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 – 12
Conducting Surveys and Ensuring Representativeness

Hello, welcome to the course Essentials of Data Science with R Software, part 2, where we are trying to learn the topics of Sampling Theory and Linear Regression Analysis. In this module, we are going to talk about the Basic Fundamentals of the sampling theory with R Software.

So, you can recall that in the last couple of lectures, we started discussion on the basic fundamentals and basic definitions which are related to the sampling theory.

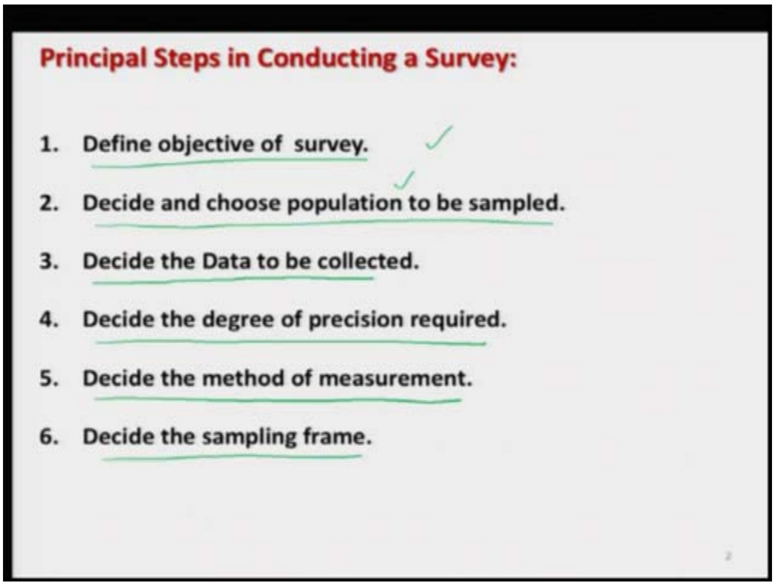
Now, in this lecture also I will continue with those topics and in this lecture, I would like to explain you that if you want to conduct a survey, then what are the different steps, means how you have to collect the data, what are the different ways to collect the data and what is your objective. The main objective is that one has to collect the data in such a way such that the variability is as minimum as possible.

So, how to ensure it? So, these are the few things which I am going to address in this lecture. So, let us start our lecture.

So, first thing which I am going to address is that what are the different steps in conducting a survey? Well, these steps are going to be like a story telling, but these are very useful when you want to conduct a survey in a real life. Well, there are several steps, but I would like to make it clear that these steps actually are not independent.

They are inter-related; I mean the first step will also indicate that what is to be done in the next step and the and something what is to be done in the further step that has to be taken care in the first step. So, let us start this thing and I will try to explain you, well, ok.

(Refer Slide Time: 02:22)



Principal Steps in Conducting a Survey:

1. Define objective of survey. ✓
2. Decide and choose population to be sampled. ✓
3. Decide the Data to be collected.
4. Decide the degree of precision required.
5. Decide the method of measurement.
6. Decide the sampling frame.

So, what are the principal steps in conducting a survey? The 1st step is you have to define the objective of survey. This is very important that first you need to decide what you really want to find out from the survey and try to write down very clearly what are the different questions which you want to answer.

And, many time people start the survey, they just collect the data and after collecting the data they come and they ask what type of information can be obtained from this type of survey. Well, those things are not really correct. So, first you have to define the objective of the survey; that means, you simply have to define different questions which you want to address. Now, in the 2nd step you have to decide and choose the population from where you would like to take the sample.

Now, you can see that step 1 and step 2, they are interrelated. Whatever are the objectives which have been defined in step 1, they will basically decide that what should be the population and what type of sampling scheme is to be used and what to be sampled.

Now, in the 3rd step, one has to decide what is the data which is to be collected. For example, if the objective of the survey is to find out the height of the students, then in the second step, we are going to draw a sample of students from the population of students;

that means, all those student who are the students of a university or a college. Then in the third step we have to find the or we have to record the height of the student.

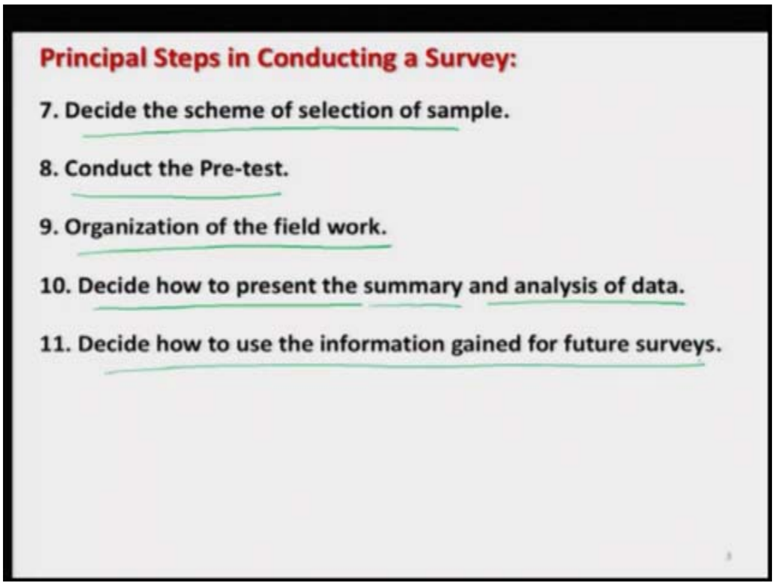
Sometime it happens that the objective is to find out the height of the student and people try to collect the data on the weight of the student, that should not happen these type of things. Now, after this we have to decide what is the degree of precision what is required. For example, if I take the same example if the height has to be measured, we have to very clearly indicate that it is in terms of meter, centimeter up to what degree of precision.

For example, if you want to collect the data on age, then one has to very clearly specify that the age has to be in terms of number of completed years or number of years number of months or number of years, number of month, number of days and so on. Or it is or is it going to be in terms of number of days whatever is the objective based on that whatever is the requirement that has to be mentioned very clearly before the survey starts.

Then in the next step, we have to decide what method of measurement we have to adopt; means, how are we going to record the observations. You want to find out the age of the person suppose, then would you like to record the age or you would like to record the date of birth and based on that you have to decide in your questionnaire that how are you going to record the data. Then you have to decide the sampling frame.

Sampling frame means in the step number 2, you already have decided that what is the population which has to be sampled. Now, you would like to identify the sampling units and give them a particular tag, a particular address so that the same sampling unit is collected in the sample which is drawn by some sampling technique.

(Refer Slide Time: 06:22)



Principal Steps in Conducting a Survey:

7. Decide the scheme of selection of sample.

8. Conduct the Pre-test.

9. Organization of the field work.

10. Decide how to present the summary and analysis of data.

11. Decide how to use the information gained for future surveys.

After this you have to decide which scheme you want to opt for selection of the sample. There are different types of sampling schemes simple random sampling, stratified sampling, cluster sampling, two stage sampling, etc.

So, depending on your objective, depending on the type of population and depending on the environment and the condition in which you are going to sample decide the sampling scheme. Every sampling scheme has some conditions and every sampling scheme works under a special type of environment, a special types of requirement. So, you have to choose.

Once you decide all these things so, you are basically ready to start the survey, but do not go to the field, but conduct the pre-test. Pre-test means before going into the field, try to conduct your survey as if you are in a real field. For example, if you have prepared a questionnaire, try to give the questionnaire to your colleagues to your friend or to people in your office and ask them to fill.

It is possible that they may face some problems in giving the answers to different questions which are written inside the questionnaire, then you have to revisit them. You have to see that how you can modify them, so that the answers to those questions are clearly obtained.

This is very important, suppose if you do not conduct the pre-test and if you go directly to the field then it is possible that the surveyors may get stuck or the people do not understand the question, they will try to give the wrong information and which will result into wrong statistical inferences.

So, it is important that you try to conduct the pre-test and then based on the feedback which is received from the pre-test you try to modify your questionnaire your sampling scheme or whatever type of this drawbacks you observe. Once you are completed with this pre-test, now you are revise your questionnaire and other steps, now you are in a condition to enter into the real field.

But, before going into the real field you have to look in the organization of the field work; that means, how are you going to work when people are working in a real field. For example, if I take an example suppose if the government if of India wants to conduct a survey at a so, the main offices are located in the capital of the city say in New Delhi. Now, they have to administrator administer the survey in the entire country. So, what they can do or what they have actually done?

That they have their offices in different states in the capital cities mainly then they have offices in different cities then I mean they can also have some small setups in different locations. Now, whatever they want to get it done from the main office they will send it to the states the states will understand those requirement and they will try to divert it to different cities and then that may go to different districts, different villages and so on.

And, at every step they have to very clearly define the role of the investigator or role of the persons who are involved in the survey, means everybody is not allowed to take a decision in case of any confusion. It has to be very well defined that who is going to address problem at that particular level and based on that, the reverse process happen. For example, the data is collected into the villages that is submitted to the district. The district will compile the data of all the villages, then that will come to the city. The city will compile the data of different districts and then city will send it to the main office in the state. The state will compile the city of the data from all the cities and then will finally, transmit it to the main office. So, these type of organizations are required for a successful conduct of the survey.

So, and then once you have this data then it has to be defined before the start of the survey that how are you going to present the summary and how are you going to analyze the data. Now, you can see once you come to this step, then all the previous steps are interrelated. For example, in the analysis if I want to have only the frequencies or if I suppose if I want to conduct some test of hypothesis, then the requirement of the statistical tool will be different.

For example, if the requirement of the statistical tool is that the data has to be in the quantified way, then this has to be immediately into the questionnaires and they have to be mentioned very clearly that what should be the precision of the data because that is the quantified data height say 160 centimeter or 160.2 centimeter or 160.23 centimeter, whatever it is.

Means if you want to conduct a statistical tool whose requirement is the data has to be quantitative, but if you have collected the data which is qualitative then you are stucked and this happens in real life. So, you can see first you fix your objective, based on that you have to decide that what you really want to find and then what statistical tools are going to be utilized.

Then whatever are the requirement of the statistical tool, based on that you have to; you have to decide for the variable, then you have to decide whether the variables are going to be qualitative, quantitative or indicator variable, whatever it is.

Then those types of variable have to be embedded into the questionnaire, so that the answers of those questions are going to give you the same data set which you are going to use at the end for the analysis, ok. So, once you do this thing then possibly you will be able to conduct a good survey.

So, the next step is this that we need to decide how the data has to be present, presented, how the summary of the data has to be presented and how we are going to analyze the data, what type of statistical tools are needed.

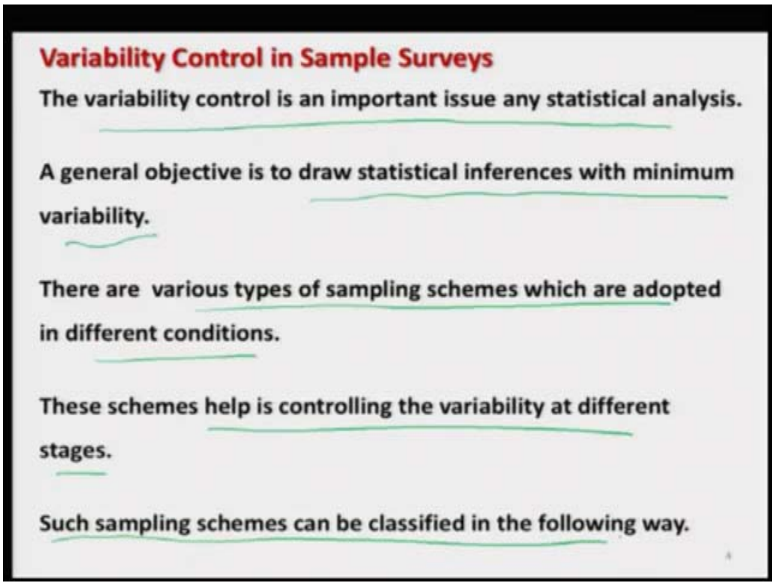
And, this is also an objective that once somebody is trying to conduct the survey, they are trying to collect different types of data, they are trying to collect different types of information if that information can be collected in such a way such that it is also useful in near future or to fulfill some other objectives.

So, these are the main principle steps which one needs to follow in conducting any samples survey. Well, these steps look very simple, but when you really try to conduct a real survey, different types of problem come, they will crop up and you have to handle them. What I have given here these are the basic steps, but depending on the environment, depending on the objective, depending on the type of survey, depending on the place of survey, there can be different types of things.

For example, if in a country like India, if people are conducting the survey in different states, they have got different types of language, the same vegetable which is being called as name X might be called as name Y in say in another state or name Z in different state. The unit of measurement of a variable may also vary. For example, say oranges at some places they are sold in numbers by counting 1 dozen, 2 dozen at some places they are sold by weight.

So, if you really want to have the information on the oranges, you have to be very careful that how are you going to measure it, how are you going to prepare the question, so that the question is clearly understandable by the surveyor as well as by the respondent.

(Refer Slide Time: 14:29)



Variability Control in Sample Surveys

The variability control is an important issue any statistical analysis.

A general objective is to draw statistical inferences with minimum variability.

There are various types of sampling schemes which are adopted in different conditions.

These schemes help is controlling the variability at different stages.

Such sampling schemes can be classified in the following way.

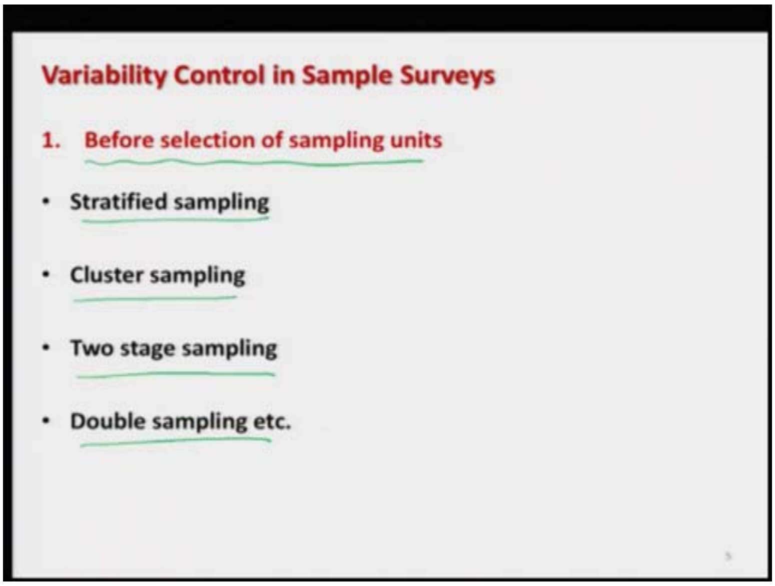
Now, when we are going to conduct the sample survey, one of the main objective in any sample survey is that the data has to be collected in such a way such that the variability which is going to occur when the data is used in the statistical tool has to be as minimum as possible. So, control of variability is an important aspect in any statistical analysis.

This variability can be controlled at different steps, for example, I can use a statistical tool which will give me a statistical inference with lower variability, but at this moment we are trying to conduct the survey. So, in case if I start controlling the variability right from the beginning then it is expected that at the end the variability is going to be as minimum as possible.

So, the first step where you can control the variability is the survey or the sample survey where you are trying to collect the data. So, the question is that how we can do it in sampling theory, ok. So, now the variability control is a very important issue at any statistical analysis and we would always like to have a statistical inference which gives us the result with the minimum variability.

And, in order to control the variability through the sample surveys there are various types of sampling schemes which are adopted under different types of conditions. And, these schemes essentially help us in controlling the variability at different stages. Now, what are those stages? Based on those stages we have classified the different sampling scheme.

(Refer Slide Time: 18:21)



Variability Control in Sample Surveys

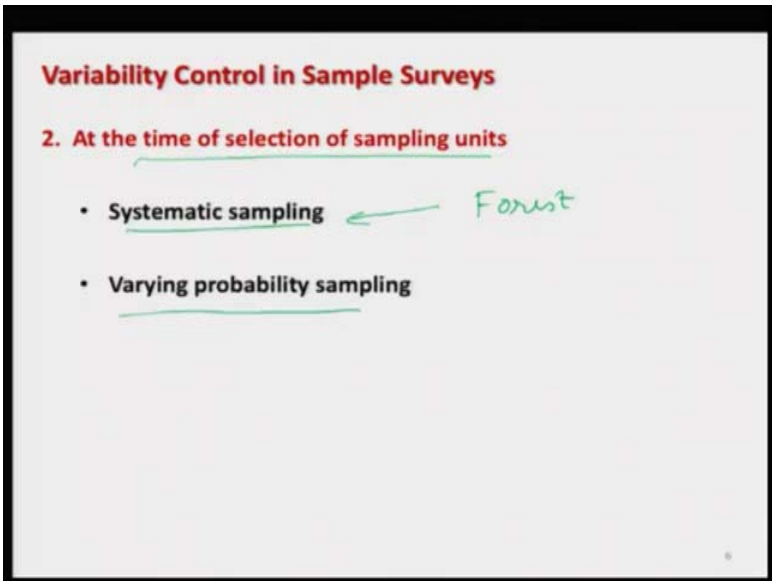
1. Before selection of sampling units

- Stratified sampling
- Cluster sampling
- Two stage sampling
- Double sampling etc.

For example, there are three stages the variability can be control before the selection of the sampling unit, after the selection of the sampling unit and during the selection of the sampling units. So, when I try to control the variability in the sample before selection of sampling units, then there are different types of sampling schemes which help us in doing so.

For example, there are, there is a stratified sampling scheme, there is cluster sampling scheme, there are there is two stage sampling, there is double sampling and so on. There are multi-stage, multi-phase sampling schemes which help us in controlling the variability in the data, right. What are those thing that is the topic we wish will come in the further lectures.

(Refer Slide Time: 19:15)



Variability Control in Sample Surveys

2. At the time of selection of sampling units

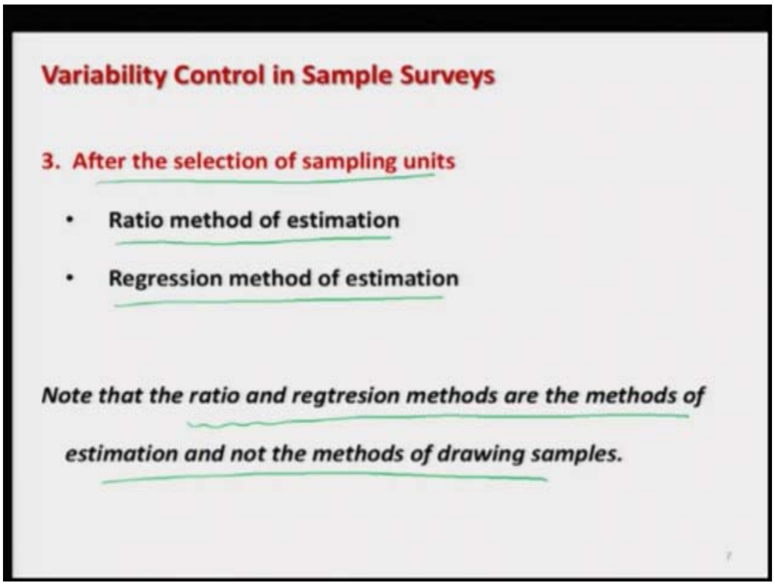
- Systematic sampling ← Forest
- Varying probability sampling

Now, sometimes the variability can also be control at the time of collection of the sampling units. For example, the systematic sampling varying probability sampling, etc. they help us. For example, systematic sampling is used in a condition like forest, right. Suppose you want to see what is the amount of the forest wood or the amount of wood of some particular trees.

Now, inside the forest you cannot create the sampling frame because the sampling frame has to be created before the start of the survey. One cannot enter into the forest first count the number of trees, indicate the number of trees, create the sampling frame and then start sampling that is not possible. So, under those situation what somebody has to do?

The person has to enter into the forest and then with some pre specified rule, he have to he or she has to identify the trees and then the required measurement have to be taken. So, similarly varying probability scheme is also there that assigns say different types of weight to the sampling unit depending on their size or probability of selection and so on.

(Refer Slide Time: 20:26)



Variability Control in Sample Surveys

3. After the selection of sampling units

- Ratio method of estimation
- Regression method of estimation

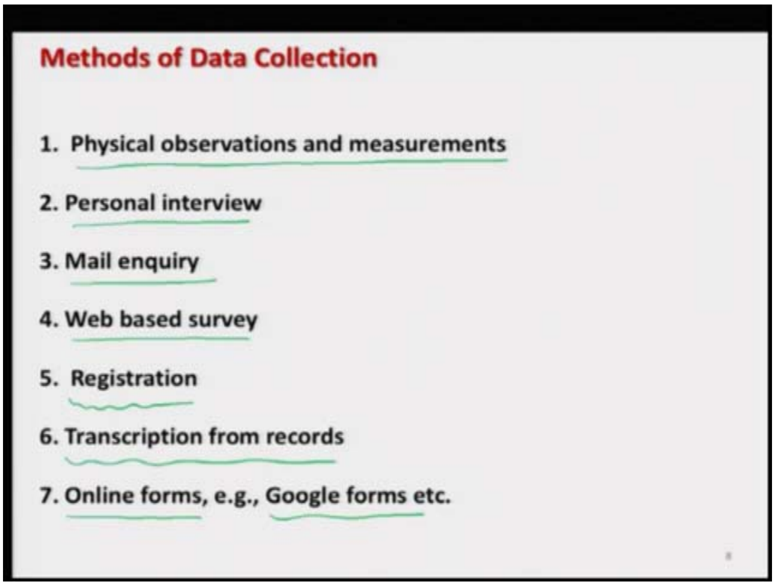
Note that the ratio and regtresion methods are the methods of estimation and not the methods of drawing samples.

And, 3rd option is this once you have collected the sample, then we always try to estimate some parameters. Those parameters can be population mean, population variance and other things.

So, once we have got the sample then which of the statistical tool can be used which can give us the statistical inference with lower variability. And, under this clause there are two popular methods in sampling theory which are very popular one is ratio method of estimation and regression method of estimation.

One thing here I would like to make very clear that sometime people try to use the terminology like as ratio and regression methods of sampling, but these are not the methods of sampling actually. Actually they are the methods of estimation of parameters. They are not the method for drawing the sample. These are not the samples sampling scheme these are the estimation methods. So, this is what you have to keep in mind.

(Refer Slide Time: 21:25)



Methods of Data Collection

1. Physical observations and measurements
2. Personal interview
3. Mail enquiry
4. Web based survey
5. Registration
6. Transcription from records
7. Online forms, e.g., Google forms etc.

Before I move further, let me try to tell you here that as I said in the beginning that sampling theory has lots of topics and this is a big course, but here in this course I am not going to cover all sorts of sampling scheme. I will try to cover only couple of sampling schemes and my objective is to give you the basic fundamental and to train you so much so that if you want to learn the remaining sampling schemes, you can learn it yourself, ok.

Now, the question comes, in a typical survey how would you collect the data? There are different ways by which we can collect the data. So, I will try to give you here a brief introduction about those methods now that will be your choice that which method you would like to use and that depends on the objective, that depends on the environment, that depends on the requirement and practically that is a decision which the surveyor or the person who is conducting the survey has to take, ok.

So, there are different methods of data collection the first one is physical observations and measurement; that means, whatever is the variable on which the observation will be required. For example, if the variable is height or weight, then the height or weight of the persons have to be recorded on those using some tool for example, weighing machine or the scale to measure the height and so on.

Second option is personal interview that the surveyor will contact the persons and they can contact now our phone or something like that or they can personally visit the person and they will ask the question. The person will reply the question and the interviewer or the surveyor will fill up the questionnaire or in some cases means the person can give the questionnaire to the respondent. And, in consultation with the surveyor the respondent can fill up the questionnaire and can give the data.

Third thing is mail inquiry that can be an electronic mail or a physical mail or a postal mail. So, the questionnaire is sent to the people and they try to answer the questions. Nowadays, web based surveys have become very popular. There are some websites which helps in creating the questions and their possible responses. The advantage is that whatever are the response responses they are automatically compiled and the time to compile the data is saved.

Sometime the data is collected from the registration forms. For example, if there is a contest the then different people want to participate they are asked to fill up a registration form in which they will give all their personal details and then from there we can collect the data, right.

And, then there are some transcription from the record for example, we have municipality offices, there are different types of government offices where various type of information's are available and if one needs to go there and then from those records one can draw or record the required data set.

Nowadays, some online forms are available are also available for example, Google forms they are one of the most popular things nowadays. Similarly, it is not only Google, but then different sites also help us in creating a different types of form they can I mean they can be sent to the responder, through E-mail. And, a link is actually sent by clicking on the link, they come on the website and they try to answer all the queries and based on that their data is recorded.

So, now these are some possible ways. Now, another very important source of data collection are different types of cards. For example, I had discussed this thing in the last lecture also. Some shopping store, some different types of a chains, office stores, they have started issuing a card and they give you some incentive. For example, if you buy something for 100 rupees they will give you 1 point, 2 points, 3 points, 4 points and they can be redeemed with some gifts or with some cash or kind something like this.

So, now what happens? You go to the store, you try to from form a very big store, you try to buy something and whatever you buy that will indicate your choice and preference and your requirement also. Now, once you come to the counter to pay the bill they will ask you do you have a card? And if you say yes, they will swipe your card and those amounts.

And, entire detail whatever you have purchased today all the details will be entered in the database of the card and the database of the card is being managed by some other company which is trying to collect the data. And, they are also not giving they are also not collecting the data from us for free for that you are given some points which you can reimburse.

So, but this is one of the say cheapest way to collect the data and now using that data, they can analyze the pattern of your shopping, they can analyze what type of requirement do you have, based on that sometime they will send you offer to attract you to the store for doing more shopping and so on. For example, if they say if they see from the pattern that this particular card makes a good shopping on clothing from this store.

So, they can sometime they can give you an offer that if you purchase the clothing for say X amount of rupees, then you will get this offer some more discount and that will help them in increasing their sales. So, that is another way of collecting collection of the data.

So, this is how the data is collected in a real survey. Now, I have given you almost all the basic backgrounds different types of definitions, concepts which are related to the sampling theory. Now, from the next time I will try to take up the sampling schemes and I will start with simple random sampling.

And, then I will show you that how the data is collected or how the sample is collected and when we are trying to implement a scheme, how it is ensuring that the statistical principles or the basic assumption of statistics are also satisfied. And, remember one thing they are not going to happen automatically. You have to conduct the process in such a way such that those assumption get satisfied.

So, you revise your basic concept and I will see you in the next lecture. Till then, good bye.