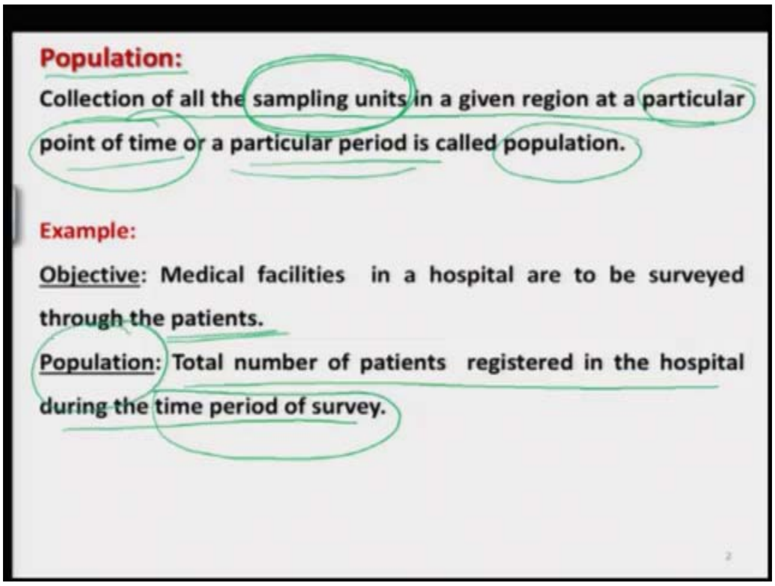
**Essentials of Data Science with R Software -2**
**Sampling Theory and Linear Regression Analysis**
**Prof. Shalabh**
**Department of Mathematics and Statistics**
**Indian Institute of Technology, Kanpur**

# Sampling Theory with R Software
## Basic Fundamentals
## Lecture – 10
## Terminologies and Concepts

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 talk about the Basics Fundamentals related to the Sampling Theory with R Software.

So, in the last lecture I started a discussion on the basic fundamentals related to the sampling theory. So, I will continue with those topics and I will try to take some more topics which I would like to explain you with example and the basic fundamentals behind those concepts, ok.

(Refer Slide Time: 01:09)



**Population:**

Collection of all the sampling units in a given region at a particular point of time or a particular period is called population.

**Example:**

Objective: Medical facilities in a hospital are to be surveyed through the patients.

Population: Total number of patients registered in the hospital during the time period of survey.

So, now let me take the first definition in this lecture. This is about population. What is population? This is essentially the collection of all the sampling units in a given region at a particular point of time or at a particular period and this is called as a population.

Note one thing here that I have used here the word sampling units and you can recall that in the last lecture, I had given you a detailed explanation about the sampling unit. This is the unit on which someone would like to collect the observation.

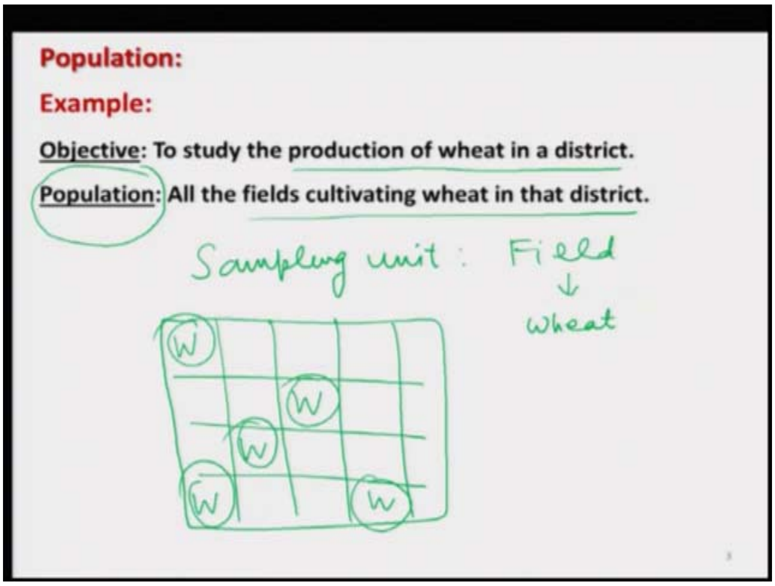
And when I try to take all the units together from where we would like to choose a sample then this collection of all the sampling units will be called as population. And here I am using a sentence, a part of sentences particular point of time or a particular period. The definition of the population basically depends upon the objective of the study and the environment under which we want to conduct the experiment.

For example, suppose we want to conduct a survey to know about the medical facilities available in a hospital. So, this can be done through the patients because patients are the one who can we give us the right report, right observation, right information that what are the medical facilities available in the hospital which are working, right.

So, now the question is this we want to have the information about that hospital. Now what will be my population? I have got several option. The population can consist of the people staying in that city or the people staying in that locality where the hospital is located or don't you think that we should consider the total number of patients which are registered in the hospital during the time period of the survey.

And I am saying here very clearly- time period of the survey because the patient will be getting registered every day in the hospital, but all those patients are not really going to help us only those patients when we are going to the hospital to collect the data or to conduct the survey only those patient have to be considered, they have to be entertained and the collection of all such patients will constitute my population.

(Refer Slide Time: 04:19)



**Population:**

**Example:**

**Objective**: To study the production of wheat in a district.

**Population**: All the fields cultivating wheat in that district.

Sampling unit : Field ↓ wheat

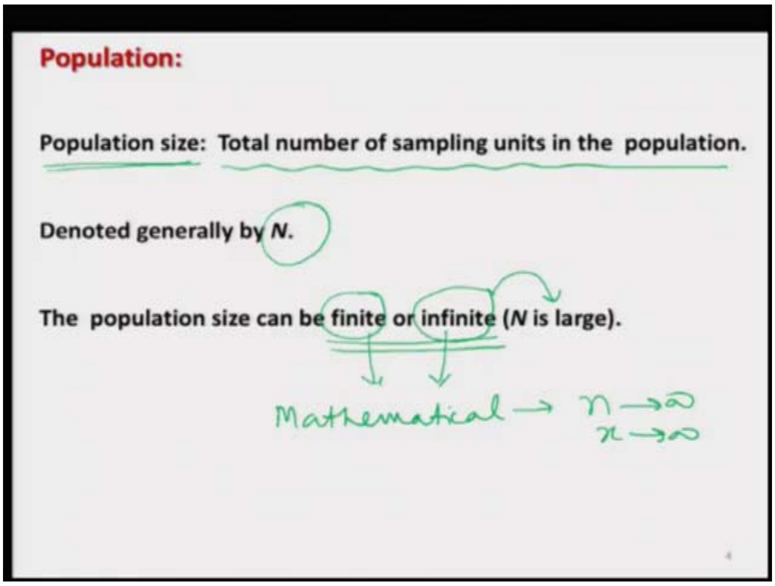
Similarly, in case if I want to study the yield or the production of wheat in a particular district or particular area then we have to first collect what is my sampling unit and then what will be the collection of sampling unit. So obviously, the production of wheat is coming from fields. So, my sampling unit will be in this case will be field.

Now, remember one thing which will in district there can be different fields which are growing wheat, which are growing rice, which are growing pulses or which are growing something else, but our objective is this we are interested in the cultivation of wheat. So, all those field where wheat is being grown, wheat is being cultivated that will constitute the sampling unit.

Now, try to consider all those fields which are cultivating wheat in that district. The collection of all such fields will constitute my population right. For example, if I say suppose hypothetically this is my some district or this is some area where we have got these types of fields and different fields are say cultivating different types of crop.

Suppose the wheat is here in this field, wheat is here in this field, wheat is here in this field, wheat is here in this field, wheat is here in this field. Then all these field this one, this one, this one, this one, and this one they will all together constitute the population and this individual fields they are my sampling unit.

(Refer Slide Time: 06:14)



**Population:**

**Population size: Total number of sampling units in the population.**

**Denoted generally by *N*.**

**The population size can be finite or infinite (*N* is large).**

Mathematical →
$n \to \infty$
$x \to \infty$

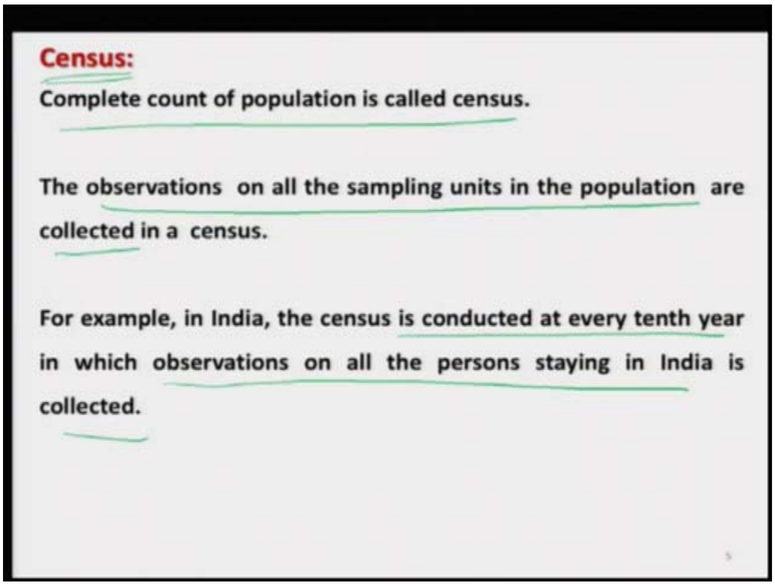
Now, another concept is population size, how do we define the population size. So, this is actually the total number of sampling units in the population. This is called as population size and in general, in sampling theory, we will denoted by capital N, right. Now, this population size can be finite or this can be infinite. The question here is this what do we really mean by finite or infinite.

So, there are two aspect; one aspect is mathematical. Mathematical means, whenever we are trying to do any algebra mathematical analysis to develop a statistical tool or when we try to derive different types of statistical properties of a tool then we try to make different types of mathematical assumptions.

And in mathematics we say for example, if something is for example, if n is going to infinity or x is going to infinity that becomes that n is becoming very very large. But in sampling theory when we say the population is finite there is no issue in understanding it.

But when we say the population is infinite practically we mean that n is very very large, n is extremely large and that is what we for all practical purposes that is what we try to consider that if capital N is very very large we will say the population is infinite, ok. So, that is our understanding in sampling theory.

(Refer Slide Time: 07:51)



**Census:**

Complete count of population is called census.

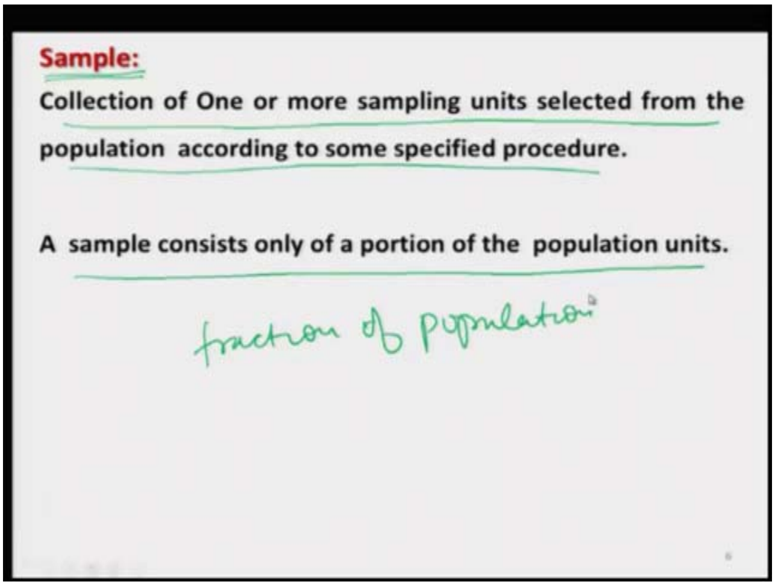
The observations on all the sampling units in the population are collected in a census.

For example, in India, the census is conducted at every tenth year in which observations on all the persons staying in India is collected.

Now, the next topic which I would like to address is census. What is census? This is the complete count of population. And whenever we are trying to collect the observation on the sampling unit then when we are conducting the census then all the units are considered and the observations on all the sampling unit in the populations are collected.

For example, if you say in a country like India the census that is the counting of the total population or the total persons staying in the country that is conducted after every tenth year. And what we try to do? That there are people who will go to all the houses they will try to contact each and everybody whose is staying in this country and they will try to collect observations from them. So, this is called actually census.

(Refer Slide Time: 09:03)



**Sample:**

**Collection of One or more sampling units selected from the population according to some specified procedure.**

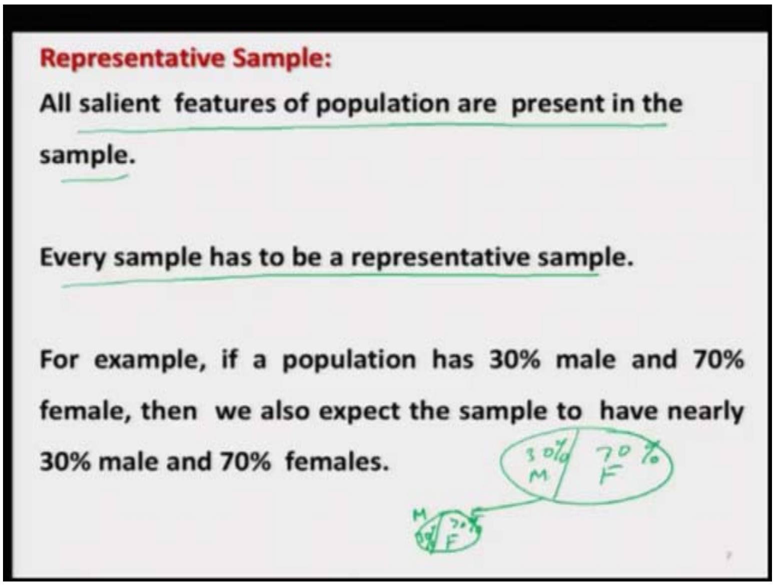
**A sample consists only of a portion of the population units.**

fraction of population

So, in census we try to consider the entire population. Now, the next aspect is this. When we are trying to study the sample surveys or the sampling theory we are interested in finding out a small fraction of the population and that will be called as sample. We will work on the basis of a sample.

So, the collection of one or more sampling unit which are selected from the population according to some pre specified procedure this is called a sample. So, in very very simple words a sample consist only of a portion of the population unit that is a fraction of the population, a small fraction; fraction of population, right. So, this is basically the sample.

(Refer Slide Time: 09:59)



**Representative Sample:**

**All salient features of population are present in the sample.**

**Every sample has to be a representative sample.**

**For example, if a population has 30% male and 70% female, then we also expect the sample to have nearly 30% male and 70% females.**

Now, when we are talking of the sample then what happens, that goes without saying and this is our basic understanding that we assume that the sample is representative that goes without saying that nobody writes that sample is representative, but it is understood, that is default, this is automatic. So, what do you really mean by representative sample?

Now, you have to first understand why are you trying to collect the data. You are trying to collect the data through a sample which is going to be exposed to some tool, some statistical tool or some mathematical tool. So, in case if your data is not having all the characteristic of the population, the same input will be given to the tool and since this information is lacking, so, definitely the tool will not be able to entertain it. Means, any statistical tool does not have a facility to take information from any other source, except than the sample.

So, in case if the sample does not contain the information, that information will not be transmitted to the statistical tool and obviously, when the outcome will come that outcome will be based on the absence or that the information will be based on the information which is provided by the sample only. So, that is why sample needs to be representative.

For example in the earlier lecture, I had taken a sample where a class has male and female students and if I try to take sample or if I try to select some student and suppose all the students are male then by looking at the sample, we will have a feeling that only male students are in the class.

Or similarly in case if my sample has only female students then by looking at the sample we will have an idea that as if there are only female students in the class. So, in this case what will be a representative sample? That male and female students both should be in my sample. So, that will give me an idea that what is in the population.

Why? Because you have to always keep in mind. Population is beyond your reach, we cannot go to the population, we cannot look into the population. What we have, what we can see and what is in our hand is only the sample. The sample means the sampling units. Sampling units means a small number of sampling unit on which we have collected the observation on the random variable.

So, just by looking at the small sample, we have to guess or we have to know we have to retrieve the information for the entire population which is very very huge. So, that is why we always talk of representative sample. So, now, I try to explain you what do you really mean by representative sample.

So, first the basic feature is that that are the one, primary condition is that that all silent features of the populations are population are present in the sample and that sample will be called as representative sample. And this goes without saying by default that we assume that every sample whatever we draw is a representative sample or rather it has to be a representative sample.

For example, if I say, if there is a suppose here population where there are suppose 30 percent here male student and suppose 70 percent are female students and suppose we try to draw here a small sample out of that then we expect that that 30 percent persons in the sample will also be male and 70 percent persons in the sample will also be female.

So, this is what we expect. Well, in practice this 30 and 70 is not exactly 30 and 70, but it is close to 30 and 70. And I can give you one more example, means you all know what is the representative sample. Whenever you go to buy some wheat or rice or some flour what do you do? There is a bag of 100 kg of wheat or say 40 kg of bag of rice, do you open the entire bag and see the quality of the wheat or rice? Really no.

You simply try to take a sample. You just try to take couple of grains some grain 50 grams grain in your hand and then you try to look into those grains. And you say if those grains are good there are no hole there are no insects or the grain is not affected by the insects that mean there are no holes you assume the entire bag is the same, there are no holes.

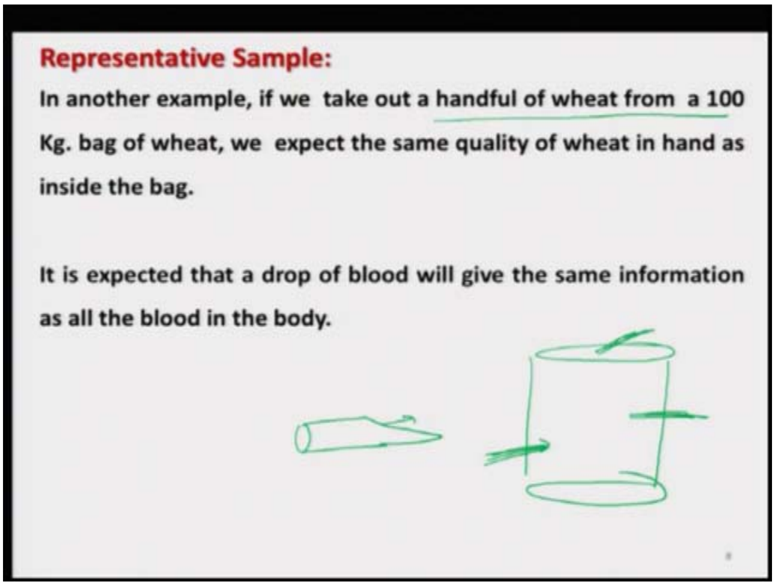
But in case if you take say about 100 grains of wheat in your hand and you say there are 5 to 10 grains which have got the small holes that mean they have been infected by some insect you automatically assume that this 100 kg of bag will also have 5 to 10 kg of the grains which are say affected by the insects.

So, this is what we mean by representative sample, right. And similarly if you to observe and try to think that in your day to day life you have many more examples of representative sample. Have you ever seen that when somebody goes for a blood test. The person working in the pathology will take some say 5 ml or 10 ml of blood or in fact, even a couple of drops of the blood also from the finger from here or from here somewhere it will take. And then after that it will analyse only that small fraction of the blood and it will give you the details about the entire blood in the body. Suppose on the basis of blood samples, somebody gets the value that the haemoglobin level is twelve.

Do you think that the person has ever assumed that if the blood is taken from this finger then the haemoglobin level is 12, but the blood in this hand has a haemoglobin level which is different or if I try to take the blood somewhere from my ear then the haemoglobin level is different. We assume that this blood is a, blood sample is a representative of the entire blood in the body.

So, that is also a representative sample. So, if you try to just think, so, spend some time in thinking from your day to day life you will observe that you are taking samples many times each and every day and without thinking you always assumed as if the sample is representative sample, right.

(Refer Slide Time: 18:30)



**Representative Sample:**

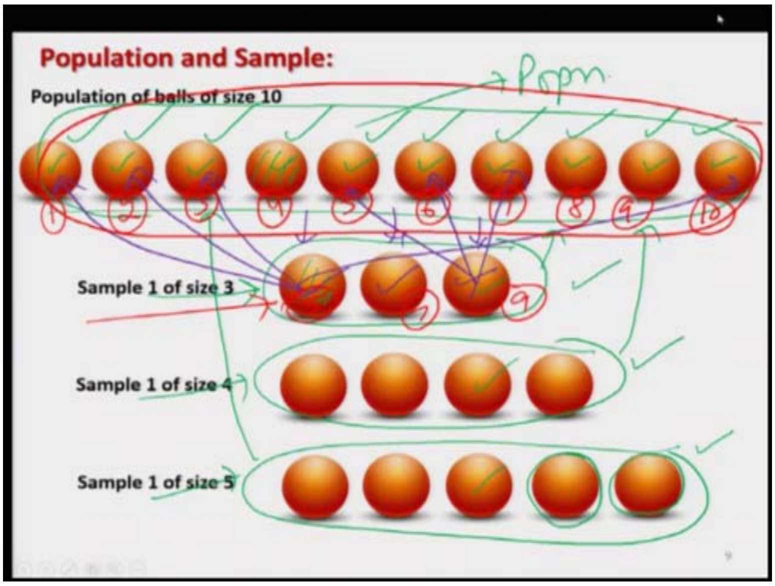
In another example, if we take out a handful of wheat from a 100 Kg. bag of wheat, we expect the same quality of wheat in hand as inside the bag.

It is expected that a drop of blood will give the same information as all the blood in the body.

So, so these are the two example which will convince you that what is a representative sample. For example, this example of handful of wheat from 100 kg of wheat, we expect that the same quality of wheat is present in the bag what was in my hand.

And some time even you have seen if my bag is like this, if means 100 kg of bag like this. sometime you try to take a sample from here sometime you try to take a sample from here, some time you try to take a sample from say here and there is a special type of say equipment something like this which is like this. You try to hold it here and then try to insert this thing inside the bag and it will give you some grains of wheat from the different parts of the bag.

(Refer Slide Time: 19:29)



Population and Sample:

Population of balls of size 10

Sample 1 of size 3

Sample 1 of size 4

Sample 1 of size 5

So, now it is clear that you actually knew what is called a representative sample, I have just made it clear. Now, let me take some more example to convince you that whether that what is the difference between representative sample on a or a non representative sample.

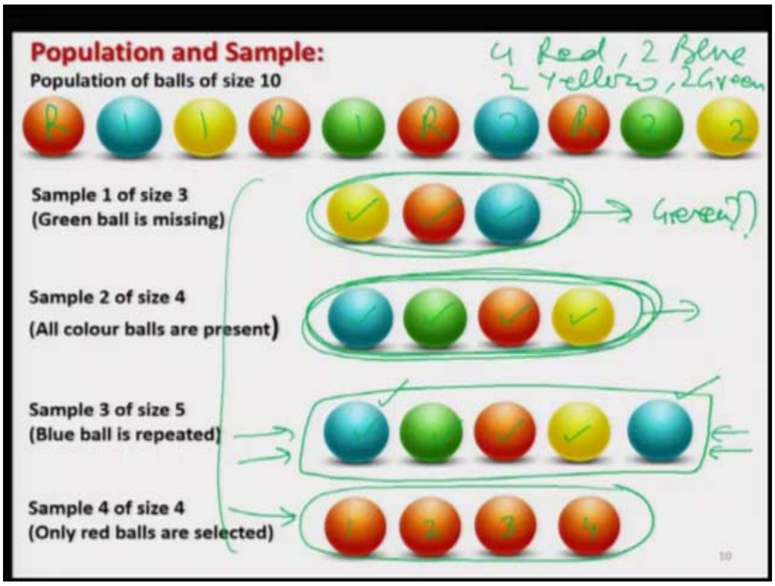
Suppose I take a population of say 10 balls. So, you can see here I have here 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 red balls and suppose I try to, and you can see here they are identical. When that is a very ideal condition that they are like this, they are they are exactly the same. Suppose, I try to take here a sample of size 3 means I can choose any 3 balls out of this 10 balls. So, this 10 balls, this is my here population.

And now you can see here I have taken here a sample of here 3 balls and similarly I can take sample of here 4 balls, if I take here the sample of here 5 balls and so on. And you can see that the balls in this sample, this sample and this sample are they the same as in the population? This is the very ideal condition. You can see that all the balls are the same. There is if this ball is here, this ball is also here, something like this and whatever and howsoever this ball looks like, these balls in the population are also the same.

So, you can see that you can say very clearly that all the statistic of the balls in the population is also present in the sample. Well, I am not talking here of the sample size, right. Sample size also is an important thing, but we will discuss about it later on. At this moment I am simply trying to show by looking at the sample. Can you say that this sample is giving all the information which is contained in the population? The answer is; yes.

By looking at the sample I can say that there are 3 balls or 4 balls or 5 balls and all are in red colour and of the size like this one, size like this one and in the population also we have all the balls which all of red colour only. So, that is the representative sample.

(Refer Slide Time: 21:45)



Population and Sample:

Population of balls of size 10

Sample 1 of size 3 (Green ball is missing)

Sample 2 of size 4 (All colour balls are present)

Sample 3 of size 5 (Blue ball is repeated)

Sample 4 of size 4 (Only red balls are selected)

Now, in case if I try to charge the colours of the ball. So, you can say here I have taken here 1 red ball, 2 red ball, 3 red ball 4 red ball. There are 4 here red balls. There is here 1 blue, 2 blue; 2 blue balls. There is 1 and here 2; 2 yellow balls and 1 and 2 here; 2 green balls.

So, they are all altogether 10 balls, but now you can see the balls are the same, but they are differing with respect to colours, right. So, now, I try to choose 3 balls from this population. Suppose I get here this yellow ball, red ball and blue ball. Do you think that this sample of 3 balls is representative? You can see here that we have here 4 characteristics in the ball. They are the colours red, blue, yellow and green.

But now in this sample this green colour is missing, this is not available. So, I cannot consider this as a truly representative sample because by looking at the sample, I cannot have an idea whether there is a green ball in the population or not or are there any other ball of different colours. This type of information cannot be obtained.

Now, in case if you try to take here a sample number 2. Here I have taken one ball of each colour. So, now, the question is this. Can you really take it as a representative sample? Means, if you go by the colours, then yes, this is a representative sample.

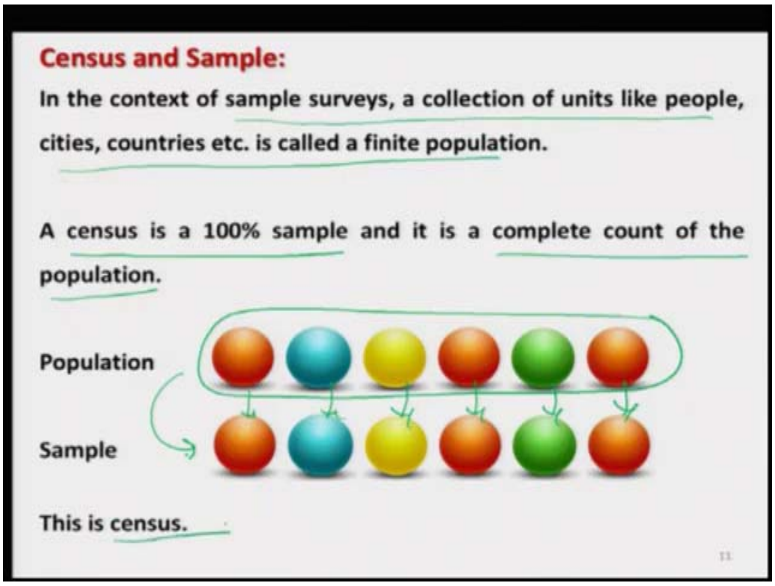
It is indicating that there are blue balls in the population there are green balls in the population there are red balls in the population and there are yellow balls in the population. But, definitely you can also notice here one thing that the number of red balls is 4, where as the number of balls of other colours they are 2.

So, this number is not being here representative because that is by looking at this sample you are getting a feeling that if as if all the balls of different colour are in the same number in the population. Now, I take another sample over here, sample number 3. So, you can see here you have here all types of ball; blue, green, red, yellow, but there are 2 blue balls here.

So, if you try to look here in this sample this sample is trying to give us an information that as if the number of blue balls in the populations are more than the green, red or yellow balls. Is that really true? It is not actually correct because actually there are the, there are 4 balls which are red colour balls and all other ball like is blue, yellow or green they are only two in number. So, yeah up to certain extent this is representative, but it is not 100 percent representative.

Now, if you try look at the sample number 4, right, this is taking all the red balls from the population. There are 4 red balls in the population and you are taking here all the 4 red balls in your sample. So, if you try to look here in this sample, this sample is it is indicating as if the population has got only one colour of ball which is red. There are no balls of any other colour like as blue, yellow or green. So, in this case this representative sample cannot really be consider as a representative, right.

(Refer Slide Time: 26:03)



Census and Sample:

In the context of sample surveys, a collection of units like people, cities, countries etc. is called a finite population.

A census is a 100% sample and it is a complete count of the population.

Population
Sample
11

This is census.

So, one thing is clear by this example that when you are talking of the representative sample, the 100 percent representative is very difficult to say achieve, but there will always be some percentage or proportion of the representativeness in the sample.

So, obliviously if you are working in real life there are so many constraints and once you do not know the entire population means you really do not know that how many types of ball are really in the entire world. So, when we talk of the representativeness usually, it is not taken means ideally it should be 100 percent, but in practice we try to make it as large as possible, as maximum as possible. This is what you have to always keep in mind, right.

For example, I have taken here different types of say sample and every sample is and most of the samples are representative, but you have to see that the degree of representativeness is varying in every sample and all are the sample they are drawn from the population.

So, now here comes the role of sampling theory. Means, you have to device a methodology by which you can ensure that whatever sample you are going get that is going to be a representative sample up to a great extent as much as possible, as maximum as possible and this is the job of the sampling theory. And different types of sampling techniques you might have heard about different names like simple random sampling, stratified sampling, cluster sampling, systematic sampling, two stage sampling, two phase sampling, population proportion to size sampling, etc.

So, these are different types of sampling schemes which work under different types of condition and which helps you in getting a representative sample. But, then the question come why should I study all these types of sampling techniques, why one is not sufficient. Because the populations is changing, the characteristic of the population is changing from one population to another population.

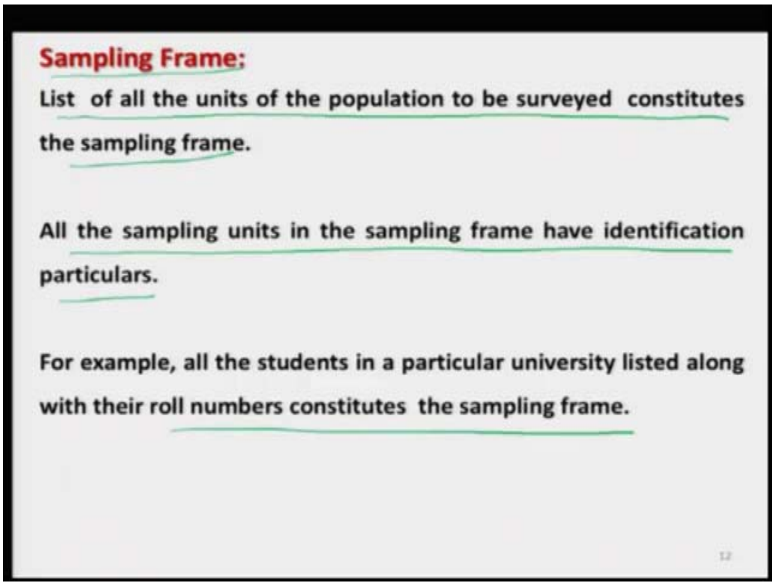
So, depending on the behaviour, nature, environment of the population, you have to choose an appropriate technique which can ensure that it will give you sample with high degree of representativeness ok. So, now I try to compare the census and sample with this example of ball.

So, now, as we have discussed that when we are trying to conduct the sample surveys usually collection of units like people, cities, countries, etc. this is called as a finite population, right. For example, if you say the population of the country like India is also finite because we know at any given point of time that what is the exact number of person staying in the country, ok.

So, the relationship between census and sample is very simple. A census is simply a 100 percent sample and this is a complete count of the population. For example, if I try to take here 6 balls in my population and if I try to draw here a sample in which all these balls are here. You can see here, yeah, at the most what will happen there? Their ordering may change, but that does not make any difference at this stage, right.

So, in this case the sample and population they are exactly the same. We have drawn a sample of the size of the population and no unit has been replaced back. So, this is actually here census. So, this is basically the relationship between census and sample.

(Refer Slide Time: 30:11)



**Sampling Frame:**

**List of all the units of the population to be surveyed constitutes the sampling frame.**

**All the sampling units in the sampling frame have identification particulars.**

**For example, all the students in a particular university listed along with their roll numbers constitutes the sampling frame.**

12

Now, the next concept is sampling frame. You see when we are trying to conduct a sample survey what we have to do? We need to draw a sample. Now, you can see here in this example if you see into my slide in this example here, means I have this unit, this unit, this unit, this unit, this unit, this unit, this unit, this unit, and this unit in my population.

Now, I am taking here a sample here. Now, can you tell me by looking at these 3 balls that which ball is this one. this one, this one, this one or this one and same is true for whether this ball is this ball, this ball or this ball. I do not know and I cannot even know. So, one simple option is that means, I can identify all the ball.

For example I can give here a number. For example, I am now writing in red colour, this is my ball number 1, this is my ball number 2, this is my ball number 3, this is ball number 4, this is 5, this is 6, this is 7, this is 8, this is 9 and this is here 10. And suppose when I try to draw here 3 ball, suppose this is ball number 5, this is ball number 7 and this is ball number 9.

Now, I can determine that from the population, these 3 balls in my sample, what are they representing. So, this collection of balls when we are trying to give an identification tag, this is called as sampling frame. It is as simple as that, right. So, I can define it formally that, list of all the units in the population to be surveyed constitute the sampling frame.

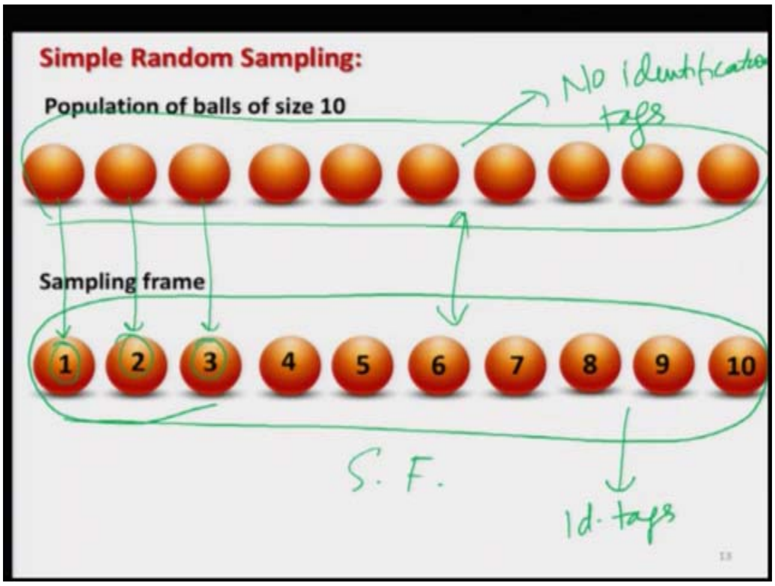
And there is one more condition that all the sampling units in the sampling frame have identification particulars.

For example, this identification particulars will vary from one population to another population. For example, if you are trying to conduct a survey in a university or a college, then it is possible that there are students who have got the same name, right.

So, their name cannot be cannot be the be the identification unit that can be assigned to the students, but if I go with their roll numbers then we know that the roll numbers are unique,  there cannot be two students who have got the same roll number.

So, when I try to take the collection of students along with their roll numbers, this will constitute a sampling frame. Similarly, if you go to the hospital and if you want to conduct a survey on the on the patients then their registration number will act as an identification tag and if you try to prepare a list of all the patients available in the hospital and try to write down the corresponding registration number then it will create a sampling frame.

(Refer Slide Time: 33:40)

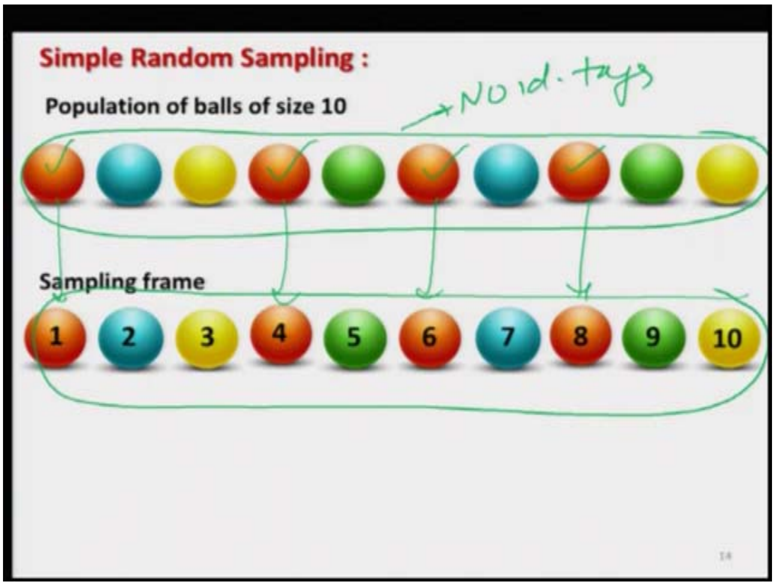


For example, in the same example which I just shown you here you can see here, now this is the population of 10 balls which are all in colour red, but you cannot identify all the balls. All are similar or say all the balls are the same.

But now what I try to do? I try to assign an assign a number. For example, this ball has been given the number 1, this ball has been given the number 2, this ball has been given the number 3 and so on. I have given the number 1 to 1 to up to here 10.

So, now this is my sampling frame, this is my sampling frame and, but you can see that both this and both these two figures are the same, but they are , but in this case there are no identification tags. So, this is only a population and it is not a sampling frame, whereas, here we have identification tags attached to all the sampling units in the population and this constitute the sampling frame.

(Refer Slide Time: 34:51)



Simple Random Sampling :

Population of balls of size 10

Sampling frame

And similarly in the in the next example where I took different colours of ball. You can see here we had 4 red colours ball; 1, 2, 3 and here 4, but now these four balls are now different. They have been assigned number 1, number 4, number 6 and number here 8.

So, similarly all the balls have been assigned a number from 1 to 10 and now these all the balls in this group they are with some identification tag. So, this is my sampling frame. Whereas, this one this is all the collection of the balls, but there are no identification tags. So, this is only a population. So, this is the basic difference between the population and sampling frame, right.

Now, I stop here in this lecture. So, I have tried my best to give you some example and I have tried my best to explain you the basic concepts of the sampling theory through various example. I am sure that you will be comfortable and it is not very difficult to understand them.

But still I would say now, you please try to think. Think about some surveys which are happening in this country. Try to think how you have defined the sampling units, how you have defined the sampling frame and once these concepts are there inside your mind it is not difficult to understand the further topics because later on, we will be using these system and all these. For example, I will simply be saying that we have say some population and the corresponding to which we have a sampling frame.

So, you please try to settle down these terminologies in your mind, so that you become comfortable in using in the next lecture. So, you practice it and I will see you in the next lecture. Till then good bye.