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

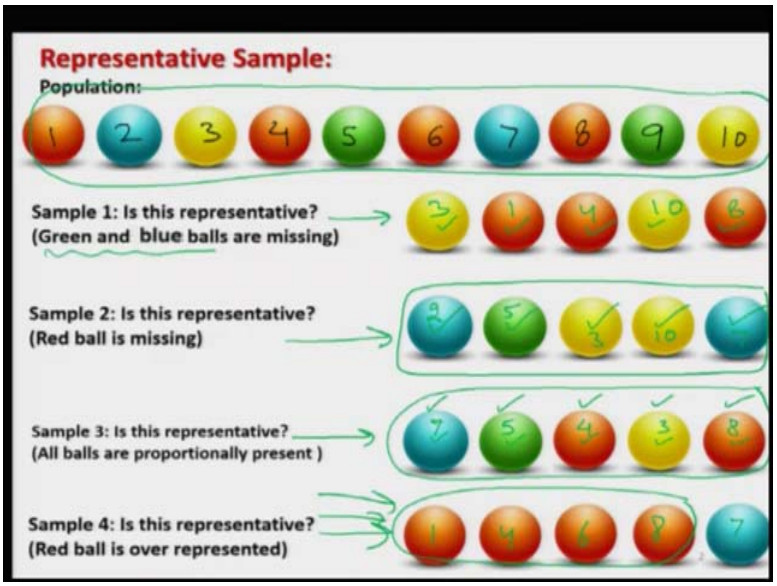
Sampling Theory with R Software
Basic Fundamentals
Lecture – 11
Ensuring Representativeness and Type of Surveys

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 continue with the different topics and different Basic Fundamentals of the Sampling Theory with R Software.

So, you may recall that in the last lecture we had talked about several concepts and we consider the topics of representative sample, sampling frame, population etc. Well, that is my experience that whenever I teach this concept of representative sample in the class, usually what happen that when I go next day in the class then student usually ask some more question and they want a couple of more examples.

So, I thought that why not to start this lecture once again with the representative sample and I will just take one or two more example to make you comfortable with the concept of representative sample and then I will move forward with some new more topics, ok. So, now let us start this discussion.

(Refer Slide Time: 01:33)



And, so, you remember that on the last term we had talked about representative sample. So, now, I am taking here similar example. Suppose I take here a population of different colours of ball say 1, 2, 3, 4, 5, 6, 7, 8, 9, 10. So, now, remember one thing when I am selecting these balls, this is my here population. So, obviously, this is not a sampling frame. If you want to convert them into a sampling frame then a better option is this.

Let me write here the number 1, 2, 3, 4, 5, 6, 7, 8, 9 and here 10. Now, this becomes a sampling frame, but anyway I am not considering it as a sampling frame, I will simply take it as a sample because my idea is to tell you about the representative sample. So, now I try to take here four different types of sample.

So, this is my sample number 1. So, you can see here there is a yellow ball there is a yellow ball. So, there are two yellow balls and there are 1, 2 and 3 red balls. Do you think that is it a representative sample? Well, you can see here that green and blue balls are missing.

So, well this is not a very good representative sample or I will say that the degree of representativeness is very very low. Well, that can be a sample. There is no doubt there is no problem in calling this as a sample because means I can always choose because in my sample. For example, this is my ball number 3 and ball number here 10 and this is my ball number 1, 4 and say here 8.

So, this is a sample, this and this can be a random sample also. There is no issue, but the question here which I would like to consider here is it representative or not. So, answer is not really ok. Now, I take sample number 2.

So, you can see here there are two blue balls, 1 and 2; there are two yellow balls and one green ball, but there is no red ball and this can also be a sample. There is no doubt for example, this can be ball number 2, this can be ball number 5, this can be ball number 3 and this can be ball number 10 and this can be ball number 7. So, obviously, this is a legally, this is a sample.

But, it is not a representative sample because red ball is missing. By looking in this sample you cannot guess or you cannot conclude you cannot infer that the population also has a red colour ball. Now, if you try to look in this sample number 3, there is one blue ball, one green ball, one two red balls, yellow ball. Is it representative?

Yeah, that looks actually nice; because you can see here, there are two blue balls in the population and you have selected here one ball, this can be ball number 2 or ball number 7. Let us call it ball number 7. There are two green balls in the population. So, you have selected here one green ball in the sample. So, this can be either ball number 5 or 9. So, let us take it to be 5.

And, similarly, there are four red balls. So, you are taking here two red balls in the sample. These balls can be say ball number 4 or ball number here 8 whatever you want to take and there are two yellow balls in the population. So, you are trying to take here one yellow ball which number say 3 or 10.

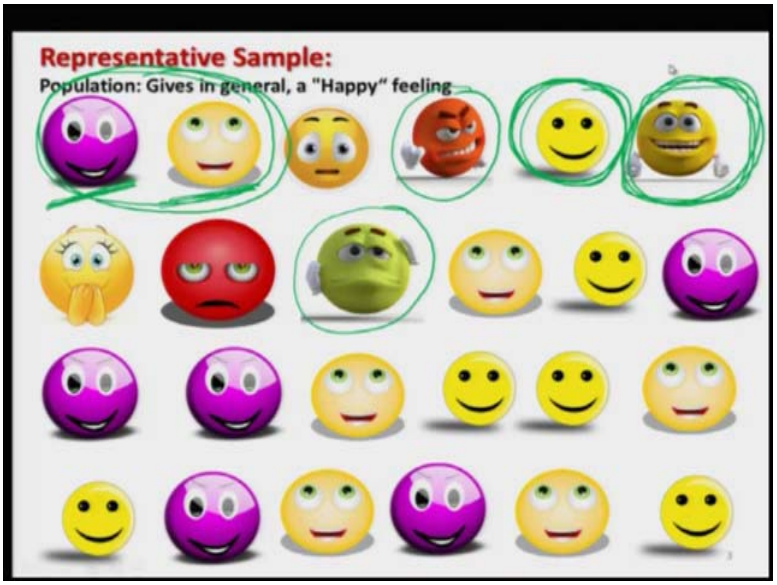
So, you can see here in this sample, all colours of balls are present and it is also indicating that if there are more numbers of balls in the population which are red, so, the sample also has more number ball which are red. So, this can be considered as a good representative sample. Why? Because, all the colours of balls are proportionately present in the sample.

Now, if you come to the sample number here 4, you can take here there are four red balls. So, they can be ball number 1, 4, 6 and here 8 and there is one blue ball, let it be ball number 7. Would you call this as a representative sample? Well, legally once again

this is also a sample because all the balls are coming from the population, but red ball here is over represented and blue, green and yellow ball they are not represented at all.

So, by looking in this sample it gives us an impression as if the population consist only of blue and red balls and red balls are much more than the number of balls of blue colour. So, this is about the representative sample.

(Refer Slide Time: 06:49)



Let me try to take here one more population to give you an idea of the representative sample. The difference between this example and this example is that you can see here all the balls are identical, say this ball, this ball, this ball and this ball they are identical and in fact, all the balls are identical, they are differing only with respect to the colour.

But, here you can see I have taken here different types of smileys and you know that by looking at this smileys; you can infer different types of information. For example, what do you think what is this representing? This is smiling. So, yeah, it is indicating that the person is happy.

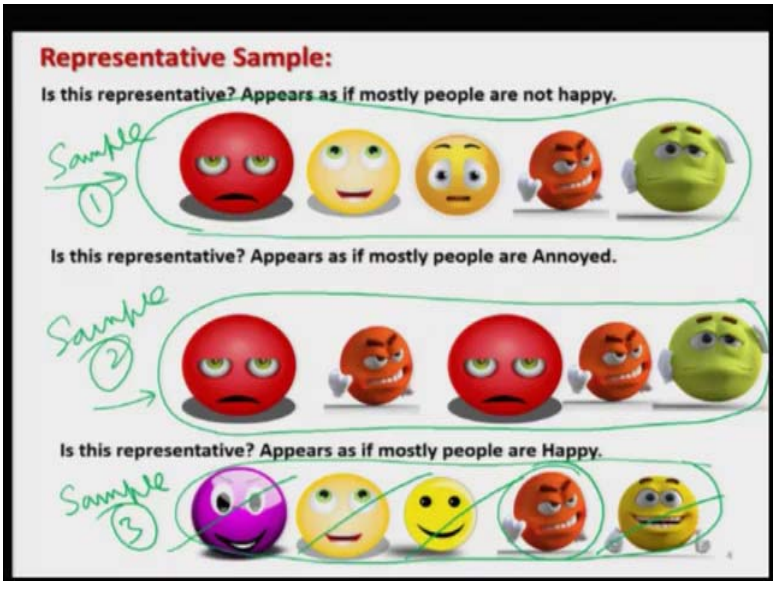
If you look at this smiley this is also indicating that the that the person is happy, but you can see both these first and second units they are indicating that the person is happy as the smileys are indicating a happy face, but they are different, right. Similarly, if you come here on say this is one, this yellow one also where I am making a circle this is also indicating a happy smiley, right.

So, you can see here that there are different types of faces, which are indicating the happy smiles. There are some faces which are saying that the person is not really happy. For example, this one, this is indicating that the person is not really happy. This is also indicating that the person is surprised and you can see here there is here one more smiley which is saying that it is very happy, right.

So, you can see here that there are different smileys and these are all together different, but different smileys are also indicating the similar information. So, my objective here is I want to consider the smiles or the facial expressions to indicate the satisfaction level for example, right.

So, earlier I had taken all the balls which were identical they were differing only with the colour, but now I have got different expressions, the faces are different smiling's are different, their expressions are different, but they are conveying a similar type of information.

(Refer Slide Time: 09:19)



So, now you can see here that I am trying to classify them into different groups. For example, if you try to classify all these things in one group all these five smileys, five types of smileys. So, they are indicating as if the people are not very happy, right either they are not happy or they are surprised or they are sad.

So, if I try to take a sample of five smileys from this population and suppose, my sample is like this let me call this sample number here 1. So, looking at this sample it will indicate as if people are not very happy, they are not satisfied, ok. So, is it really representative? Not really.

And, if you try to take here the sample number 2 which I am taking here, so you can see here means all the faces they look as if they are not happy, they are annoyed with something. So, suppose if I say that these faces are indicating that how are the medical facilities in the hospital, the same example which I took in the last lecture, then these faces in the sample number 2, they are indicating that as if the medical facilities are not good and people are not happy with them, they are really annoyed.

But, on the other hand, if you try to take here the sample number if here 3 so, you can see here this 1, this 2 and this 3, they are, and this is a 4; they are indicating that the people are happy whereas, this one which is the orange one that is indicating that the people are not so happy, they are annoyed. So, looking at this here sample number 3, in this five smileys it is giving us information or it appears as if most of the people are happy, right. So, this is how we try to represent a representative sample.

And let me honestly accept that I am taking here different types of faces to represent the same behavior or the same feeling which is representing actually the variability in the units, right.

In general, in practice all the sampling units in the population they are not really going to be identical, but there will be some small variation in those values and through this example, I am trying to convey that the variability will also be there in the sampling unit and we have to take care of this. For example, here I have taken care by indicating them by different smileys – even the happy smiley is represented by say couple of faces, ok.

So, next question comes that how we have understood what is the concept of a representative sample. So, then the question is how are we going to ensure that the sample which we are going to draw using any sampling technique will result in a representative sample. Or in other words, what type of methodologies we have to opt so that the resulting sample is a representative sample.

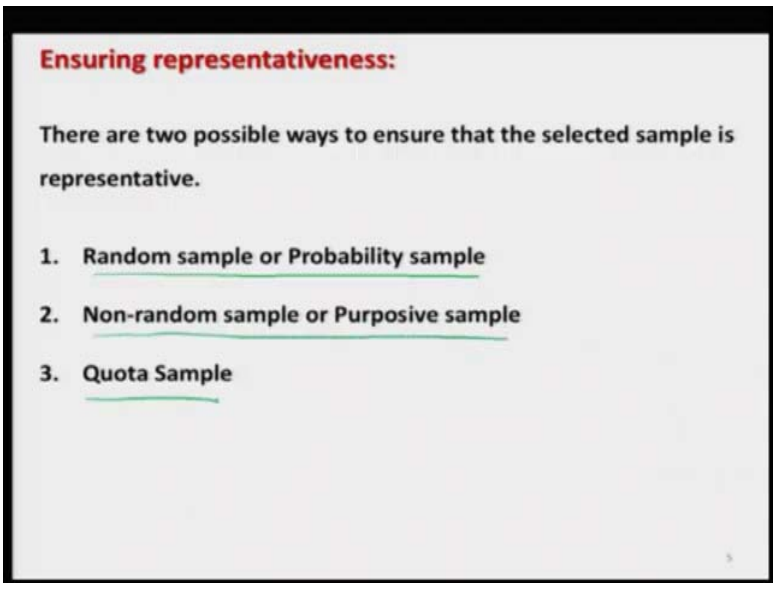
Beside this thing, there is one important aspect which I would like to address here that is related to the statistical tool. Whenever we try to use any statistical tool, then in most of the cases you will find that we assume that the sample is a random sample.

So, the question is what is actually here a random? So, when we say the word random, this means there are certain properties which are hidden or which are to be satisfied. For example, one aspect is that that whatever units we are trying to draw, they are independently drawn, they are identically drawn etc. The concept of independence identically distributed, they are used in the development of the statistical tool.

So, now the question is this if that statistical tool has to be used, then the condition for example, like random sample had to be satisfied by the data. So, now, the reverse happen that at this stage when we are trying to draw the sample, we have to ensure that we have to draw the sample in such a way so that it is random. So, the question here is this how we are going to ensure the representativeness in the sample and for that, there are different approaches.

So, I will briefly try to explain you here how we are going to achieve this concept through the sampling ok.

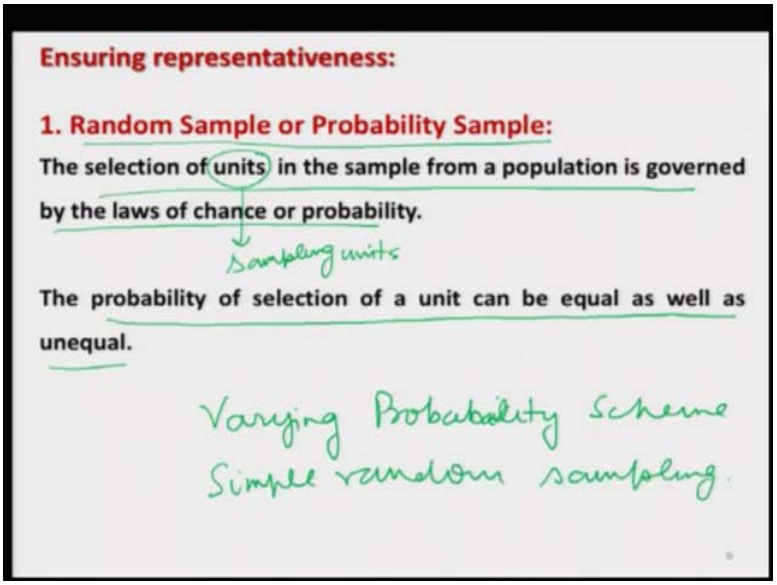
(Refer Slide Time: 14:32)



So, there are three possible ways which I am considering here which are needed to ensure the representative sample. 1st is or we call it also a probability sample; 2nd one is

non random sample or we call it as a purposive sample and 3rd one here is the quota sample.

(Refer Slide Time: 14:51)



So, I try to address them here one by one. So, when we talk of the random sample or we also call it as a probability sample, this means the selection of the units; remember one thing now, whenever I say units that it means these are sampling units and this will be our terminology in the remaining course, units means sampling units. I will not be writing sampling again and again, but that is imbedded in this word, ok.

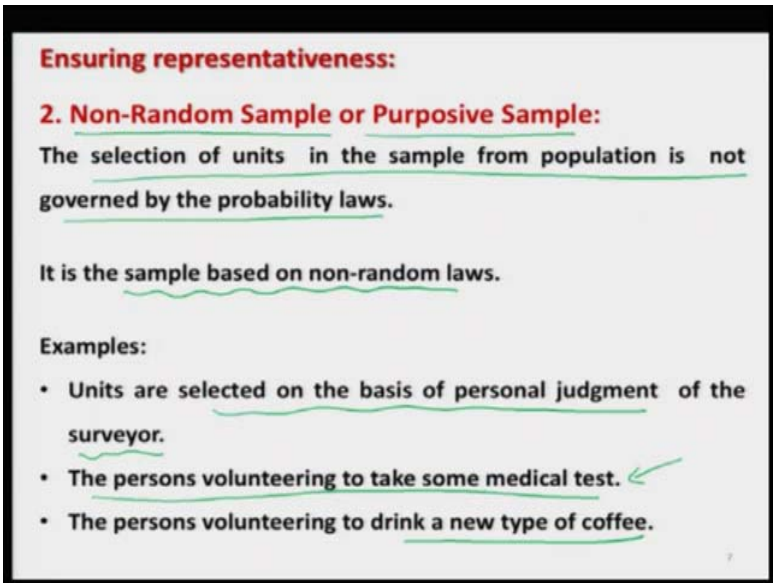
So, the selection of units in the sample from a population is governed by the laws of chance or laws of probability. You must have read about the probability theory and we want to ensure that when we are trying to construct the random sample for that we have to draw the sampling units, required number of sampling units from the population then those units are drawn in such a way such that the laws of probability are satisfied.

For example, the laws of probability does not mean that all the sampling units have to be of the should have the same probability of selection or not selection. The probability of selection of a sampling unit can be is can be equal or this may be unequal and based on that we have different types of sampling scheme.

For example, when the probability of selection of a unit are not equal, then we have a sampling scheme which is called as varying probability scheme, right. And, when we

have the equal probability of selection we have different types of sampling techniques like as simple random sampling, stratified sampling and yeah other things that we will try to consider them later in the course, ok.

(Refer Slide Time: 17:02)



Now, the second thing is this non-random sample or this is called as a purposive sample. So, in this case the selection of the sampling units from the population is not really governed by the probabilities law. This is just opposite to the case 1. But, the sample has been drawn on the based on the rules which are non-random laws.

For example, sometime the units are selected on the basis of some personal judgment of the surveyor. For example, the surveyor has been given some idea that what he has to look into the unit and suppose he goes to the field and he tries to and he or she tries to see whether that criteria is satisfied or not; if the criteria is satisfied in his or her opinion only then the unit will be selected in the sample.

Sometime the persons who would like to volunteer to take some medical test that is also not really a random sample, but it is a non random laws. Why? Because if there is a new medical test and in order to see the efficacy of the test, the test has to be conducted on some number of people and yeah, everybody would not like really to volunteer for the test because there is a risk that the health may have some problem, right.

So, in this case the sample is selected from only those people who are willing to participate in the medical test. So, this is also a non-random sample. Sometime if you go to a shopping mall, in some food store they try to offer some drinks to taste, right, but then everybody does not drink those stuff.

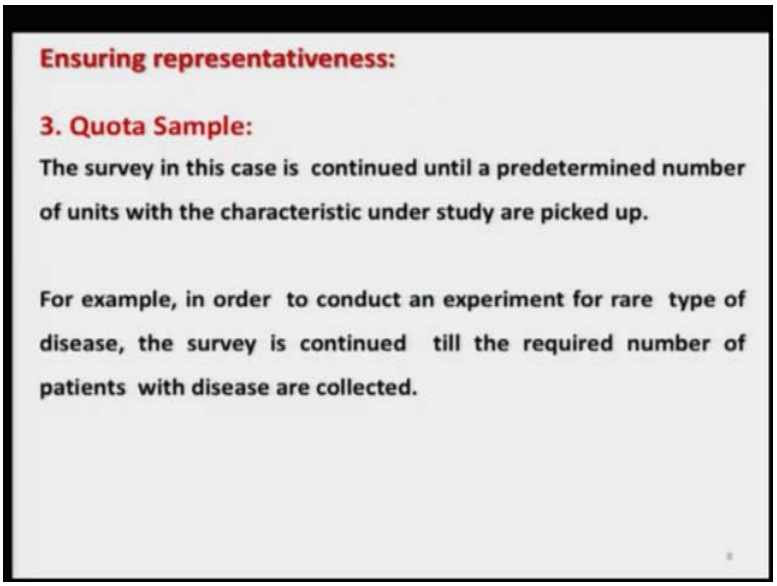
So, if they really want to suppose check, that how is the taste of a new type of coffee, then first they have to ask the person that would you like to drink the coffee or would you like to taste the coffee and the coffee will be tested by the person only if he or she agrees for it. So, this is also not really a random sample.

The random sample in these cases would have been something like in the case of medical test, they just select some patient by chance including the laws of probability. Suppose there are 100 patient and they select that patient number 1, 10 and 15, they are going to be given the medical test, then those patient have to begin without any exception and their choice has been made by some by some laws of probability, right.

Whereas in the case of non random sample, the person first has to say yes, I wish to participate. Similarly, in the case of coffee experiment means you just cannot decide yourself that who are the person who have to compulsorily drink the coffee, but first you have to ask their wish their decision and only then you can offer the coffee.

So, from that point of view from that angle this will be a sample, but this is not really a sample which is based on the laws of probability. And so, this is called as non random sample or purposive sample. Why purposive sample? As the name suggest, the sample has been collected with some purpose. Well, the purpose is also there when you are trying to select a random sample, but there is a difference that here the laws of probability are not being followed strictly, ok.

(Refer Slide Time: 20:47)



Third thing is quota sampling and this quota sampling is used in those type of surveys where it is hard to get the people who are carrying the characteristic or who have that characteristic. So, in this case the survey is started and the survey continues until a predetermined number of sampling units having the characteristic are picked up.

For example, if there has to be a medical experiment, a clinical trial, where they want to give the medicine for some rare type of disease, then definitely all the patients who are coming in a general hospital may not have that rare type of disease. So, in that case they have to ask if the person has disease and if the person is willing to participate in the study they will consider it.

And, they will continue it till the predetermined number of sampling units are collected and the difference is that in order to collect the predetermined number of sampling units, how many sampling units they have to draw, how many person they have to ask, that is not known to us. So, this is called quota sample. For example, suppose in a country like India, if I want to take a sample of say 20 people who have got blue eyes, well, we know that blue eyes are not so commonly popular, are not so common in India, but if you go to say European countries or there possibly, those type of eyes are more common.

So, now if you want to have 20 Indians, who have got blue eyes it is not so easy to get, but suppose if you decide that you have to conduct the experiment only with them then

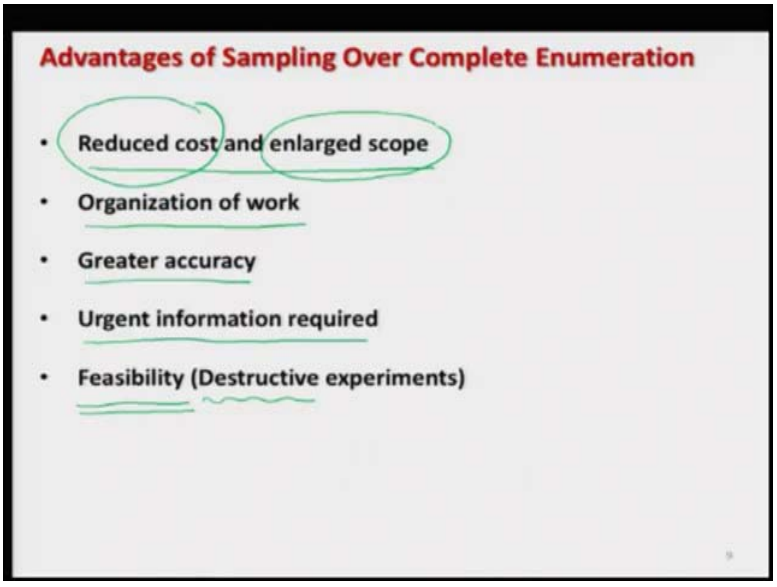
you will start your survey and you will start verifying the colour of the eyes and then you will stop only when the 20th sampling unit is added in your sample.

But, in order to add the 20th unit in the sample, how many sampling unit, sampling units you have to survey that is not known to us. That will be known to us only at the point when the 20th unit is collected, 20th sampling unit is selected in the sample. So, this is the these are the three possible ways to ensure the representativeness in the sample.

Now, next question comes why should we do this sampling? Because in case if you try to increase your sample size means we know in a statistic that the efficiency will increase the quality of statistical inference will get better. So, with this concept, why not to go with the census, why not to go with the complete enumeration?

But, definitely it is not possible, there are certain advantages in the complete enumeration, but on the other hand there are more advantages in conducting the sample survey. So, let me try to briefly address those advantages, the advantages of sampling over the senses or complete enumeration.

(Refer Slide Time: 24:38)



So, first point is this reduce cost and enlarged scope. When you are trying to conduct a sample survey, the number of units on which you have to collect the data they are much smaller than the population and we all know that whenever we want to collect the sample there is a cost involved in the sample survey.

For example, some cost can be fixed; for example, the rent of the office or the salary of the persons etc., but beside those things suppose if you say that means I will give 100 rupees per day to some person to collect the data and suppose the person is able to collect 5 observation per day. So, as a rule of thumb, I can say that category observation is costing us 20 rupees.

So, definitely if you are trying to go for complete enumeration, then you have to collect the data for all the units in the population. So, definitely the cost will become very high. On the other hand, if you go for the sampling, then you have to collect the data on a smaller number of units and so obviously, the cause of the experiment will be less.

And, the sample will have an enlarged scope, what does this mean? When you are going to collect for example, the data on some person, so, you can also collect data on some more variables. And, then whatever data you have obtained in the sample, that has to be used in the defined objective, but there can be some more objective which will also be satisfied by this data set.

For example, if I have collected the data on say blood sugar, blood pressure height and weight of persons. So, I can also check whether the disease of or the problem of say blood pressure, blood sugar, how they are present in the population beside checking whether the person is healthy or not, ok.

The next point is organization of work. When we are trying to conduct a sample survey, we have to collect only smaller number of sampling unit and when we are going for the complete enumeration, then we have to collect the data on higher number of unit or large number of units.

So, obviously, when you have to collect the data on a smaller number of units then organizing such a sample survey is easier. For example, you can say the census in India that is happening after every tenth year, why? Because it requires lot of organization. So, that is another advantage.

Then the third aspect is third advantage is greater accuracy. Well, definitely if a person has to collect only 20 observation or if the person has to collect 20,000 observation, you can very well expect that the person will be more careful when only 20 observations

have to be collected and so, when the data is more accurate, your statistical outcome will also be more accurate.

Fourth option, fourth advantage is urgent information required which means suppose if there is some need and you want to collect the data, then in those cases collecting the data on a smaller number of sampling units is always a better option than going for the complete enumeration.

The complete enumeration is always not possible; it will take more resources more time, more labour, more etc. So, whenever you need the some information urgently, then sample surveys will have a (Refer Time: 28:28) and they will give us a better information. And, another advantage is feasibility. Feasibility means here what?

For example, if you try to see when we are trying to conduct some experiment, some experiments are destructive. Destructive means suppose you are trying to conduct a medical experiment. Now, in the medical experiment you have to give the medicine to some birds or to some animals. And, suppose after the medicine, some infection will be developed in their body or something will happen, so those animals have to be practically killed.

Similarly, if you for example, in another example suppose you want to find out the average life of say here say any equipment like a bulb or tube light or LED lights then what you have to do? You have to just keep them open or keep them working till they get fused.

So, those sampling units will not be usable after the experiment. So, in these cases, well if you if the sampling units are going to be destroyed after the experiment we would like to use them in the means as minimum as possible. Well, just for your information that when whenever we are trying to use some animals in our in the experiment then there is usually an ethics commission.

And, whenever such an sample survey or a or an experiment has to be conducted, it needs the clearance from that ethics commission also ethics committee also, so that they can decide that that you are not using unnecessarily large number of birds or say or say

animals or you are not wasting the resources. So, that is another advantage of sample survey that you can conduct the experiment with smaller number of sampling units, right.

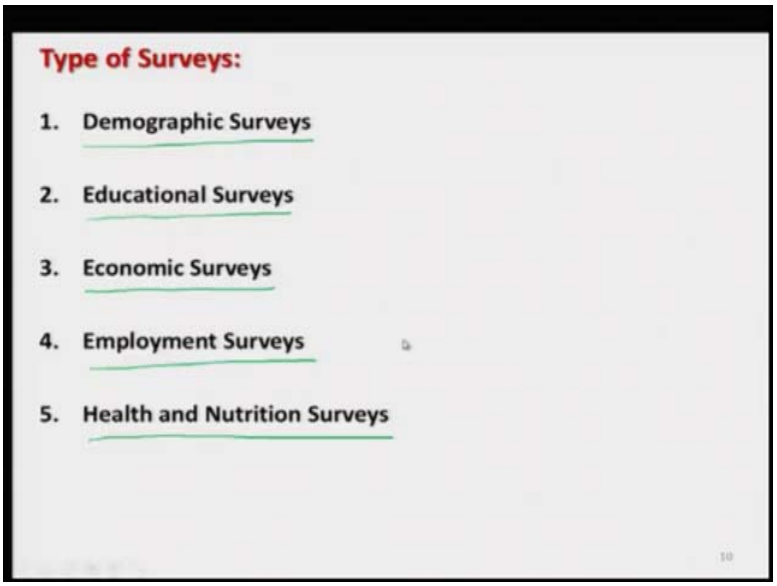
Now, after this I will try to address another issue. We have got different types of surveys and those surveys have some objectives and based on their objective they are conducted from time to time. Some surveys are conducted by the government; for example, in India we have national sample survey organization which conducts different types of demographic surveys, economic surveys etcetera.

There are some private organization they also try to conduct the survey and beside those things there is a new method which also conduct the survey and people do not understand it that they have been surveyed.

For example, nowadays you can see that in shopping stores, there is a culture that they will issue you a card, possibly it is free or they will charge a very small amount and they will tell you if you say spend some amount of rupees then you will get some points and you can encash those points and so on. So, people try to get those cards.

Now, whenever they are making any shopping they have to swipe the card and as soon as they swipe the card, their sample is collected that what type of shopping they have made. So, this is another type of survey and in their card they have the entire information name, date of birth, age, address, telephone number etc. etc. So, this is another way of type of survey which is now happening in happening nowadays, right, ok.

(Refer Slide Time: 32:37)



So, now let me briefly try to address these surveys. So, first survey is demographic survey. So, whenever there is a need related to the demographic information that how many people are living in the city, in the country and so on then we try to conduct the demographic surveys. So, usually in India they are conducted by mostly by the government institutions.

Similarly, we have educational surveys, whenever the objective is that we want to find out the education level of the people staying in the city, district or country then these educational surveys are conducted where the information on education level say different types of qualification etc. etc. from different schools, colleges, university etc. that is collected.

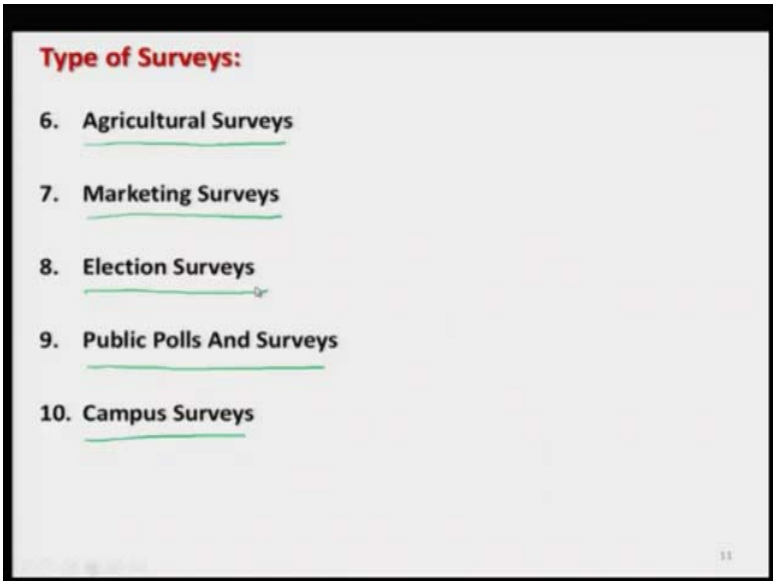
Similarly, there are economic surveys and these surveys try to collect the information on information on the economic indicators. For example, how much the person is consuming, how much the person is spending and there are various economic variable on which the data is collected. So, these are the economic survey and in Government of India also, there are ministries which conduct such a surveys from time to time.

Similarly, there are employment surveys, whenever people want to know the employment status in the city, district or country they try to conduct a survey in which they try to find or ascertain that how many people are employed and what are the

different types of things which are lacking, so that the employment rate can be improved or the people can be helped. So, these are the employment related surveys.

Similarly, we have health and nutrition surveys and under this type of surveys, people try to know that what are the health condition of the people in the city, state or country and they try to collect the information on health and nutritional parameters.

(Refer Slide Time: 34:37)



Similarly, there are agricultural surveys also and these agricultural surveys try to collect the information on the agricultural status that what are the different types of crops, which are being shown, what is the area under cultivation, how much is the agricultural land and what are their requirement, what are their needs. So, that will essentially help the agricultural part of the country.

Similarly, there are some marketing surveys, right. For example, if you go to a shopping mall, sometime you will see somebody will come to you and they will try to give you a small questionnaire and they will ask you different types of questions and they are basically related to the marketing of their product. So, this type of marketing surveys help the people in understanding the needs of the market and the consumer, right. So, these are essentially the marketing related surveys.

Similarly, there are elections surveys. For example, you have seen that whenever there are election, they try to make a forecasting that this time this political party is going to

win or lose. So, in that case also they try to conduct a survey they try to collect the data from the eligible voters in that survey and then they try to analyze the data to inform or to forecast the result of the possible election which is going to happen after some time.

Similarly, there are public polls and surveys, sometime government wants to or some organization or some office wants to wants to employ some policy or some new rules, but before that they want to know how the people are feeling about those things. So, they try to conduct the public polls and based on that they will try to seek the opinion of the people.

And, based on that they will try to take say appropriate surveys and similarly, there are some other type of surveys which are the campus surveys. For example, different types of campuses are there in the country which are educational campuses, medical campuses and so on.

And, in campus there are different types of needs and requirement and these type of survey, try to collect the information on those things, so that they can make the place better that was a very quick brief introduction on various types of surveys, the advantages of sampling over senses and so on. And, my idea was that I wanted to convince you that whatever you are going to do is really fruitful.

And, these are the thing which will which you will require in future. So, you should not be blank that what are those things. You must understand that all these surveys whether they are at city level, state level or country level they are based on some statistical laws.

And, there will be a need that after sometime you may be asked to conduct a particular type of survey, then how to start the survey, what are the different things, which you have to keep in mind – well, these are the thing which one should know as a student of sampling theory. So, you try to think about these concepts. All these things are happening around you. Try to observe them and try to identify whether similar type of incidents are happening around you or not.

So, you think about it and I will see you in the next lecture. Till then, good bye.