

Essentials of Data Science with R Software -2
Sampling Theory and Linear Regression Analysis
Prof. Shalabh
Department of Mathematics and Statistics
Indian Institute of Technology, Kanpur

Sampling Theory with R Software
Basic Fundamentals
Lecture - 09
Sampling, Sampling Unit, Population and Sample

Hello, welcome to the course Essentials of Data Science with R Software- 2, where we are trying to learn the topics of Sampling Theory and Linear Regression Analysis. And in this module, we are going to start with the Basic Fundamentals related to the Sampling Theory with R Software.

Now, from this lecture, I am going to take up the topics from statistics. The first topic which I am going to take here it is about sampling theory. Now, before I go further into the topics, I would like to make some observations which you have to keep in mind while you learn from this course.

When I am trying to prepare this lecture, I am thinking of my audiences which consists of undergraduate students, data professionals, those students who are from the statistics stream, who have taken statistics as one of their main subject and there are students for whom statistics is not the major subject, but this is a part of their curriculum.

Beside this audience, I have another group which are the practitioners. So, my efforts are going to strike a balance among all of them, their needs. As a student, you would like to learn the theory, the formula as well as the proofs of those results. As a practitioner you would not be interested in the mathematical proofs of the results.

So, what I am going to do? I will try my best to strike a balance. And the topics of sampling theory; this itself is a full semester course actually; and if I try to take all the topics from sampling theory, possibly this entire course will be only on sampling theory which is not my objective.

So, what I am going to do? I am going to float the concept, I will try to explain it in my best capacity and I will choose some specific topics. And those topics have been chosen

to the best of my ability in such a way which will give you the basic concept and they will explain you the basic fundamentals behind the subjects methodology and their statistical background.

And I feel that if you learn these many topics, then if you want to extend it further; you want to learn more, then you can take the help of books, notes etc. Just for your information I have one course on NPTEL website on sampling theory also where you will get all details of the theory as well as mathematical proofs.

So, I will try to strike here a balance between the theory application and mathematical proof. So, let us begin our lecture from here.

The first chapter which I am going to consider here is about the basic fundamentals and in this lecture, I will be covering several small topics related to sampling, sampling units, population and sample.

You can observe here one thing more that I am trying to give the lecture number with respect to the chapter. For example, this is my first chapter basic fundamental. So, now, I will have lecture 1, lecture 2, lecture 3 and so on. When I start the chapter 2 on say simple random sampling, then again I will have lecture 1, lecture 2, lecture 3 on that topic. So, this is how I am going to proceed further, ok.

(Refer Slide Time: 04:43)

Sampling:

Why do you need sampling?

Description of any statistical tool starts with "Let x_1, x_2, \dots, x_n be a random sample from population...."

Based on this sample, the statistical analysis is conducted.

As a matter of fact, statistics has utility only because it can provide statistical inferences for the entire population using the sample data.

2

So, the first question comes over here, what is sampling and why do you need sampling? For example, those who are from statistics background or though those who are using the statistical tool you will find that the first line of the topic is let x_1, x_2, \dots, x_n be a random sample from some population say normal, binomial or some given population.

And then based on this sample whatever you have observed, the entire statistical analysis is conducted and means if you try to see as a matter of fact, the statistics has got its utility and importance because it can work only on the basis of a sample and it can provide the statistical inferences for the entire population using the sample data only.

So, now first we try to understand, why do we need sampling and how it connects to the data science. Now, in case if you try to understand that whenever we are trying to handle any statistical issue or any issue; we want to answer certain queries, certain questions and for that we try to collect the data.

Why do we collect the data? Because we want to extract the information which is contained or hidden inside the data, but remember one thing, what is the objective of a statistical tool? Do you want to take a decision only for that sample or a much bigger thing which is called as population.

Suppose I want to test that a medicine is effective in controlling the body temperature or not and suppose there is a new medicine developed for this. Now, definitely I need to test this medicine through clinical trials on some animal on some human being and so on. What do I do? Shall I take all the human beings in this world? Or shall I take all the animals on this planet and I try to give this medicine to them?

Answer is no. Rather we try to choose a sample out of that we try to select some or we try to identify some animals or human being and then we try to give the medicine to them. Now, after getting the results, after doing all the statistical analysis, suppose we have a result.

Now, what do you think this result will be valid only for the sample which you have selected or will it be valid for the entire population? Suppose medicine is developed in say US. Now, they would try to conduct a clinical trial possibly in US and if their funds

permit, their condition permit, possibly they may take sample from some other countries also.

But have you ever heard that this medicine can control the body temperature of Indians or of US citizens or of European or something like this? No. Once a medicine is developed in most of the cases that is effective or that has got the same effect on each and every human in this world.

So, my objective is this I want to draw the statistical inferences, I want to make some decisions, I want to make some recommendations for the entire population on the basis of that small sample. So, that is your starting point of study. So, this is the first point because of which the sampling theory plays a very important role in data sciences.

Now, when I come to data science, definitely you would try to collect the observation through some automated way. For example, now you have seen something like Google forms or some web based surveys and so on; that is first thing. Second thing is this in data sciences when you are trying to deal with big data, I mean the data size is not in GB or terabytes, but it is in petabytes.

Then mathematically it is a very big challenge that how the mathematical tools can be applied over such a huge data set. There are several approaches which are coming which we which are being developed nowadays that how to handle the big data. And one of the approach is that from that big data they ask that you try to take a small sample, try to work on it and then try to extend the findings of the sample to the entire population.

And that is possibly the reason that either the classical statistics or the latest data science cannot proceed further without sampling. Why sampling? Because sampling is the science, is the subject which is going to tell us that how should we draw the sample. Why? Because, I want that the sample should have the same characteristic which a population has. So, that is why this sampling theory is important for the classical statistics as well as for the data sciences, ok.

(Refer Slide Time: 11:54)

Sampling:

How to obtain these " x_1, x_2, \dots, x_n "?

If these " x_1, x_2, \dots, x_n " are good, we get good inferences. $x = \text{value of } X$

If these " x_1, x_2, \dots, x_n " are bad, we get bad inferences.

Entire success of statistical tools depends upon the outcomes and the outcome depends upon the quality of sample used in the analysis.

Handwritten notes on the slide include:

- X Height of persons
- 1st person 150 cm $\equiv x_1$
- 2nd person 155 cm $\equiv x_2$
- ...
- nth person 165 cm $\equiv x_n$
- Population: Boys M M ... M, Girls F F ... F
- Sample: M M ... M, F F ... F

So, now the question comes over here very simple. We are going to work on a sample of data which is usually denoted as x_1, x_2, \dots, x_n . Let me first give you an idea what do you really mean by this x_1, x_2, \dots, x_n . Suppose, I try to take a variable here x which is the height of persons. Now, you choose some persons and you try to observe the height. Suppose you take the first person and measure its height suppose this is 150 centimeters.

And let us denote by here x_1 . Suppose now you take second person you measure the height of the person suppose this comes out to be 155 centimeters. So, this is going to be your small x_2 which is the height of the second person on the variable x . So, x is denoting the variable that is indicating x is height and x_2 is the value of the second observation on height. And this process can be continued till n th person and say the height of n th person is suppose 165 centimeter. So, this is going to be my x_n .

And you know that in statistics we always try to denote the variable or random variable by capital alphabets. For example, I have denoting here by x and the values of the variable they are denoted by say small alphabet say x is the value of x and this x_1, x_2, \dots, x_n they are denoting the first observation, second observation or say n th observation.

So, this is what I mean when I try to say x_1, x_2, \dots, x_n . So, obviously; once you obtain this x_1, x_2, \dots, x_n . Now, this is your sample after that you will not look back into your

population. You have only these n observations in your hand and based on that you will try to conduct all your statistical analysis, you will try to choose good statistical tools and you will try to take correct statistical inferences out of that.

So; obviously, if this selected sample values that is your x_1, x_2, \dots, x_n they are good, they are representative, they are indicating all the characteristics which are inside the population then; obviously, we will get good statistical inferences. And on the other hand if these x_1, x_2, \dots, x_n are not so good and suppose they are bad then; obviously, we will get bad statistical inferences.

Bad statistical inferences means they are not really representing the properties of population. For example, I can give you a simple example suppose there is a class of students and suppose there are some boys and there are some girls. So, boys are supposed to be indicated by male their gender is denoted by here male and girls are indicated by here F that is female gender. Now, this is my here supposed population. And from there, now I try to draw here a sample here and sample is supposed to be only here boys male, male, male, male, male, no F. Or I take another sample and if I try only the girls all F F F F F. So, is it a good sample?

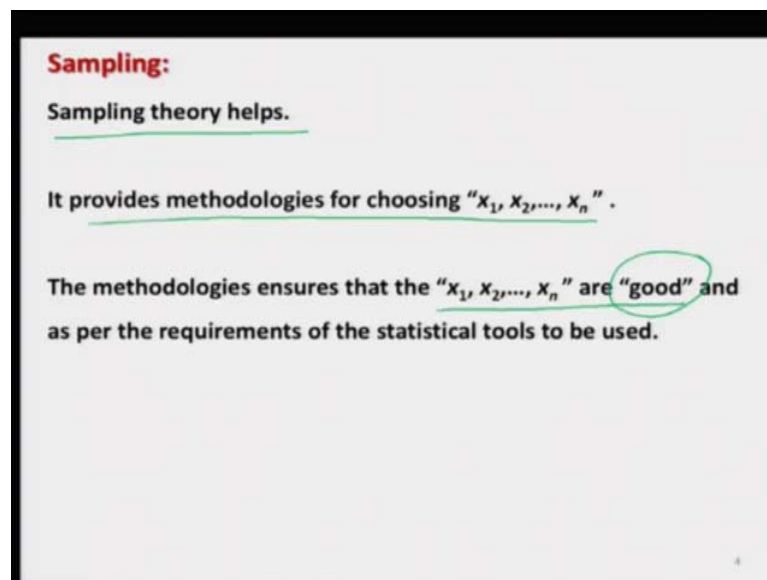
If you try to look in the sample number here 1 or the sample number here 2; that is by looking at this sample values only here you will get an idea that there are only boys in the there are no girls and vice versa if you try to look into the second sample over here you will get an impression that there are only female students in the class, there are no male students. This is what I mean by good or bad sample.

So, if you try to work with sample 1 which is consisting only of male students then you will get statistical inferences only for the male students which may or may not be valid for the female students. And similarly is for the sample 2 also, all the statistical inferences will be primarily indicating for the female students, they may or may not indicate for the male students, right.

So, the entire success of the statistical tool depends on the upon the outcomes and those outcomes are depending on the quality of sample which is being used in the analysis. So, now, at the end the moral of the story turns out to be that quality of the sample is very important for us.

And how should we ensure the quality of the sample? How should I draw the sample such that it is indicating all intrinsic properties of the population and all our statistical tool when they are applied over the data they are also giving us the correct statistical outcome right.

(Refer Slide Time: 17:41)



So, in order to solve this problem the sampling theory helps. And the sampling theory provides different types of methodologies for choosing the sample x_1, x_2, \dots, x_n . And these methodologies have been devised for different types of condition and if you choose the correct methodology for a correct situation then the sample drawn from that methodology will ensure that the random sample or the sample which you have drawn x_1, x_2, \dots, x_n this these observations are good.

When I am saying here good I have made it within the course. Why good? This good has several aspect. One aspect is that we have just discussed that they should be representative value; that means, whatever is my population that should be reflected in the sample also.

Beside those things one aspect is that which is very important aspect that once we get the sample, then we try to employ some statistical tool. The statistical tool is a mathematical function and this function is not coming from the sky.

But some people have used their mathematical knowledge and statistical knowledge to develop that tool. And every tool works under certain type of conditions those conditions are primarily originating from the mathematical need that when we try to develop a statistical tool we try to use the mathematical theories the fundamentals of mathematics and in order to solve them we use those tools.

Hence the tool what we develop at the end need to satisfy all the assumptions which are needed for the mathematical theory. So, now, the question is this when you are trying to employ a statistical tool over a data set. Then the data set must satisfy all those mathematical assumption or statistical assumptions.

So, when we try to draw a sample or when we try to define a methodology for the sampling, we always keep in mind that the basic assumptions of a statistical tool are satisfied and in case if these assumptions are not satisfied, possibly they may give you a wrong outcome or wrong statistical inference.

Possibly if the deviation is less, the final outcome may not really be affected too much, but in case if the deviation is high, the statistical inference will simply become invalid, they cannot be used, right. So, this is what I mean by the good sample. And when I go further I will try to take some more aspect which can be incorporated into the quote unquote good sample ok.

(Refer Slide Time: 21:30)

Sampling:
Sampling theory provides the tools and techniques for data collection keeping in mind the objectives to be fulfilled and nature of population.

These are two ways of obtaining the information

1. Sample surveys
2. Complete enumeration or census

So, now we have understood that sampling theory provides the tools and techniques for the data collection and it keeps in mind that what are the objectives to be fulfilled and what is the nature of the population.

Nature of the population for example, I have just taken an example here. For example, you can see here that in this example in this population there are male students and there are female students, but in this, but in sample number here 1, there are only male students and in this sample there are only female students, right.

So, the nature of the population is not being getting reflected in the sample. Now, in case if I want to obtain the information through the sample, then I have got two possible ways. There are two possible ways to obtain such information. One is sample surveys and another is complete enumeration or census.

(Refer Slide Time: 22:38)

Sampling:

- Sample surveys collect information on a fraction of total population whereas
- the information on whole population is collected in census.

Some surveys are conducted regularly like economic surveys, agricultural surveys etc.

Some surveys are need based and are conducted when some need arise, e.g., consumer satisfaction surveys at a newly opened shopping mall to see the satisfaction level with the amenities provided in the mall .

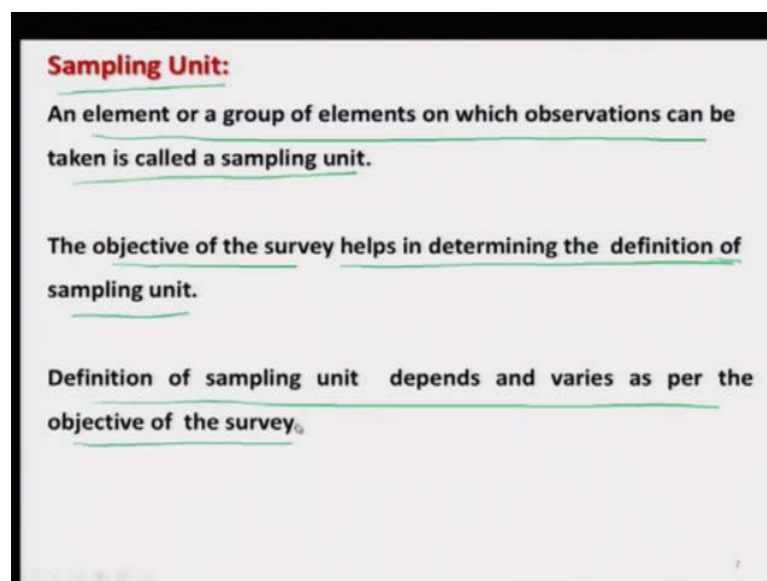
Now, what is the difference between the two? When I talk about sample surveys; so, the sample surveys we will collect information only on a fraction of the total population whereas, in census the information on the entire population whole population is collected. And you also know that there are some surveys which are conducted on a regular basis like as government of India conducts the economic surveys, agricultural surveys etcetera. We have different types of ministries, we also have Central Statistical Office, we also have NSSO National Sample Surveys Organization and these are the

offices of Government of India which conduct such a survey. And beside those surveys which are conducted on a regular interval.

There are some surveys which are need based. And they are conducted only when there is some need. For example, if a new shopping mall is opened then definitely the owner of the shopping mall would like to conduct a consumer satisfaction survey to know whether the customers are satisfied with the amenities which are provided in the shopping mall, whether customers are satisfied with the type of shops which are open in the shopping mall and so on.

There can be different types of queries. So, these type of surveys are going to take place only when there is a need. Means the possibly the owner of the shopping mall may not like to conduct the survey every month or say every year, but as soon as there is some change possibly he would like to have the opinions of the customer.

(Refer Slide Time: 24:45)



Now, I try to take a small definitions and I will try to explain them and it is very important for you to understand what are these thing beside the formal definition. So, first definition I am going to consider here is about sampling unit. So, first let me give you the definition and then I will try to explain you what is this thing and how it makes a difference.

So, an element or a group of elements on which the observations can be collected or taken is called a sampling unit. That means; this is a component of a sample survey on which we try to collect the observation. Because, in any survey the objective is that the survey helps again helps in fulfilling the objective of the study which can be done after determining the definition of the sampling unit.

Means, if you try to define the sampling unit in a different way, possibly that may give you some other information which is not included in the objective of the survey. And the definition of sampling unit depends and varies as per the objective of the survey, right.

(Refer Slide Time: 26:13)

Sampling Unit:

Example:

Objective: To determine the total income of all the persons in the household.

Sampling unit: Household.

Example:

Objective: To determine the income of any particular person in the household.

Sampling unit: Income of the particular person in the household.

So, now let me try to take some examples and try to explain you how this thing work. Suppose the objective of a survey is to determine the total income of all the person in the household. Then in that case you have several option that where should you collect your data, whose income you should collect in your data.

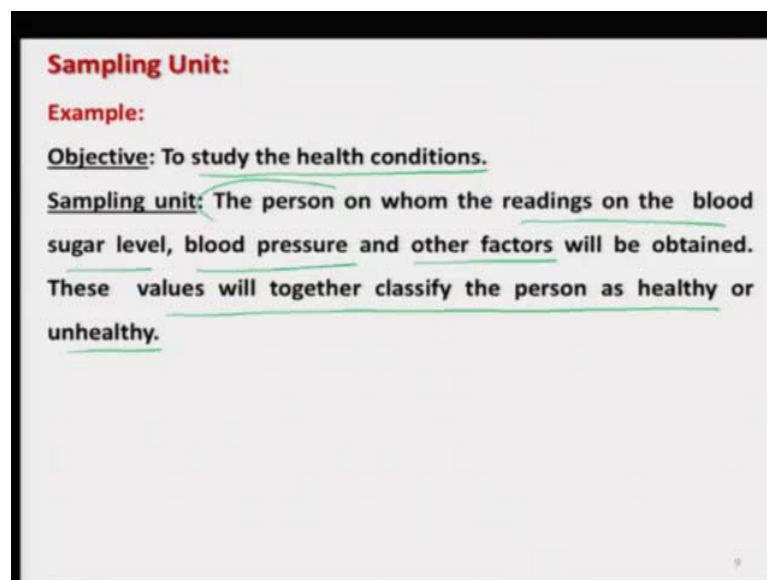
There are different persons there can be some younger children also who are who are earning, there is some elder people who are earning, there are some ladies who are also doing some house jobs and they are earning from home. So, the question is this, how would you define the total income; total income of whom, children, male, female, husband, wife or all?

So, now, I have to go with the definition of what is called a household. As far as I know the definition of Government of India says that the number of people or the group of people who takes the food from the same “choolah”, that means, the burner, right, they will create the household.

So, all such persons who come under the definition of household, their income have to be collected and then they have to be added together. So, that will give you the household income. So, my sampling unit will become household. On the other hand, in case if the objective is to determine the income of any particular person in the household, for example, I can say what is the income of the house lady who is possibly doing some job from home.

So, in that case the person who is working in offices outside home, their income is not going to be added, but I will try to collect the data only on that lady who is working from home. And in this case my sampling unit will become the income of that particular person in the household; for example, in this example; the income of the lady who is working from home.

(Refer Slide Time: 28:41)



Sampling Unit:
Example:
Objective: To study the health conditions.
Sampling unit: The person on whom the readings on the blood sugar level, blood pressure and other factors will be obtained.
These values will together classify the person as healthy or unhealthy.

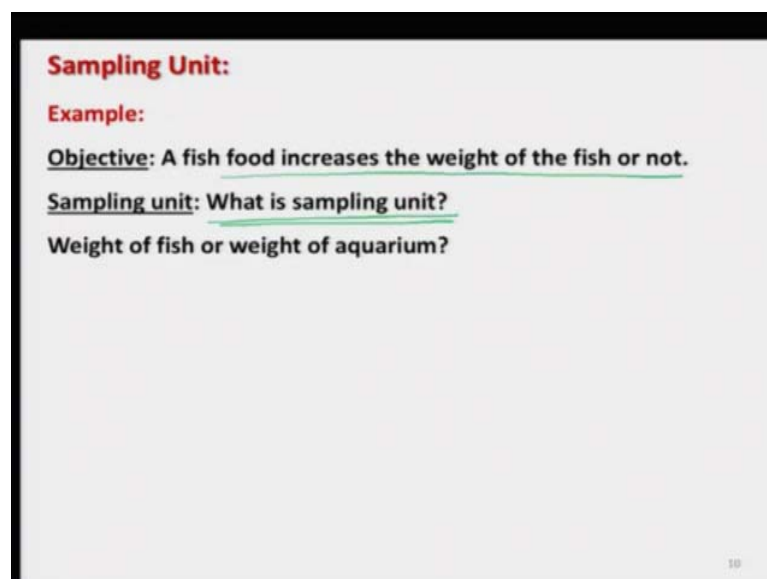
Similarly, in case if I want to study the health condition. Suppose the objective is to study the health condition in some colony, in some area, in some city or in some village

or in some country whatever you want. So, now, if you try to understand you want to study basically whether the persons are healthy or not.

But, now the question comes how would you define whether a person is healthy or not. Suppose if a person is under aged, but his blood pressure is good, blood sugar level is good, but he is under weight. Now will you call him as a healthy or her as a healthy? No. Generally, from a common sense I can say that a person is said to be healthy if all the body parameters are within the control limits.

For example, the readings on say blood sugar level, blood pressure, height, weight, age etcetera and other factors need to be obtained. And when all these values are clubbed together then jointly they will classify the person as healthy or unhealthy. So, in this case the sampling unit becomes the person on whom we are going to collect the data; data on the blood pressure, blood sugar level, height, weight, age etc. So, in this case my sampling unit is the person.

(Refer Slide Time: 30:19)



Sampling Unit:
Example:
Objective: A fish food increases the weight of the fish or not.
Sampling unit: What is sampling unit?
Weight of fish or weight of aquarium?

Now, I ask you a question. Suppose there is a food for the fish which is developed and the fish food claims that in case if you try to give it to the fish, then the weight of the fish will increase and you want to test it on the basis of sample.

So, in this case, my question to you all is; what is the sampling unit. You please give a pause to this video and try to think. How will you define the sampling unit? Your

objective is- you want to know whether the food which is being given to the fish is effective in increasing the weight of the fish or not.

So, what you will try to do? How would you conduct the experiment? You will try to give the food to the fish and then you will find out the weight of the fish after sometime, some weeks or some days, but my question is how you can get it done. Take a break, pause the video, think and then continue after at least 2 minutes, but I will continue with my answer.

Now, my bigger question is how to ensure that how much the food is being eaten by the fish. One possibility is this that you take an aquarium means aquarium is that pot where you fill the water and there are fishes inside it. So, there is a pot with the water and there are some fishes. You take out a fish in your hand, hold the fish like this, do some tickling the fish will start laughing fish will open the mouth and then you put the food into the mouth of the fish. And that is how you will come to know that that how much the fish has taken the food. Is it possible? Not really. I am sure you will simply laugh at this idea.

Second option is this; you take some known quantity of the food and sprinkle the food over the aquarium, but now the problem is this suppose you are sprinkling or giving 100 grams of food inside the aquarium. And suppose there are say 10 fishes inside the aquarium. How will you ensure that which of the fish has eaten how much of the food? And also some of the food will remain in the water, fish will not eat, it some part of the food will get dissolved in the water. So, practically you have no option, you have no medium to know that out of that 100 grams of food how much food a fish has eaten.

Now, what to do? In this case, what will be your sampling unit? Can it be a fish? Possibly now by this time you must be convinced that fish cannot be here as the sampling unit. But, then again the question is what is sampling unit; means if you ask me I will say my aquarium is the sampling unit. Because, what I am doing? I will take the aquarium, I will take the fishes into it and then I will weigh the entire aquarium.

And then I will keep on giving the food say for about a week or 2 week and I will maintain the same level of the water and after 2 weeks I will try to weigh the aquarium and if there is an increase in the weight of the aquarium then that can be contributed as if

the weight of the fishes has increased and which is due to the food which has been given to the or which has been sprinkled inside the aquarium.

So, in this case in my opinion, the sampling unit is not the fish, but this is aquarium. Now, believe me when you come to a real life experiment, how to define the sampling unit is not an easy job people will struggle. But it is not that difficult also. You have to keep in mind what is your objective what you really want to know.

And if you want to know something you have to collect the data, then the next question comes- on which item on which thing, on what you are going to collect the data; and once you decide that becomes your sampling unit, as simple as that, right.

So, you will try to understand this thing, you try to means consider these thing inside your brain, try to think about them, try to take some more example and try to decide yourself that what is the best way to choose the sampling unit in that experiment. And I will see you in the next lecture with some more topics till then good bye.