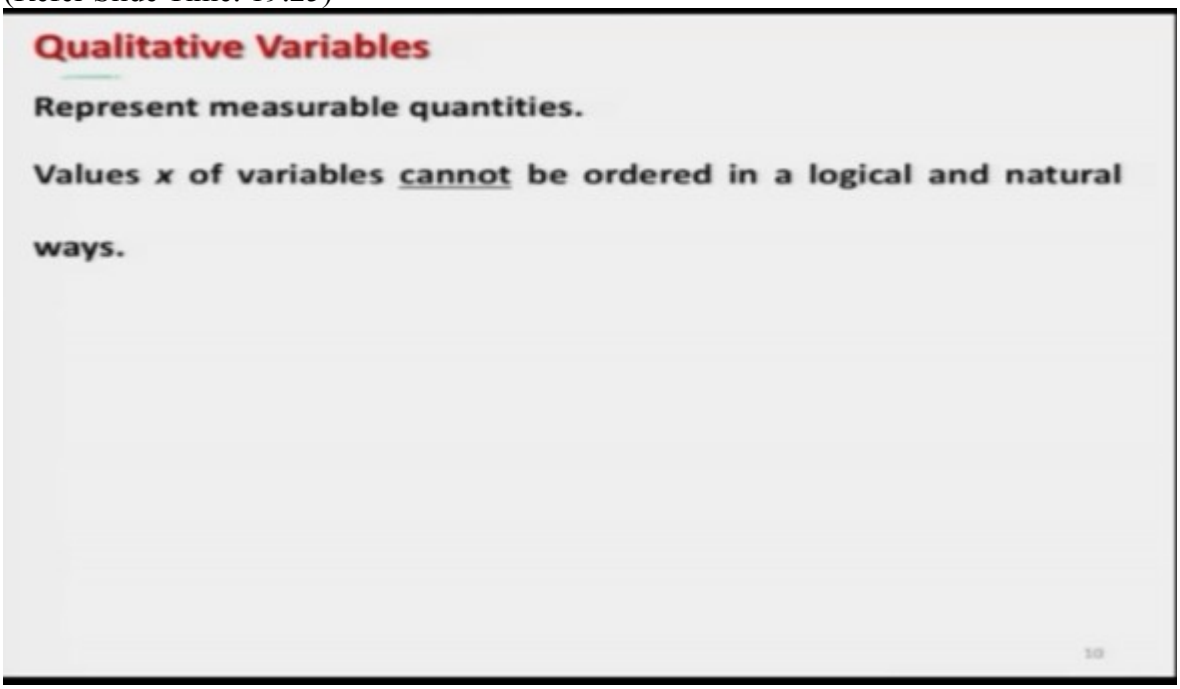
And similarly if I try to say, try to take an example of a cost or the price of the vegetable say, per KG price of say this some vegetable, this price can be 30 rupees a kilo, 35 rupees a kilo, 40 rupees a kilo, 45 rupees a kilo, what does this mean? Once I say you have to give 35 rupees for one kilo of the vegetable you know what you have to give, and in case if I say that the price of the same vegetable in one shop is 35 rupees per KG, and the shop of the same vegetable in say another shop is 40 rupees a KG then you can always make a conclusion by putting them into order that the price of the vegetable in the second shop whose rate is 40 rupees a KG is higher than the rate of the vegetable in the earlier shop where the price was 35 rupees a KG.

And similarly if I say the price of the vegetable is 50 rupees then you will say that it is more expensive than the earlier shops. And similarly if I want to count the number of colleges in a city, this number can be 2, 5, 10, 8, 12, 15, 20 so whatever is there, so now once again these values have some interpretation, if I say I have 2 cities one say city has 10 number of colleges and say another city has 20 number of colleges, then I can always order them and can say that the second city has more number of colleges than in the first city.

Similarly if I try to measure the heights of the children say 1.2 meter, 1.23 meter, 1.32 meter and so on, then you know that this 1.2 meter and 1.23 meter has some interpretation and you

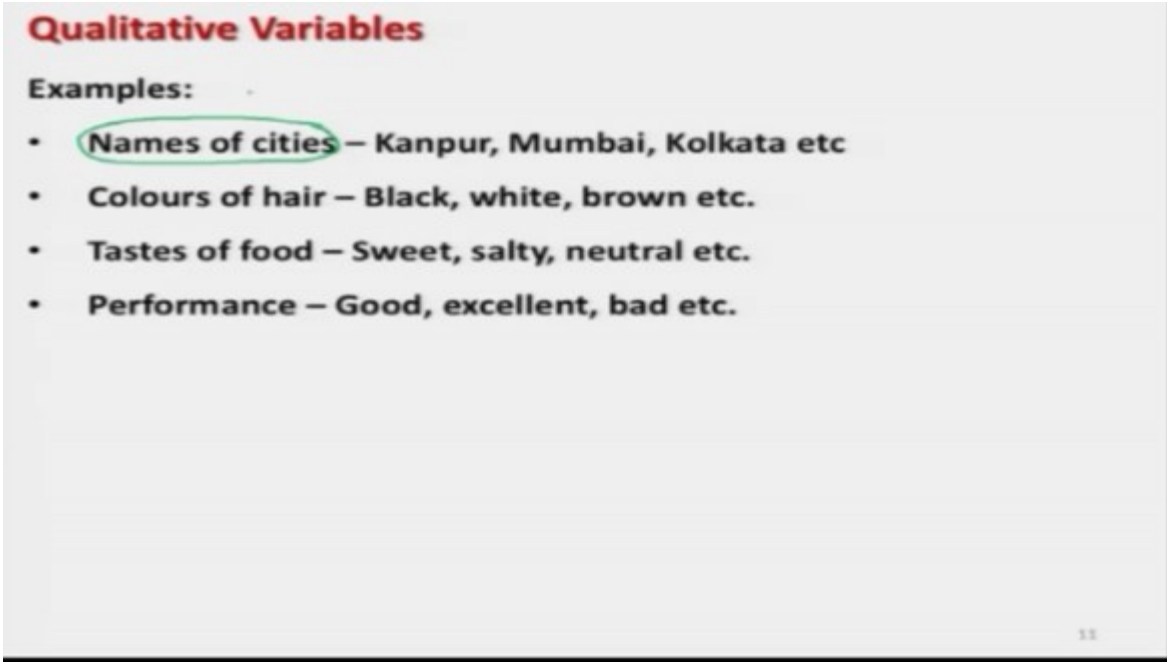
can visualize these things, suppose there are two children and suppose you record to get the height of the first child is 1.2 meter and the height of second child is 1.3 meter then you can always infer that the second child is taller than the first child, so this is what we really mean by quantitative variables, I have a variable like as price, height, number etcetera, and I can obtain the values, numerical values on them and those numerical values have some interpretation, interpretation means they can be ordered and they will have some meaning in their numerical value, right.

(Refer Slide Time: 19:25)



Now I'll try to address the aspect of qualitative variables, this qualitative variables represent the measurable quantities, so this is same as in the case of quantitative variable, then what is the difference? The difference lies here that the values of the random variable or the values of the variable which are denoted here as say small x they cannot be ordered in a logical and natural way.

Once again I would say possibly I can think of this is one of the most simple definition to understand for a common person not having a statistical or say a strong mathematical background, what is this actually mean? That the values cannot be ordered in a logical and natural way, so let me take here some example and I try to explain it to you, now suppose I want to collect the data on the names of cities in this country India, okay,
(Refer Slide Time: 20:28)



so I will define my variables as say here X, and now this variable will take different values, for example Kanpur, Mumbai, Kolkata, Delhi and so on, so these values are going to be represented as say X1, this is the first value of the variable that it can take.

Second value of the variable which it can take and third value of the variable Kolkata that the variable can take, but this variable is very well understood, but how to order them, how to put them in a natural way? Well, as soon as I say this thing you always try to associate some number with it, but here I am not associating any aspect of this variable, I'm calling only at their name, but in case if I associate like that number of person is staying in a city, then it will become a quantitative variable, or if I associate the area under that city then this will become a quantitative variable, but up to now this is only a qualitative variable.

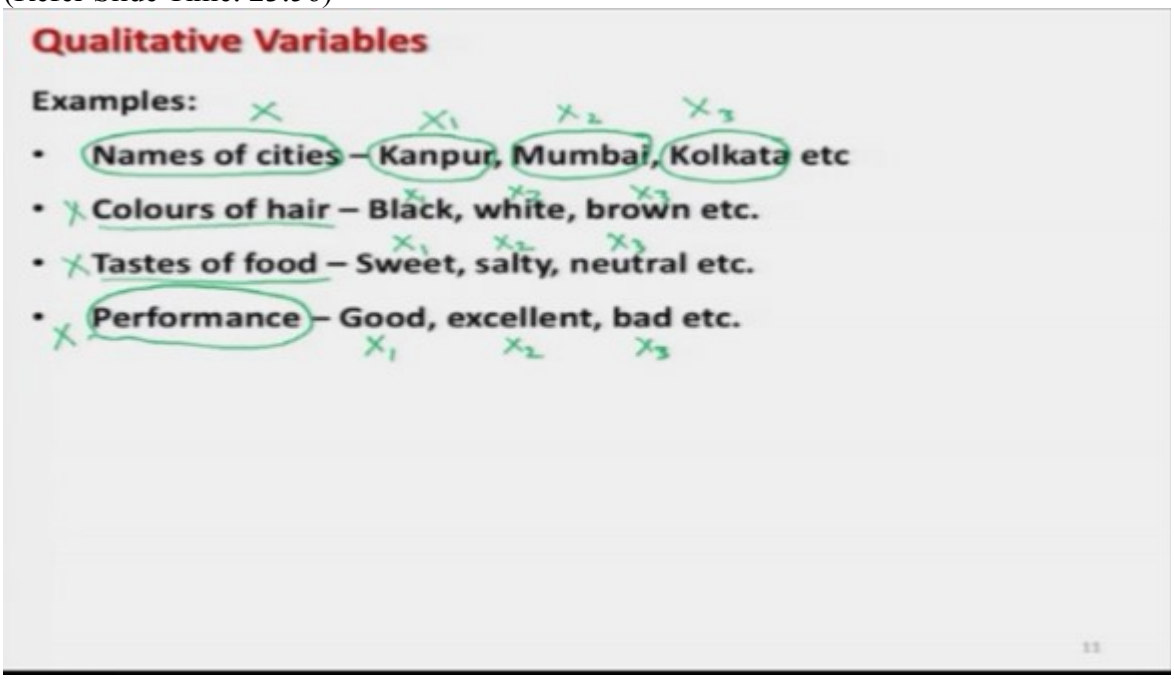
Similarly if I take another example then I say I want to record the colours of the hair, for example now they can take different values say black, so I can denote black by capital X1, white by capital X2, and brown by capital X3 denoting the first, second and third values which the variable can take.

Yes I can very well understand that the colours of the hairs are black, brown, white or something else, but how to quantify it? Unless and until I try to associate a degree or some measure on some scientific the colours will remain only as a colours and there is no way that I can order them that for example like white is better than grey or grey is better than brown and so on, but here I'm recording only the data, so that is why the colours of the hair this variable is a qualitative variable.

Similarly if I take another variable here tastes of food, which can be sweet, which can be salty or which can be neutral and so on, so I can denote the first value that the variable takes X1 to be sweet, second value of the variable X2 to be salty, and third value of the variable to be neutral, and these values sweet, salty, or neutral they are only the qualitative variable they are not the

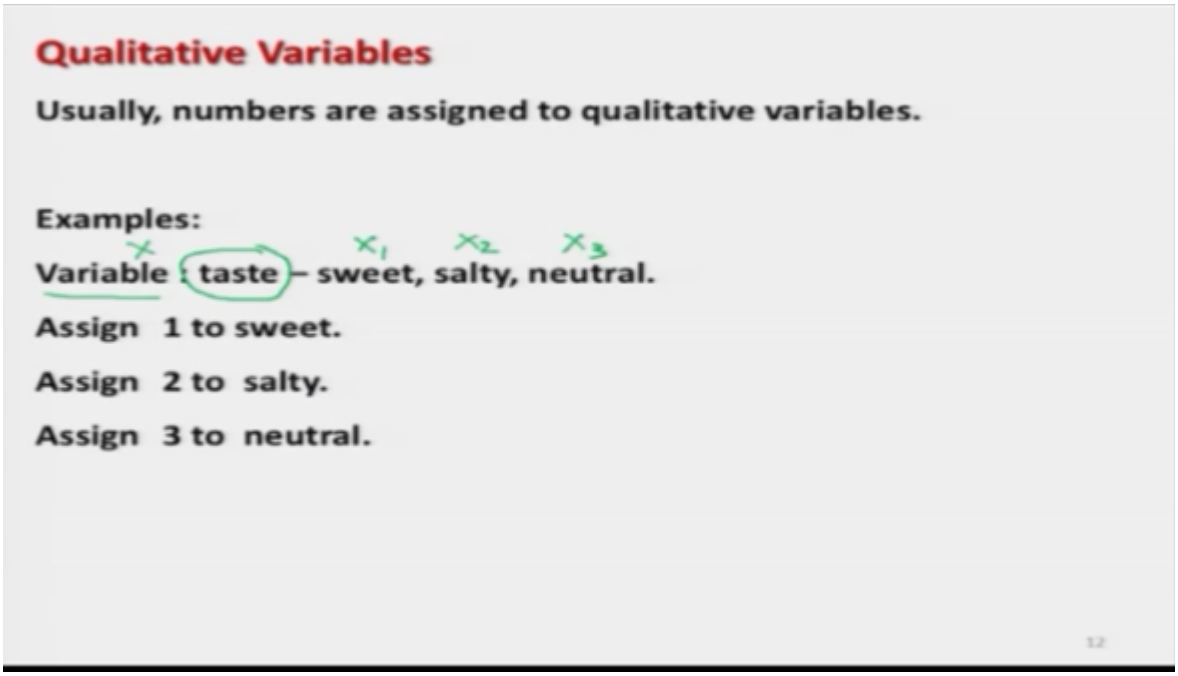
quantitative variable, I cannot say that sweet takes 20 or salty takes 30, so this is the idea behind the qualitative variable.

And similarly many times in examination or in any competition we try to adjust the performance of the candidate by making it good, excellent or bad, so in that case this performance can be variable which is denoted by the variable X, and the good this is the first value which variable takes excellent, this is the second value which variable take, and bad is the third value of X, so again good, excellent, bad,
(Refer Slide Time: 23:56)



these are only some qualitative things, I can understand them, I can observe them but I cannot quantify them unless and until I say that if a student is in the excellent category then it is better than the student who are in the category of good and bad, or similarly if a student is in the category of good student then he will be considered as better than the student in the bad category, so unless and until I make this types of rules which are again trying to denote a quantitative way up to that point, the variable will remain only as a qualitative variable, right.

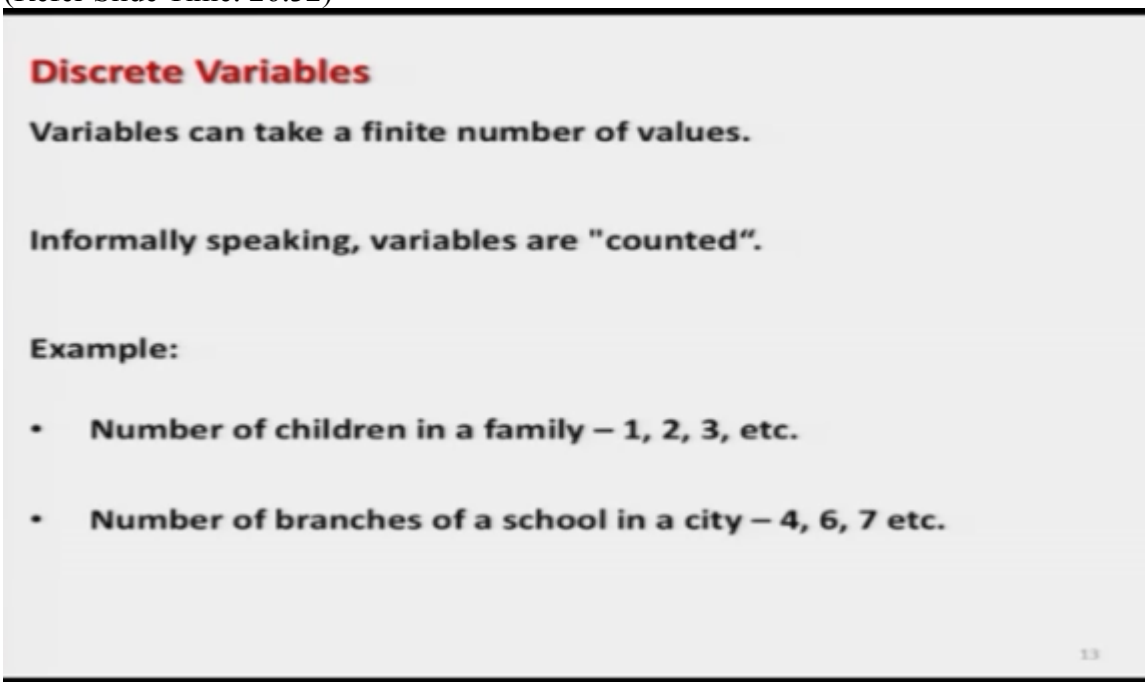
But in statistics now we have a problem, statistics work only with the quantitative data with some numbers, so in case if I have got a qualitative variables, unless and until I associate a number with this I cannot operate my statistical tool, and here at this point I will try to inform you that the statistical tool for qualitative variable and quantitative variable they are different in most of the cases, so you have to be very very careful when you are trying to use a tool whether you are applying it on a qualitative variable or a quantitative variable, so for example now I will try to take an example to show you that how we try to handle the qualitative variable by associating a number with them, so suppose I consider here the variable say here X, say here as taste, now taste is taking here 3 possible values, capital X1 it denoting the sweet, capital X2 it denoting the salty, and capital X3 denoting as neutral, so right,
(Refer Slide Time: 25:44)



so what I will try to do, I will associate a number with these 3 indications, sweet, salty, and neutral.

Suppose I decide that I would say assign 1 to sweet, I will say assign number 2 to salty, and I will assign number 3 to neutral, but remember one thing once I'm trying to assign this number 1, 2 and 3 they are only indicating the category, means if I'm assigning 2 to salty and 1 to sweet this does not mean that salty is 2 times of sweet, this will be a wrong interpretation, so here I'm simply trying to assign a number to indicate the category.

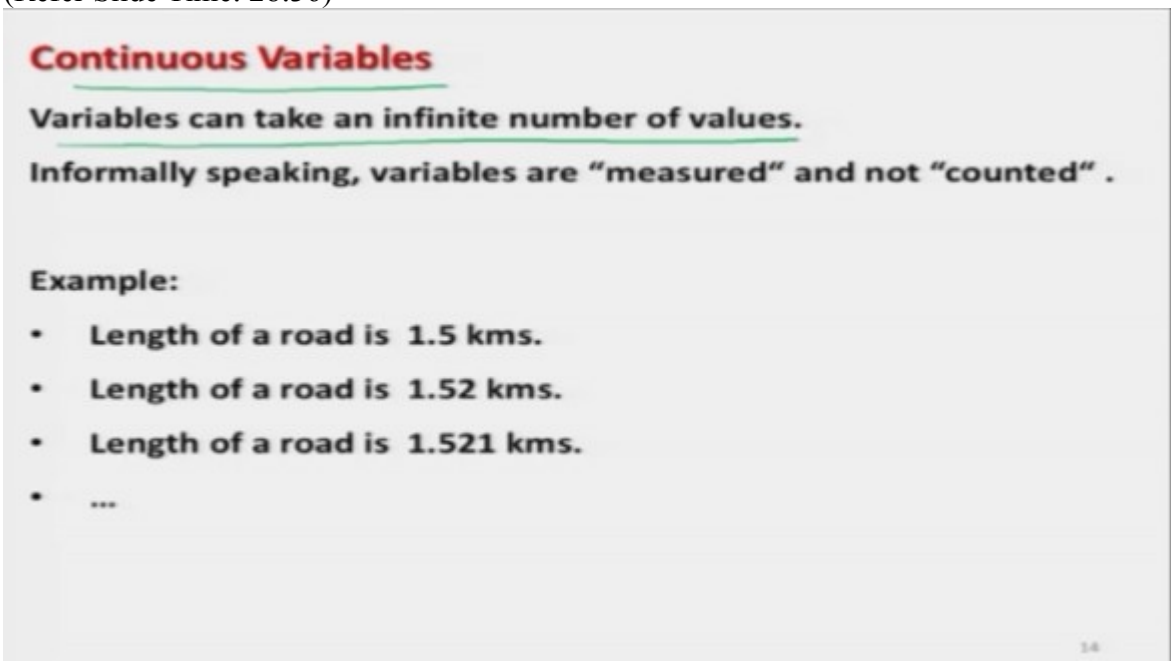
(Refer Slide Time: 26:32)



Now after this I'll try to address the discrete variables, in some situation the variable on which we want to record the data that can take only a finite number of values, and in a very simple way or in an informal way I can say that the variables are counted, for example in case if I want to find what are the number of children in different families, so this number can be 1 child, 2, 3 and so on, they cannot be that a family has 1.2 children or 2.4 children, these value will not exist or they will not have any interpretation, so in this case I am simply trying to count the number as a whole number.

Similarly if I try to find out the number of branches of a school in a city this number can be 1, 2, 3, 4, 5, 6, 7, but this number cannot be 2.5 or 5.5 or 6.7, so in this case these values are being counted, so all those variables where we are going to record the data on the basis of only counting they can be categorized as discrete variable for all practical purposes.

So now in case if you try to associate this definition of counting, then in case if I try to say there are, there is another option that a variable can also take a value in fraction, say 2.1, 2.2 and 2.3 also, so those variables which can take an infinite number of values they are called as continuous variable, so basically there are two categories discrete variable and say continuous variable,
(Refer Slide Time: 28:36)

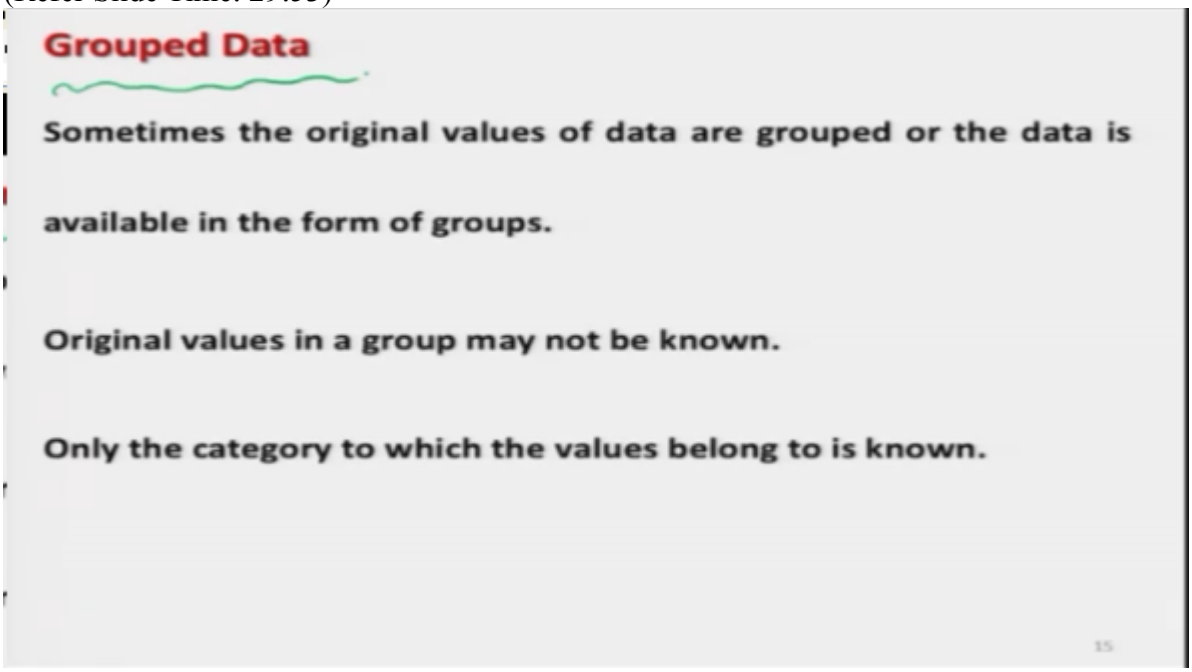


so in discrete variable the values are counted and in continuous variable case the number of values what a variable can take that can be infinite, and in simple words in formally I can say that the values are measured, and they are not counted that is very important to note, they are being measured, for example if I say suppose I want to measure the length of a road in certain fraction, it all depends how we are going to measure it by which instrument that is a separate aspect.

Now let me take an example, suppose I want to measure the length of a road, the length of a road can be 1.5 kilometer, this can be 1.52 kilometer, or this can be 1.521 kilometers and so on, so in this case you can continue as long as you want depending on the instrument, depending on

the length, so in this case I am measuring the value and this value can take infinite number of values, so this type of data usually we collect under the headship of continuous variable.

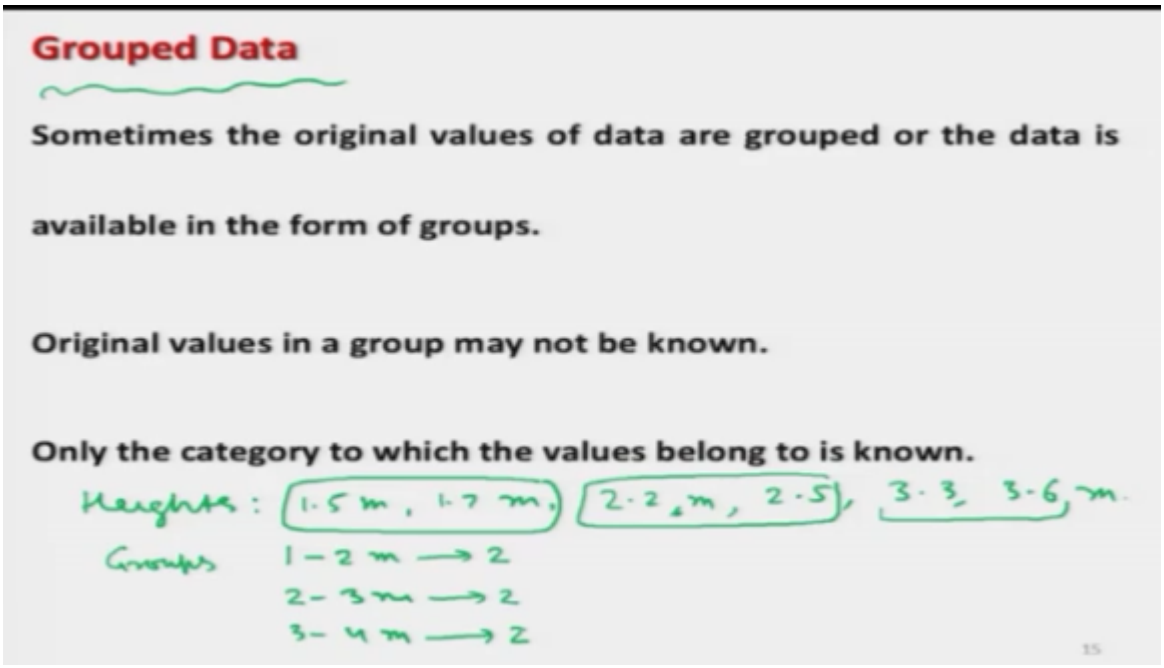
(Refer Slide Time: 29:53)



Now I'll try to address another aspect, this is called grouped data, suppose you have got large number of values, then in that case it is possible to group those values in certain categories or certain groups, and then what will happen that the original value or the behavior of the variable will be such that, that the nature of the exact value will be lost and that value will be identified only by the category, suppose I try to take an example, suppose I tried to measure the heights, this height can be 1.5 meter, 1.7 meter, so 2.2 meter, 2.5 meter, 3.3 meter, 3.6 meter and so on.

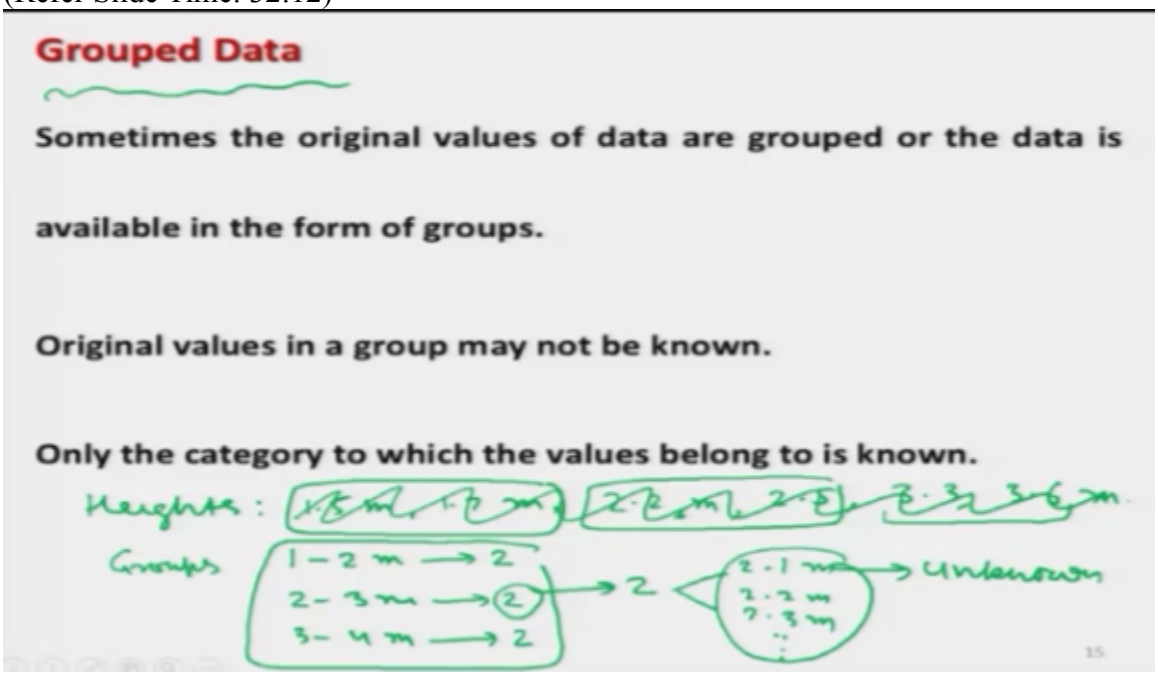
Now I can make here 3 groups, say group 1 where the heights are between 1 to 2 meters, the height are between 2 to 3 meters, and say last group in which the heights are between 3 to 4 meters, now they are here two values 1.5 and 1.7 meters which are lying between 1 and 2 meters, so these are 2 values so I can write down here 2 values.

Similarly here 2.2 and 2.5 these are two values which are lying between 2 and 3, so I can write down here 2 values, and similarly 3.3 and 3.6 they are lying between 3 and 4 so I can write down here 2 values, now I will have only this information, and this information will be hidden.
(Refer Slide Time: 31:50)



So now looking at say this value here 2, I cannot find whether this value was 2.1 meter, 2.2 meter, 2.3 meter or something else, I do not know, so these values will become simply unknown to me,

(Refer Slide Time: 32:12)



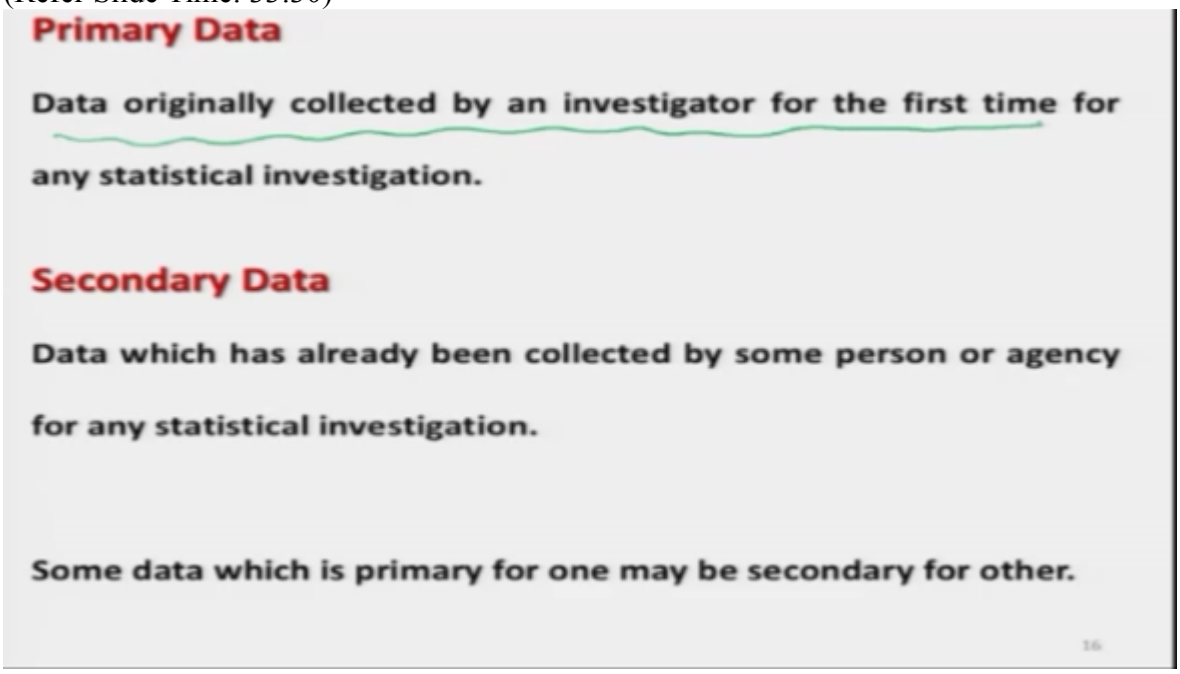
so whenever we are trying to deal with the group data we have to keep in mind that the values will be grouped together, they are individual values will be lost and that we will be working only the data that is represented in the form of those groups.

Now I would like to briefly address another aspect that is called as primary data and secondary data, so what is the difference between the primary data and secondary data, so you see suppose I will objective to study, and based on that either I go to the field or I ask some of my

investigator to go to the field and collect the data directly, and I try to work on this thing, so this data will be called as primary data.

Whereas second option is this means I can go to some offices like as municipalities of the cities or there is a, and there are different offices like a National Sample Survey Organizations who collect the data from time to time on different aspect, I can request them to give me the data and I tried to work on that data, so this data has been collected by somebody else, either a person or an agency and we are trying to use the data from that source, so this type of data is called as secondary data,

(Refer Slide Time: 33:30)

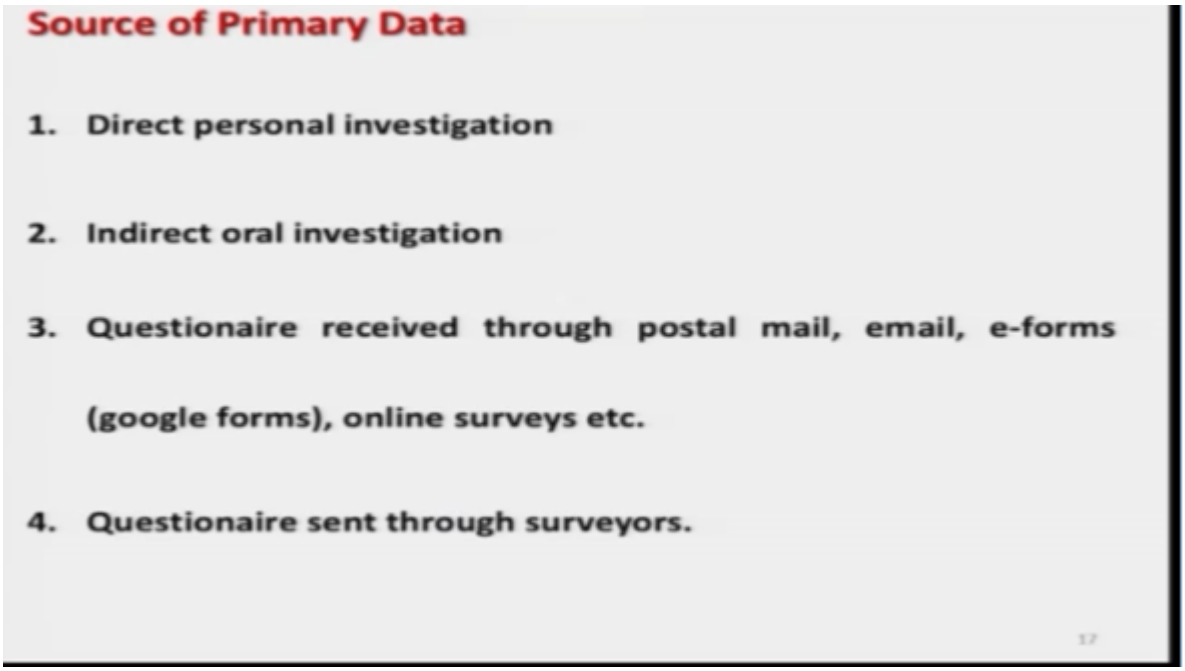


so I can say here very briefly that the data which is originally collected by an investigator for the first time with an objective to study any statistical query or statistical investigation that will be called as primary data.

And the data which has already been collected by some person or some agency for any objective or say for any statistical query or for any statistical investigation, and we are trying to borrow the data or we are trying to collect the data from their agency and then we are using, this is called as secondary data.

Well the definition of this primary and secondary data is very relative, some data which is primary for someone, and maybe secondary for say for the other person, so I'm not going into that detail but this was just for the, for your information.

(Refer Slide Time: 34:24)



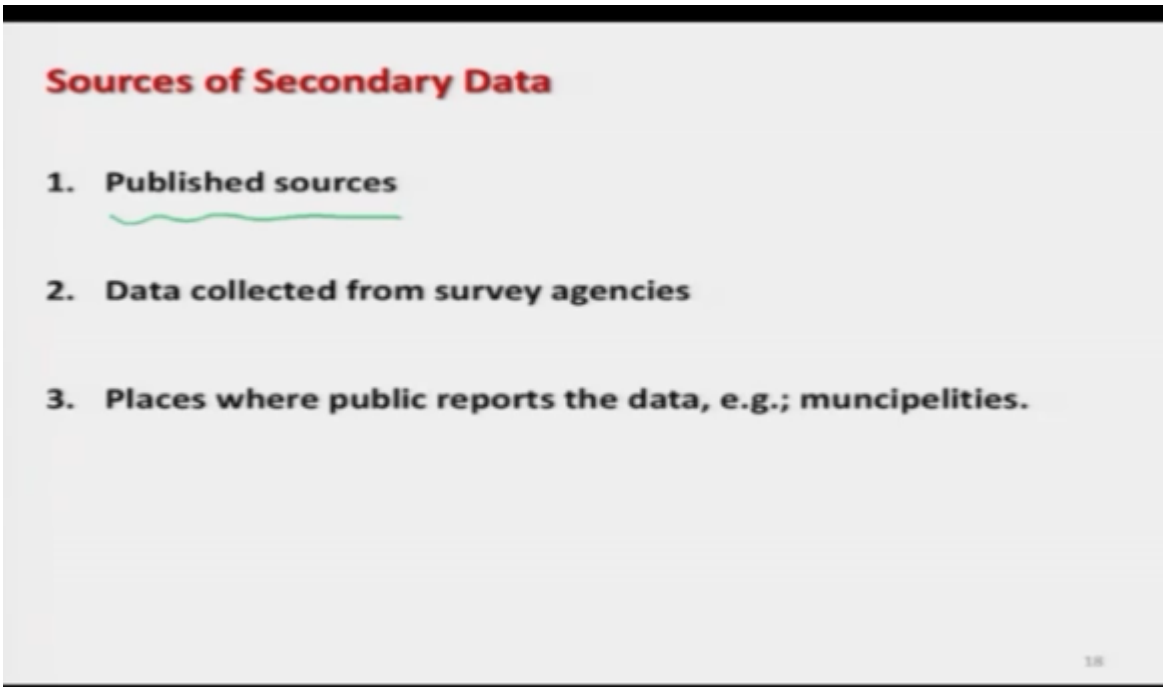
Now the next question comes, how this data comes into picture? How this primary data is obtained, how this secondary data is obtained? So very briefly I can give you different ways in which this type of data are collected, so in order to collect the primary data one of the important source is direct personal investigation that the person goes directly to the respondent, and he or she ask the question and he or she tries to record the answers directly.

Second thing is this indirect oral investigation that the person will go, he or she will ask different types of questions and based on that he will try to make a judgment that what is the numerical value of the variable.

Third popular option is this some questionnaire which I have sent through postal mail, email, e-forms like as nowadays Google forms are very popular and through online service there are some websites which tries to help us in conducting a survey, so they also try to give us the primary data.

And sometime we send our surveys, surveyors through the field and we don't allow them to ask anything but we will give them a questionnaire and we ask them that you please give it to the concern person and he will or say she will collect the data and give it back to us, so and then there are many other ways also, but this is how we try to collect the primary data.

(Refer Slide Time: 36:00)



For the secondary data there are some published sources, for example there are some reports or data sets available where the countries, offices responsible for samples survey, for example in India we have National Sample Surveys Organization, Central Statistical Organization, and in definitely at some more level we have United Nations and then there are different wings who try to collect the data from time to time and they try to published the data and we can use that data.

The second option is this the data which is collected from some survey agencies we can use them, and third thing is for example there are some public offices where we record our data for example municipalities, whenever there is a birth of a child or death of any person we tried to go to the municipality and we try to report it there and they try to keep the data, so this type of data is also available in those municipalities and we simply try to take it from there.

So now in this lecture I have given you a background under which we will be working in for the lectures, from further lectures now I will try to take one topic at a time, I'll try to give you the basic idea, I will not be going much into theory but I'll try to explain you with different example that what are the different concepts and I will try to show you that how to obtained those things using the R software, so now I will be going into the different tools of descriptive analysis from the next lecture, so you please try to review this lecture, you please try to revise the lecture and try to understand the concepts, try to settle down inside your brain and we will see in the next lecture, till then good bye.

Acknowledgement
Ministry of Human Resource & Development

Prof. Satyaki Roy
Co-coordinator, IIT Kanpur

NPTEL Team
Sanjay Pal
Bharat Lal
Ashish Singh
Badal Pradhan
Tapobrata Das
K. K. Mishra
Ashutosh Gairola
Puneet Kumar Bajpai
Bhadro Rao
Shikha Gupta
Aradhana Singh Rajawat
Sweta
Nipa Sikdar
Anupam Mishra
Ram Chandra
Manoj Shrivastava
Dilip Tripathi
Padam Shukla
Sharwan K Verma
Sanjay Mishra
Shubham Rawat
Santosh Nayak
Praduyman Singh Chauhan
Mahendra Singh Rawat
Tushar Srivastava
Uzair Siddiqui
Lalty Dutta
Murali Krishnan
Ganesh Rana
Ajay Kanaujia
Ashwani Srivastava
M.L. Benerjee

an IIT Kanpur Production