Handling Large-Scale Unit Level Data Using STATA
Professor Pratap C. Mohanty
Department of Humanities and Social Sciences
Indian Institute of Technology, Roorkee
Lecture 06
Review of Sample Techniques-I

Welcome learners, once again, for the MOOC module, on handling large-scale unit-level data with STATA. This is one of the unique platforms, where we are trying to cater to the need of the researchers, especially the Ph.D. students, or the budding researchers in different institutions, for their future endeavour in achieving or targeting for a particularly large scale unit-level research. So, myself Pratap Mohanty, have been teaching this particular module on sampling, and its design, data handling, for over several years, in different platforms.

So, I will try my best to nurture you these techniques very systematically. Wherever you have difficulties, you may raise these questions in our question forum of NPTEL there is a column called ask a question. You raise your question, will certainly cater to your need, as early as possible with the help of our TA, the TA is Shruti Shuvam and the Shruti Shuvam is a Ph.D. scholar in our department under me. and she has been preparing the details over time and these PPTs are largely also held by her. So, I will specify the particular need is usually desired at these levels.

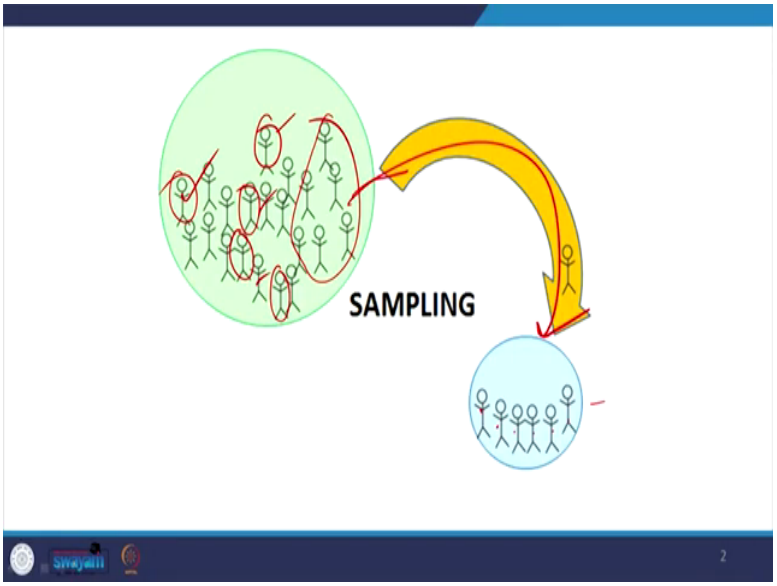
So, we are starting with the week number two, in week number two, we have a separate discussion as compared to the first week. In the first week, we discussed the familiarisation of data, understanding data, through different data sets available in India. But here we are taking you to the journey of understanding samplings because, without samplings, we may not understand the questionnaire of those data set correctly.

How those questionnaires are dealt? What do you mean by sampling? Even if a large data set is presented before you, having more than 6 lakhs observation that does not mean that is important to you. Entire data set surveyed by the government may not be important, even a small sample out of it is very important.

You may also take repeated sample, number of samples, iterate a number of times, have an experiment of some cluster of samples out of it, and test your requirement or a variable whether it is generalizable or not. At what level if you increase the number of sample units

out of the bigger sample, whether still your result hold or not is very important. So, sampling technique is very very important, I will also add what are the details of sampling. So, let me take you to the further details of it. The module, the first lecture of week number two is defined as review of sampling or sample techniques. So, review of sampling or sample technique is important. The picture itself tells story.

(Refer slide Time: 03:59)



I did not specify other details, then challenging your understanding on it because, it seems as if there is a Genesis and a species on it, the species may be discussed about the Genesis in a very different manner. There are a number of narratives defined from this particular picture.

You can tell lots of stories, like let me help you here. Here is the population, it seems that is the entire population or the number of persons within this group, the bigger group, bigger circle is more than that of the small one. We selected no filter it and derived those small numbers or I name it as species, but the right term is not species, it is simply called sampling.

It may be the case that these are 1,2,3,4,5,6. These six people are selected it may be the case that, this one is selected, this one is selected, this one is selected, this one is selected, we do not know these six are selected. There are the different possibilities of selection. So, the approach of selection matters, if I simply select these because this is very closer to my hand, or closer to my locality. So, these six are important to me.

And if I have certain experience, in selecting the persons from a group, instead of following a complicated sampling procedure, I may select some specific person and I know that they can

represent the entire population. That is also another approach of selecting the sample. The simplest format of selection is the simple random sampling, but random sampling has lots of limitations. They said, this may be selected, this may be selected, anyone, if anyone single individual has equal chance of selection in the sample, is called random sample.

But it may be the case that the population itself is not normal in the sense that they are not homogeneous. If they are heterogeneous enough, randomness may not be giving you the right result. If there are two heterogeneous, but you have selected the five or six, but a sample discussing about a specific people with some similar features, they are not covering entire one. But if the population is skewed may be positively or negatively, but you applied a random one, will result in a skewed distribution in the sampled database also.

So, first assumption of selection is that the characteristics of population. Whether it is normally distributed, or randomly distributed or not, or they have equal chances homogeneous or heterogeneous. Accordingly, different formats or different sampling design is developed. You might be also asking yourself why sampling? Why should I bother for sampling?

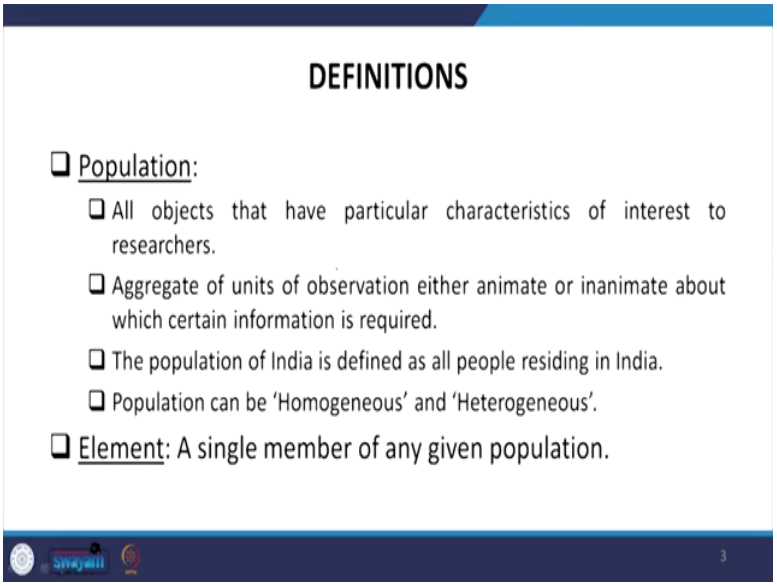
I have a slide, completely dedicated to it to answer, why sampling is important. Let me mention if you hear that it is very cost-effective. Imagine our population in India is exceeding 130 crores. 130 crores and counting the 130 crores is nearly impossible. And it is possible but if you are including a very less number of variables, if our target is to analyse the very specific information on it, then counting is fine we do follow census in every 10 years.

Even the government of India is presently thinking of NPR (national population register) of counting each population. But there are a number of questions raised or how could you be able to do it? It is consuming time, space, money, or expenditure. So, it is very difficult to deal with. So, the most important aspect of sampling, or survey through sampling is to minimize costs.

The second one, sampling also minimizes error. When you start counting all the population there are high chances of under-reporting, if you directly get that specific information and report may be under-reporting, or maybe the case that, many information are hidden at the point of counting, or it may be the case that while since its 130 crores times minimum 20 or 30 variety of information you asked.

130 crore times 20 or 30 variables. The information in each block is too huge. So, there are high chances of error in the entry of the data. So, in the non-counting procedure also while you enter and feed the data in the computer, there are chances of major errors. Where in the sample, chances of those types of errors are very less, entering data because a number of variables are limited, but features are different. So, we will discuss in detail.

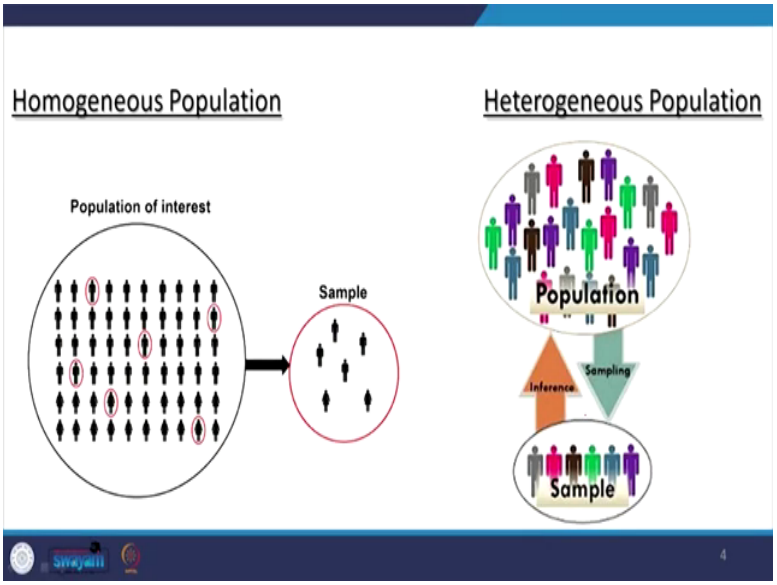
(Refer slide Time: 09:41)



So, whatever I said, I am just going in a structured manner with detailed point. So, let me read between the lines of some of the important points. I do not want to skip certain contexts. So, what do mean by population, I am referring that all the objects that have particular characteristic of interest to the researcher, or aggregate of units of observation either animate or inanimate, about which certain information is required.

I have already discussed why I did not give further explanation to it, the population of India is defined as all population residing in India, population can be homogenous or may be heterogeneous. Accordingly, the sampling design is framed. What do you mean by elements in our procedure? A single member of any given population is called an element. Within the population if a single member is referred, we term it as an element. As we have already said, population may be homogeneous, may be heterogeneous.

(Refer slide Time: 10:46)

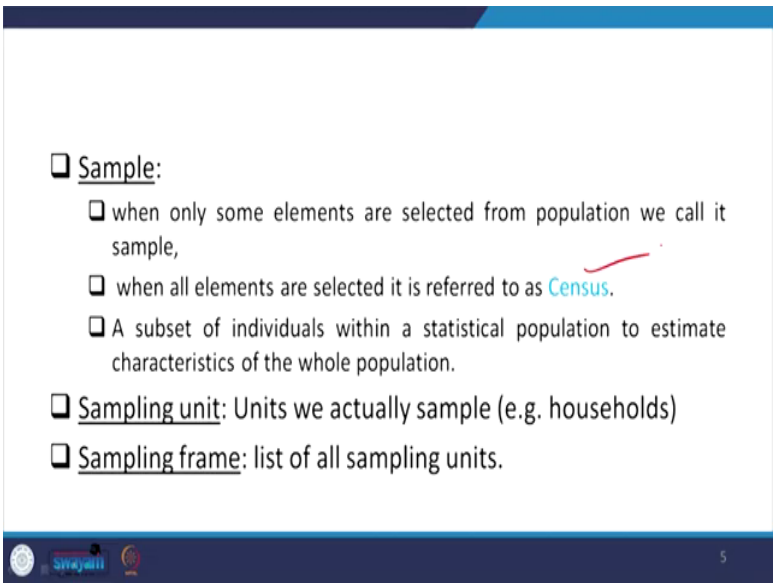


If it is homogeneous, usually this is not the character of population. So far as population are concerned. If it is so, then we can derive a sample. If it is heterogeneous, we will also derive the sample, but usually when the population is having similar characteristics, are more or less homogeneous, the sample size you need to derive should be very less. Whereas, if it is heterogeneous, the sample size is generally expected to be higher, then only you can cater to the larger population. So, different colours are given deliberately in order to describe you, how we can represent our sampled data while interpreting to a larger population.

So, it is interesting to note, it may be the case that our population is heterogeneous like there are different colours, we combine those colours or you cluster them in a different format, if you combine the same colour differently, like red colour in one place, black colour in another place, pink colour in another place, after combining or making them cluster we can than randomly select the clusters. That will simplify our sampling procedure also.

So, not necessarily homogeneous sampled data or heterogeneous sampled data should be different. But ideally, when your characteristics in the population are homogeneous, you need not go for more sample numbers. Because a small number of sample might generalize the context, but when it is heterogeneous, you have to make different strata then each strata you have to collect certain information on it. There are some standard rules by which you can decide the sample size. That will represent the whole population. Our next lecture will be on that understanding the sample size, we can able to generalize the entire population, so that the error will be minimum.

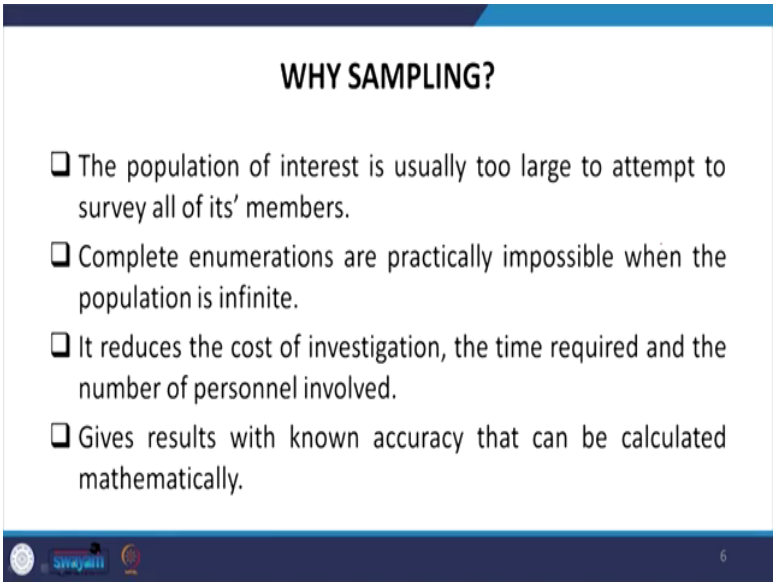
(Refer slide Time: 12:55)



Looking at sample, what do I mean by sample? I already said those are only some elements selected from the population. So, we call it sample. When all elements are selected, we are already mentioned, called as Census. A subset of individuals within a statistical population to estimate characteristics of the whole population is called sample. What do you mean by sampling units? Already mentioned in our NSS schedule that sampling unit are the units which we sample. That we collect the exact household, or the exact individual. Generally, our sample unit is a household considered in different database.

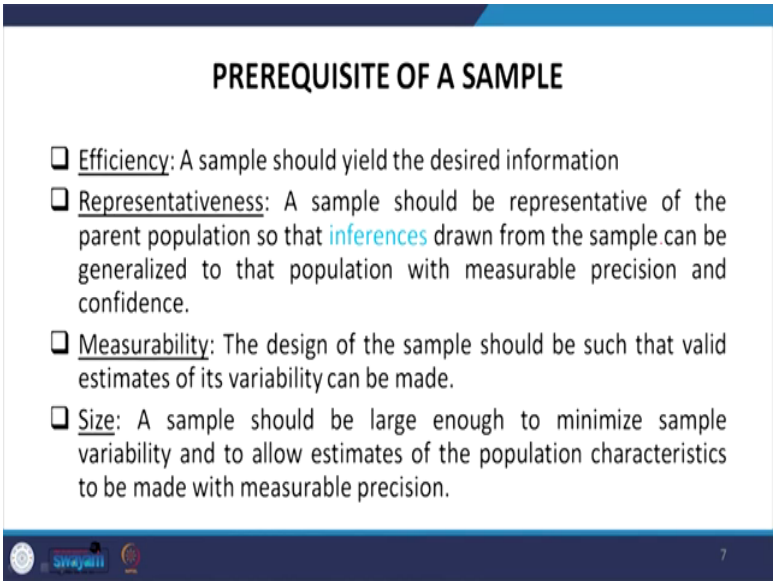
Whereas sampling frame is the entire sampling units which are collected it is a list of all the sampling units called sampling frame. What do you mean by sampling design? The way by which you have collected your sample, the structure by which you have collected the sample units or the entire sampling frame is based on a design or structure called sample design. Then, why sampling? So, is interesting because it minimizes cost.

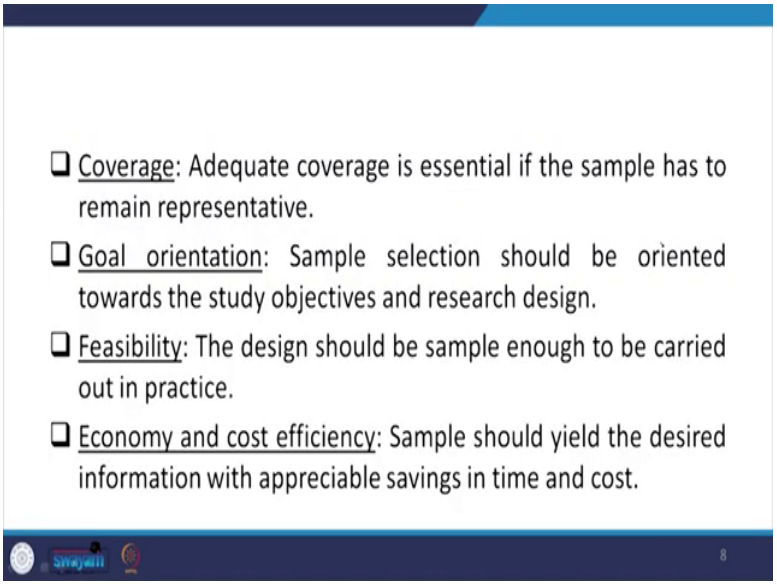
(Refer slide Time: 14:22)



So, it is the population of interest which is usually too large to attempt to survey of all its members and complete enumeration are not practically impossible when the population is infinite. It reduces the cost of investigation, I have already said, it gives the results with accuracy and which can be presented mathematically. There are no different indicators of validating of sample.

(Refer slide Time: 14:52)





So, we have a dedicated lecture for it. Also understanding the accuracy, precision accuracy validity for sample. What do you mean by efficiency in this segment in this lecture and the sampling frame or what do you mean by or what are the prerequisite for a sample? It is efficiency, repetitiveness, measurability, size, coverage, goal orientation, feasibility, economic or cost-efficiency. So, efficiency we do mean that the sample should yield the desired information, the target information you must have said from the beginning or the hypothesis you are saying or the research question you have said in your research must have dealt.

Representativeness, what I mean by that? It must be covering the parent population, it could derive inferences, the sample should be in a position to derive lots of inferences and those will be comparable to the population features. To the entire population, which we discussed with measurable precision and confidence.

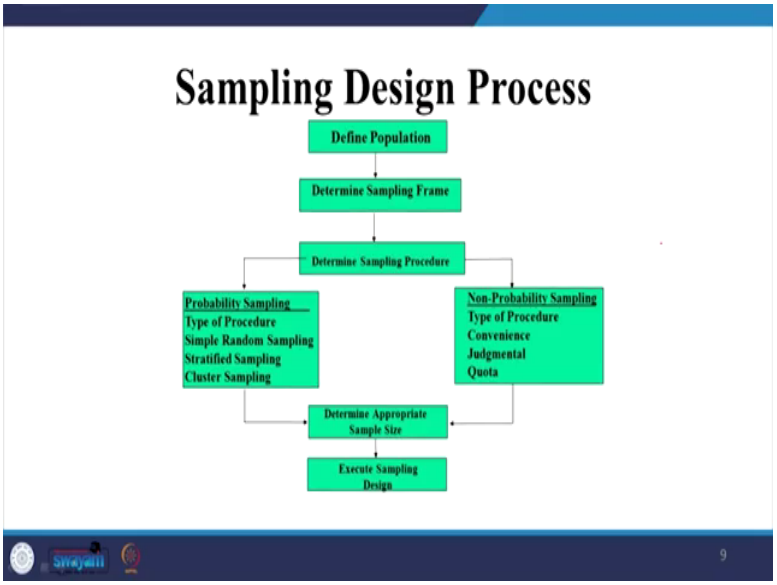
So, far as measurability is concerned, the design of samples should be such that the well valid estimates of its variability can be made. Sample size matters, it should be large enough to minimize sample variability, if the variability is less it is good to represent the population and to allow estimates of the population characteristics with measurable precision.

Coverage, so far as coverage is one of the features of sample is concerned. Adequate coverage is essential if the sample has to represent the entire population or if it is representative, it must have a defined goal sample should be oriented towards the study objectives or research design as I mentioned. Feasibility, the study should be sample enough to be carried out in practice, I mean, sometimes we frame the research question and we

propose a sample but technically it is not feasible. So, for example, if I just simply mentioned that, what is the probability of a person to be live, if the person is moving to the moon.

If this type of objective or research questions are there because we do not have a population in estimating how many are moving to the moon and what are the other variability in this regard. So, it is practically impossible to study. So, it is very important then. So, the cost-effectiveness is very important as I mentioned, because it saves time as well as cost. So, let us have a sample design.

(Refer slide Time: 18:00)



As I mentioned already, so we start with defining a population, what is our total aggregate information in this regard, and we need to define the sampling frame, I already discussed that, then we need to follow the right sampling procedure. In order to derive or execute the sampling design, we must have followed a right sampling procedure. The sampling procedure follows two approaches probabilistic structure or non-probabilistic structure.

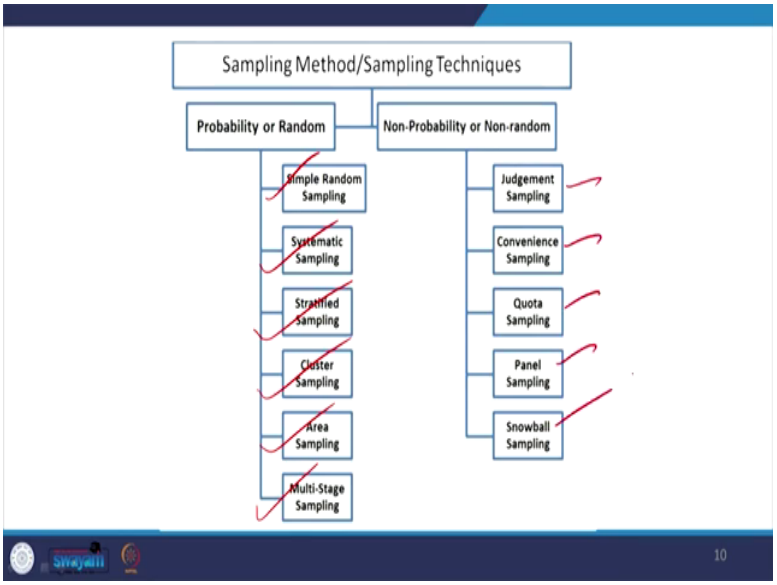
Probabilistic structure, we do mean where the probability of selection of all, the chances of being represented in the sampling frame is equal. When each person of the population, we are targeting, or the entire range or the aggregate information out of which the particular information is representing the entire sample frame should have equal chances.

So, the type of procedure followed are usually of simple random sampling, stratified sampling, and cluster sampling, these three are usually followed. Whereas, in the case of

non-probability sampling, we generally follow convenience sampling, judgmental sampling, and quota sampling.

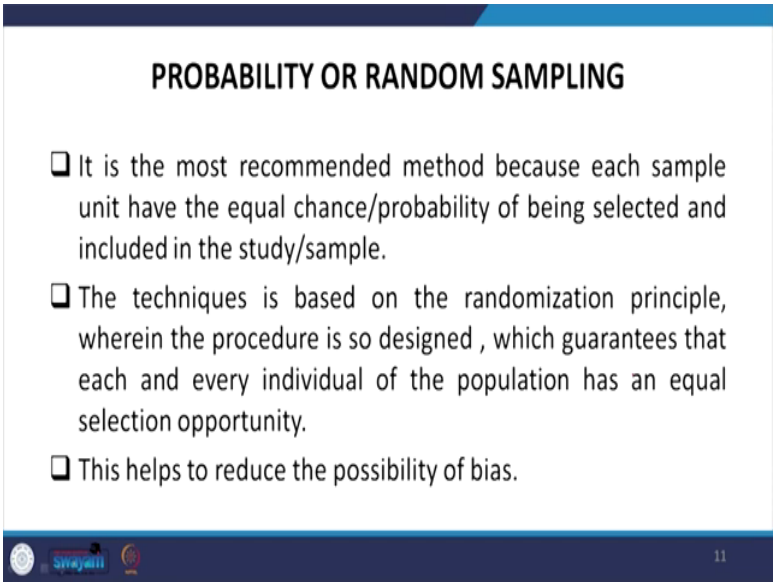
So accordingly, once you follow the procedure, the type of sample procedure we consider, then we can understand what exact size is important. Random sampling generally covers more sampling units, whereas if you follow the right procedure, multistage sampling procedure, it minimizes the units. So then, let us go for further details of sampling techniques and the procedure. So, it is based on the randomness as I already mentioned.

(Refer slide Time: 19:57)



Starting with simple random sampling, systematic sampling, stratified sampling, cluster sampling, area sampling, multi-stage sampling, these are part of probability sampling or random sampling. Whereas, nonprobability sampling covers judgment sampling, convenient quota, panel, snowball sampling. So, I will be explaining each of these steadily.

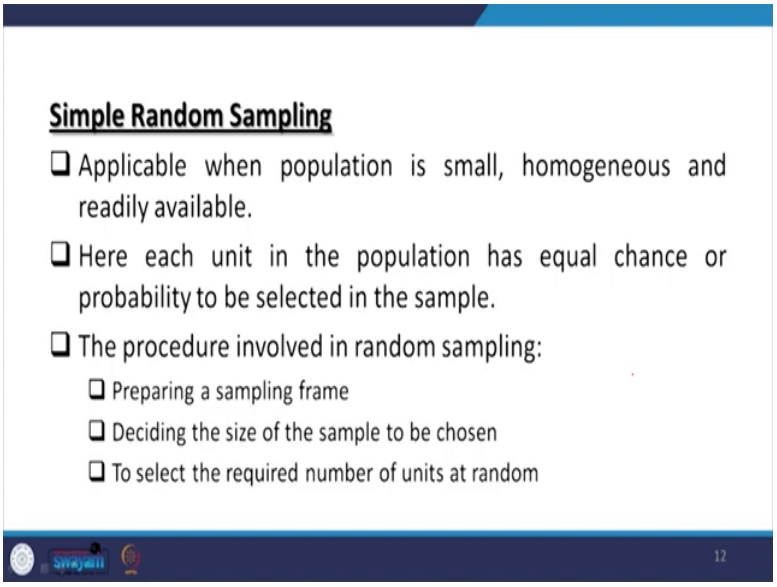
(Refer slide Time: 20:18)



Let us start with the random sampling. Random sampling, as I already mentioned the chances of selection is equal for each of the units within the population to be selected in the sample. So, each has equal chances. This is recommended, when it is homogeneous, each person has equal probability, the technique is based on randomization principle wherein, we generally follow the random table or the random data set available to start from a point of selection, I will discuss in our next slide.

This guarantees that each and every individual of the population has equal selection opportunity, this helps also in reducing possibility of bias because we are not differentiating persons from persons within the publishing space.

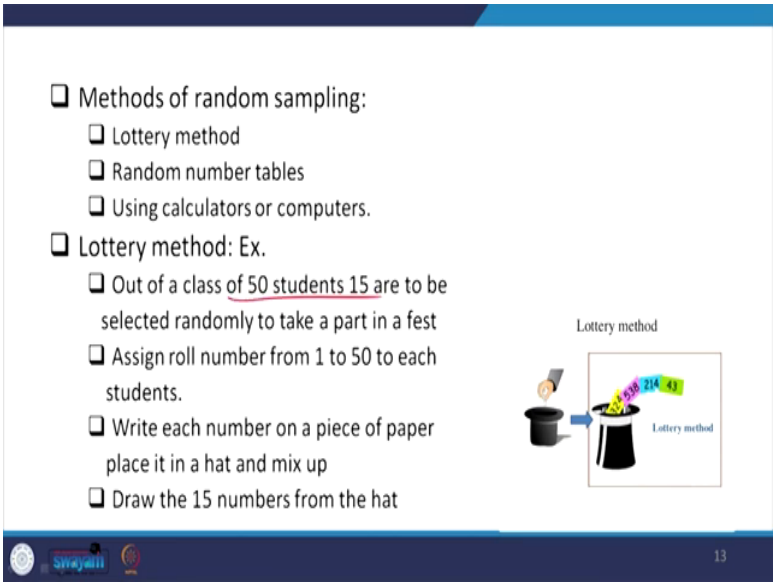
(Refer slide Time: 21:14)



So, looking at the simple random sampling, once again, this is applicable when population is very small. But if population is very large, it is generally heterogeneous, not homogeneous. So, first of all, it should be small, and it should be homogeneous. And it should be readily available, when you do not know what your population is I have just given you an example where our population is not defined.

So, sampling is very difficult. So, here each unit in the population has equal chance or probability to be selected in the sample, the procedure involved in the random sampling procedure is we do follow preparing a sampling frame, deciding the exact size to be taken and select the required number of units at random is also important, we have some techniques that will guide you to understand what is the exact sample size to be followed.

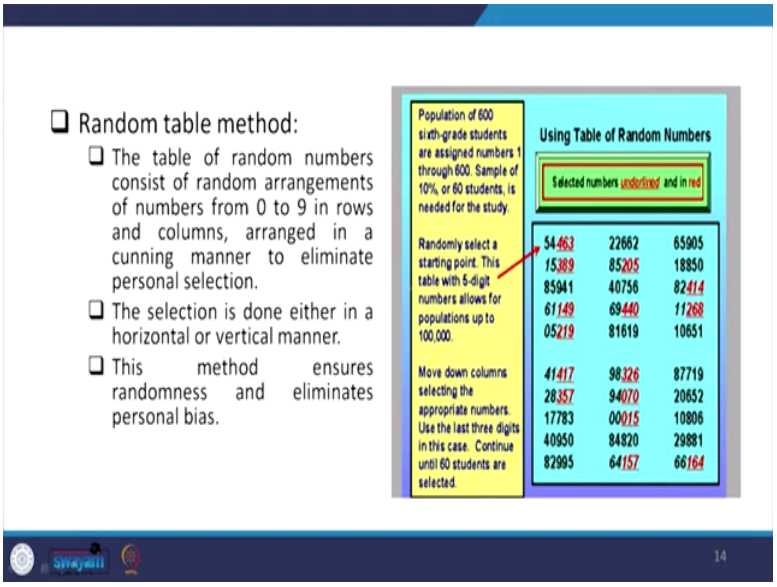
(Refer slide Time: 22:22)



Methods of random sampling, which are the general methods follow, you might have heard about bidding process, lottery. There are a large number of bidders in any public programming. Government usually opens the bidders or tenders are there, so lottery process, random number tables, using calculator or computers to get a particular number to be selected. So, I am not going into the depth of it, because it is too simple to understand. So, we may proceed.

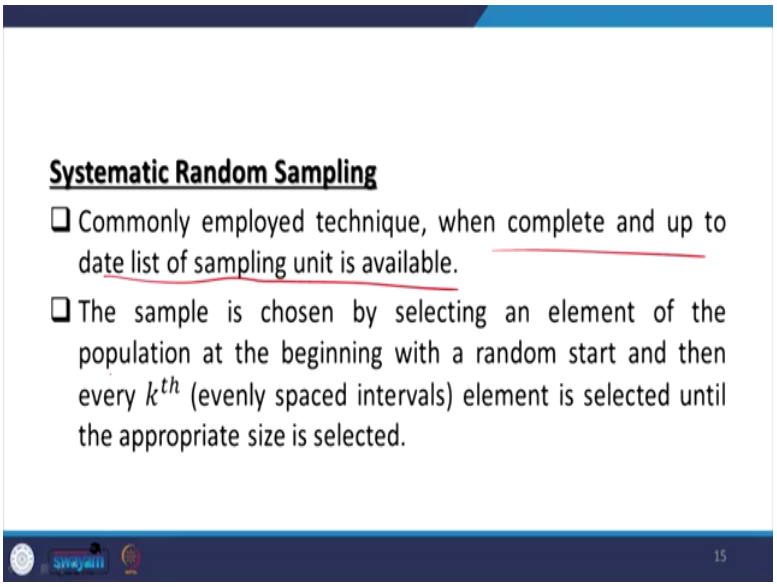
Suppose there are 50 students 15 are to be selected randomly, to take part in a fest in a class of 50 students only 15 to be considered. Who are going to be the part of this fest as per the assigned roll number from 1 to 15 first to each of the students then write each number on a piece of paper. and then mix up all each of those numbers and you can draw those lotteries out of those papers. So, 15 you have to draw those are selected called simple random sampling. So, this is one of the examples of a random number table. A random number table by the government of India is also available from any numbers from a space.

(Refer slide Time: 23:45)



If you define the population, there are population charts are given here, the numbers can be picked up. Even in the computer, it is always suggested to go by the computer and computer will give you, if you define your population space from this number to that number, the computer will give you that random number. Even if you do not know some persons in your contact, you simply give them the units and ask them to pick up some numbers, that is also sometimes defined as random numbers.

(Refer slide Time: 24:17)

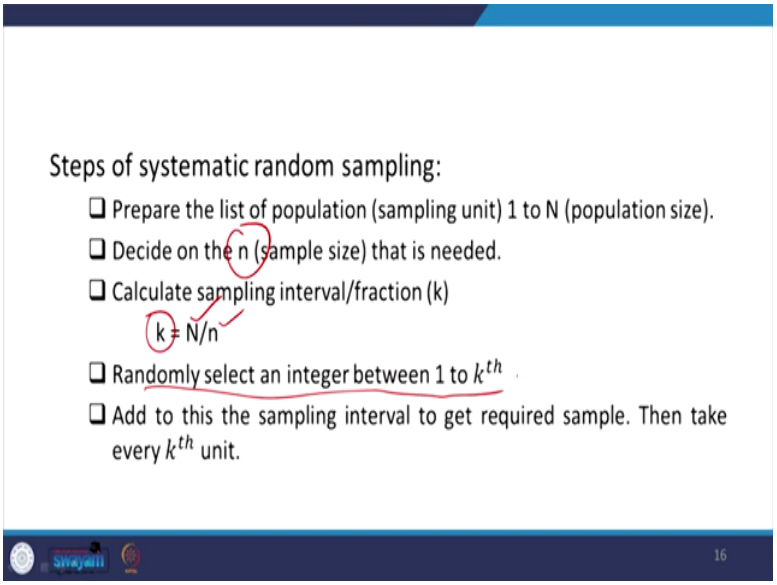


What about simple systematic random sampling? Systematic random sampling, the randomness is a task. So, it is a probabilistic sampling design. In the systematic random

sampling, let me understand very carefully that our population is divided into different clusters. Each clusters have equal chances of selection within the cluster each number have equal chances of selection.

I will tell you systematically from start, we give an example, commonly employed techniques, when complete and up to the date list of sampling is available. The sample is chosen by selecting an element of the population at the beginning with a random start and then every Kth element is selected until the appropriate size is selected.

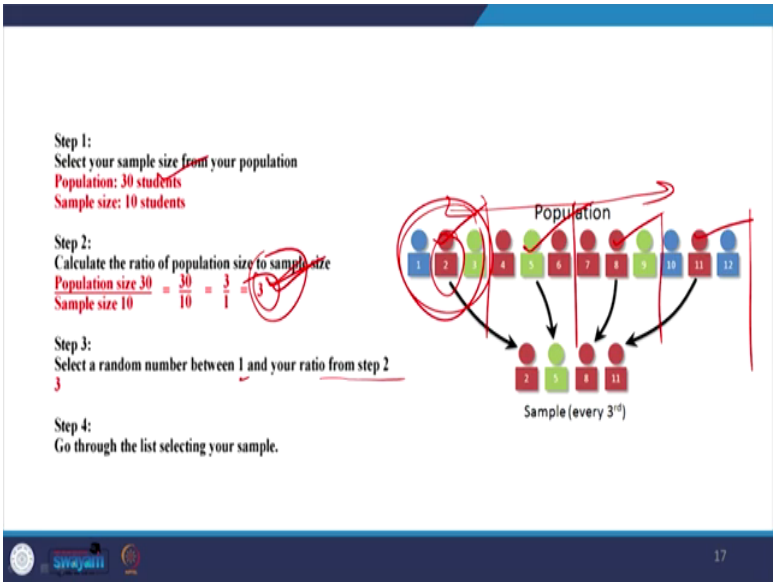
(Refer slide Time: 25:08)



Let me give you the example here. So, steps of systematic sampling. First step is, prepare the list of the population from 1 to N. So, I have already clarified systematic will follow a systematic procedure. So, first you number, the population from 1 to N and decide the sample size how much you require and how much is required sample size to be done, we will discuss the sample size procedure in the next class.

But, suppose the sample size then you pick up is n (small N) then calculate the interval. Interval that means total population if it is 100 or if your interval is 10, your required number of sample is 10. So, divide 100 by 10 So, the interval is 100 divided by 10 that is 10. So, K is the interval that is equal to N is the true population size and n is sample size. So, randomly selected an integer between 1 to Kth item, Kth interval we require.

(Refer slide Time: 26:25)



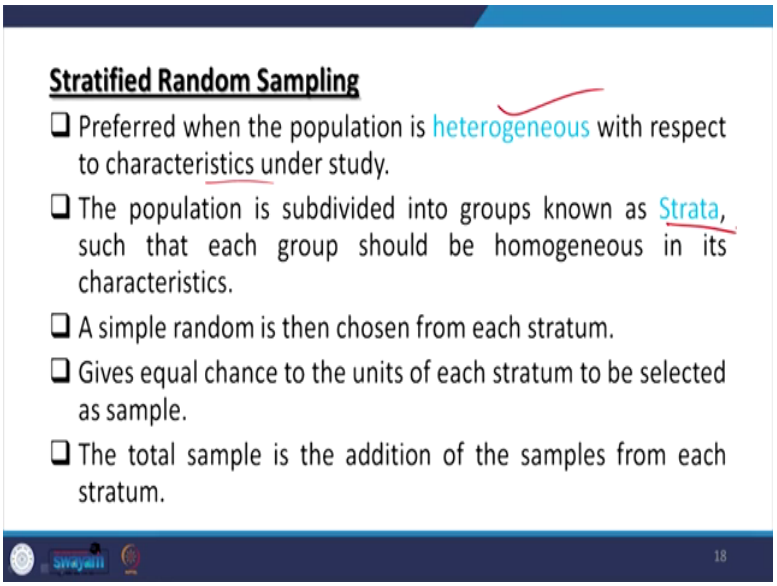
So, let us go to the example if we have a population that is of students 30 we require a sample of 10. So, suppose these are the population given here in this case, it is not exactly 30 here, 12 are given but if it is 30, we require 10. So, the sample interval, the interval is 30 divided by 10 that is 3 for every 3 you are supposed to select 1. So, you will be getting 10 students as your sampled units.

So, K here is three. So, select a random sample between 1 and your ratio from step 2, that is from the sample interval. Here we have selected 2 suppose you have selected first out of the 30. So, 2 then you add the 3 here, 3 plus 2 then 5 has been also selected, 5'th numbers of student has also selected then 5 plus 3 then 8'th is selected 8 plus 3, 11 is selected and so on. So, accordingly, you can able to get your 10 students for your study. So, this is called systematic sampling procedure because you have followed a systematic approach.

There are some problem lies in the systematic sampling, if the first group that is 1,2,3, suppose the first group then 4 to 6 by 3 intervals you have divided then to another group, this is another group then another three another group, it is there, if you have selected the middle one, probably the first cluster itself is asymmetry or it is skewed. If this is skewed, you have selected a particular person from a skewed distribution. In every other cluster, there are high chances of selection at the particular position, not other positions. It may be the case that wherever it is skewed you have selected a person which are not normal.

So, in the systematic random sampling, the errors are expected to be higher sometimes, but it is much better than that of the simple random sampling, if and only if your data is heterogeneous. Here we have different colours that represent different heterogeneous groups. let us come to stratified random sampling.

(Refer slide Time: 29:10)



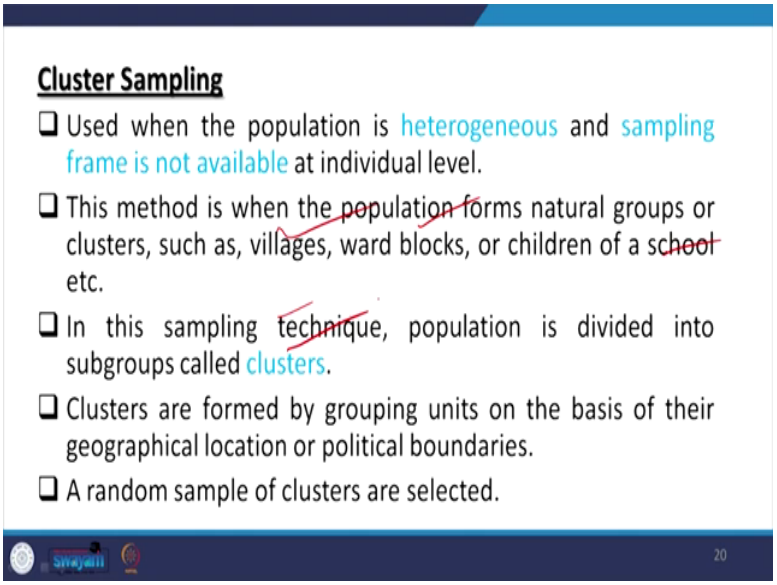
Stratified random sampling is applied when your population is heterogeneous. And we know this is preferred as I already mentioned when the population is heterogeneous and with respect to characteristics under study. So, the population is subdivided into different groups sub-groups, and groups are called strata, such that each group should be homogeneous. Divided into sub-groups like entire population, we know that in our country, we can divide the group into area by location, urban area and rural areas.

We know very clearly in almost all regard; urban areas are different than that of rural areas. In many socio-economic indicators, rural and urban should not be mixed and random sampling should not be taken. First of all, we need to rationalize by making into different strata one is rural and urban.

So, rural has a more or less a similar feature where urban has a similar feature. So, this is one form of understanding stratified sampling. So, a simple random is then choosing from each stratum and gives equal chance, we know since it is expected to be homogeneous within the strata. So, this gives equal chance to the units of each stratum to be selected as a sample.

And we have minimized our procedure we can minimize the number of samples to be covered. Had it been an entire population who followed simple random sampling and generally the entire population is not homogeneous not 100 percent homogeneous. So, you have to take more sample sizes, but in this case sample size can be minimized, because we have defined it in different groups. So, total sample is the addition of the samples from each stratum as I already said

(Refer slide Time: 31:11)



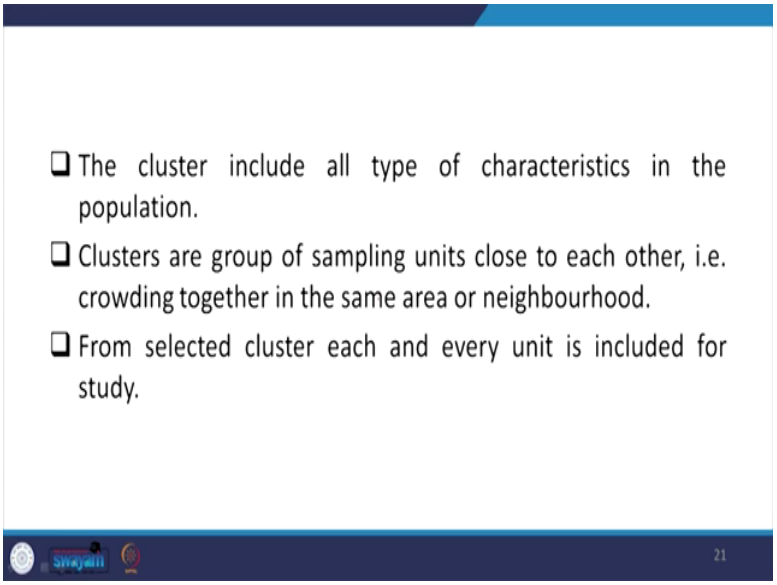
This is an example there are heterogeneous in the population. Heterogeneous in our population is given. We divide with different strata that represent almost with homogeneous groups, three different groups are defined from each group we can then take randomly. Therefore, it is called stratified random sampling, because randomness is still there. In the third part, if you are not referring third part then it is simply called stratified sampling. But if you are including, then you are attaching randomness to it then it is called random sampling.

If you follow after stratified, if you follow another procedure, not random purely again further dividing based on certain other features, then still that is not random. If the final frame is of random then that is called stratified random sampling. Similarly, cluster sampling is there, we will discuss that let me finish cluster sampling and non-probability sampling. We will keep it for the next lecture.

There is non-probability sampling also, we will discuss in the next lecture. So, far as cluster sampling is concerned the population is also heterogeneous but here the sampling frame is

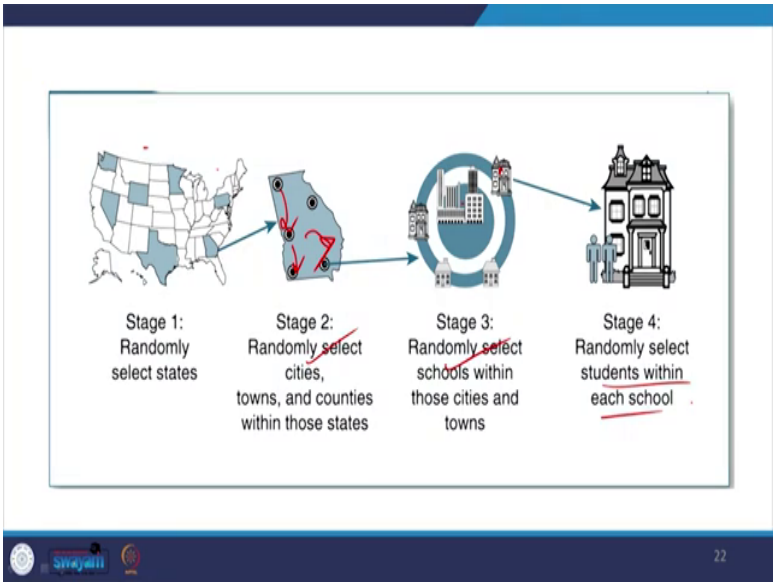
not available at the individual level we simply get the clusters. I will explain what we mean by cluster. like, NSS follows they define into some UFS blocks, blocks are there based on urban areas, rural areas, schools. The sampling technique here is population is divided into subgroups are called clusters, clusters are formed by grouping units on the basis of their geographical location or boundaries. So, as a random sample of clusters isolated not the persons are isolated.

(Refer slide Time: 33:10)



The cluster includes all type of characters in the population clusters are group of sampling units close to each other, crowding together in the same area or neighbourhood are called clusters. From the selected cluster each and every unit is included in this study. This the example.

(Refer slide Time: 33:30)



If the states are united states, for example, all the states have equal chances of selection. So far as policy implementation is concerned, let us select randomly some cities out of it then each city has equal chances of selection. So, each city is comparable because we are assuming that, each clusters have equal probabilities, then randomly select schools within the cities, then schools have an equal probability of selection, then randomly select students within each school. So, randomness is followed therefore it is called a cluster and random sampling. And therefore, it is called probability sampling.

So, multi-stage sampling is there, since we are a little ahead of time, I am deliberately closing here. We will carry forward multi-stage sampling. Even in multi sampling, we are including, stratified, cluster, random sampling there. And we will be discussing, non-probability sampling also we will be discussing probability proportional to size sampling in the next class. With this, let me stop here, will carry forward non-probability sampling and other details of sampling with sample size in the next lecture.

Thank you.