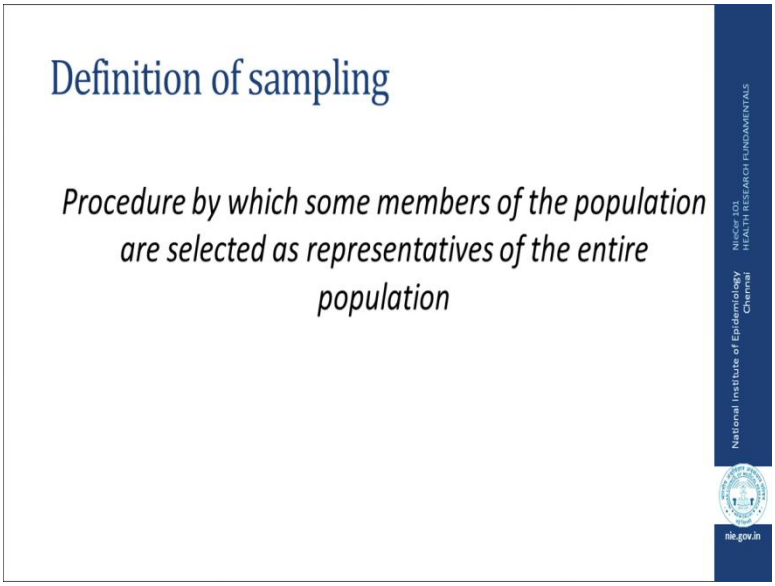
Health Research Fundamentals
Dr. R Ramakrishnan
Former Scientist G
ICMR-National Institute o Epidemiology

Lecture - 11
Sampling Methods

Welcome to Research Fundamental Modules, the NIeCer course one. Today, we are going to talk about some aspects of Sampling Methods.

(Refer Slide Time: 00:22)

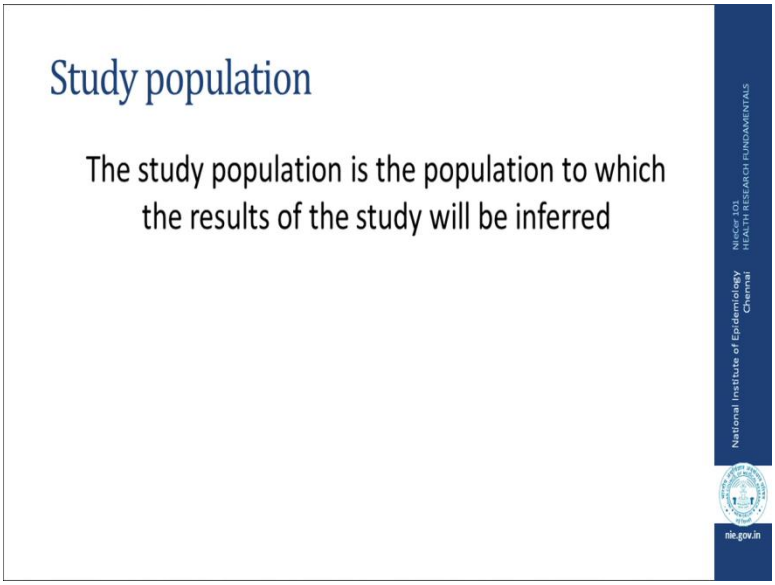


Sampling is really required, whenever you are dealing with a very large population and you want some quick information. Now, look at the definition of sampling. The sampling is a procedure by which some members of the population are selected and they are supposed to be the representative of the entire population.

(Refer Slide Time: 00:47)

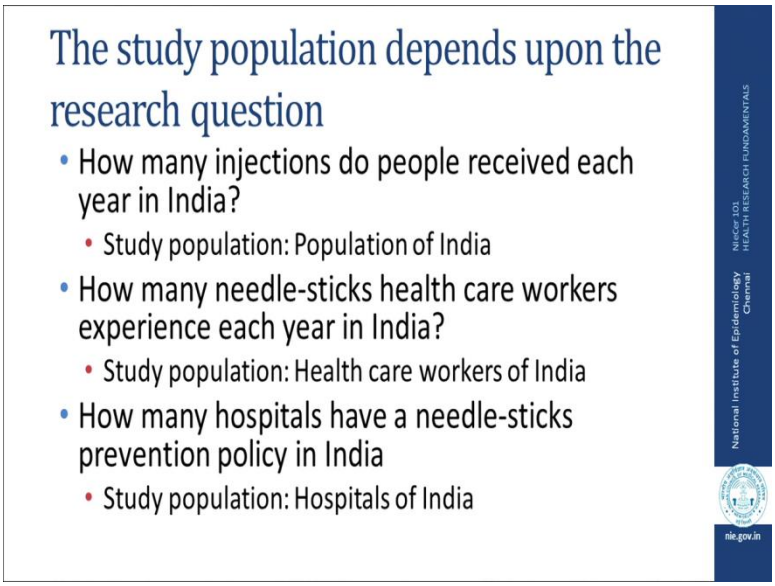
See, if you have a population of people like this and you are looking at a portion of them and that is called a Sample.

(Refer Slide Time: 00:58)



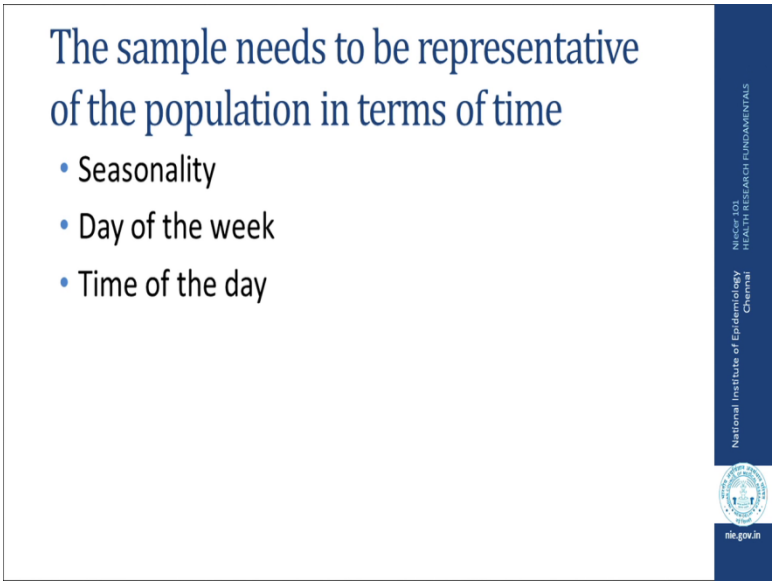
I would like to introduce you to some concepts and one of them is the Study population. What you mean by Study population? The study population is the population to which the results of the study are to be inferred.

(Refer Slide Time: 01:18)



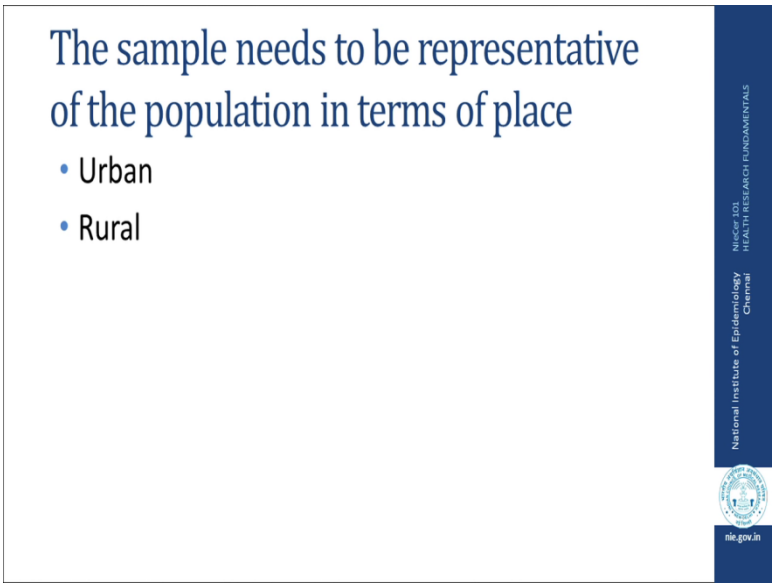
Say for example, how many injections do people receive each year in India? The study population in this case is the entire population of India. Suppose, your research question is, how many needle-sticks health care workers experience each year in India? Then the study population becomes health care workers of India. Suppose, if your study question is how many hospitals have a needle sticks prevention policy in India? Then your study population in this case becomes hospitals of India.

(Refer Slide Time: 01:55)



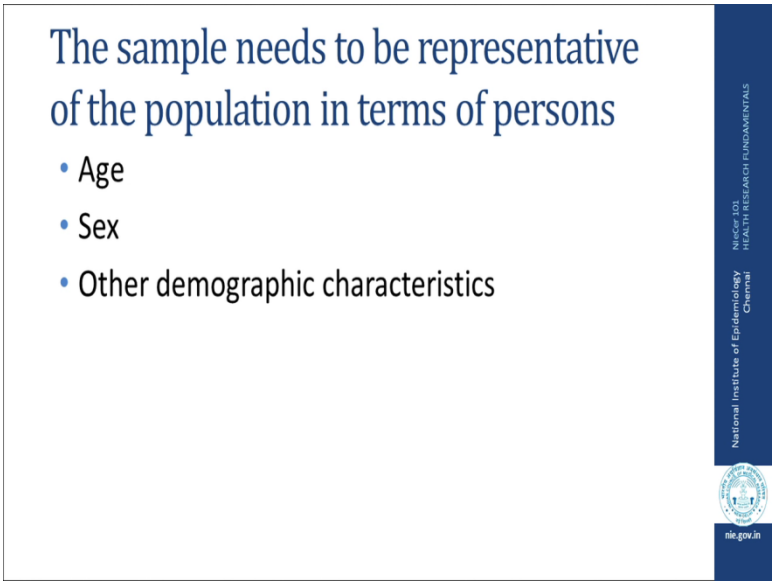
The sample which we select should be representative of the population for which we require an answer and this representation should be in accordance with , seasonality, the day of the week, the time of the week.

(Refer Slide Time: 02:11)

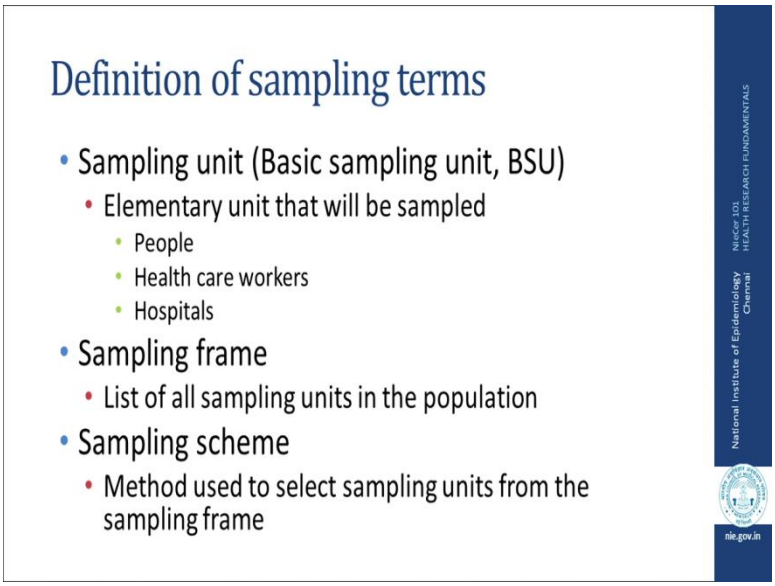


Whether it is urban or it is rural, or it is should rather match the composition of age, sex and other demographic characteristics of the population.

(Refer Slide Time: 02:14)

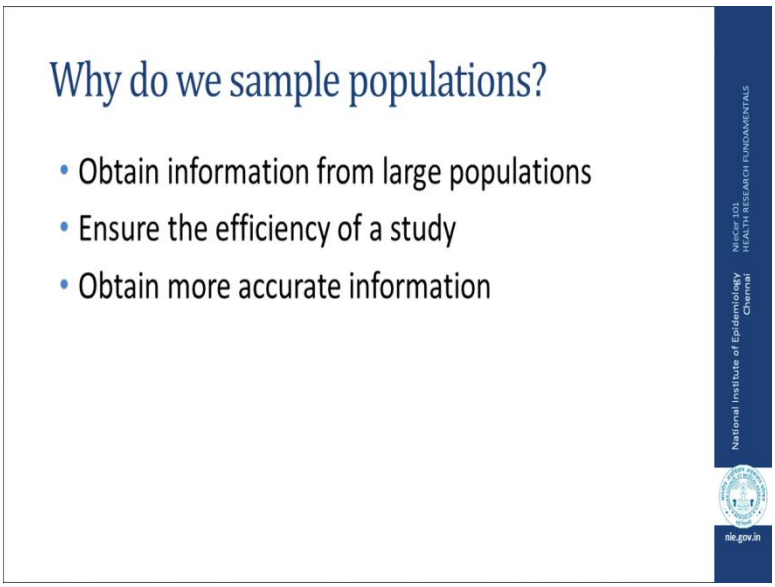


(Refer Slide Time: 02:24)



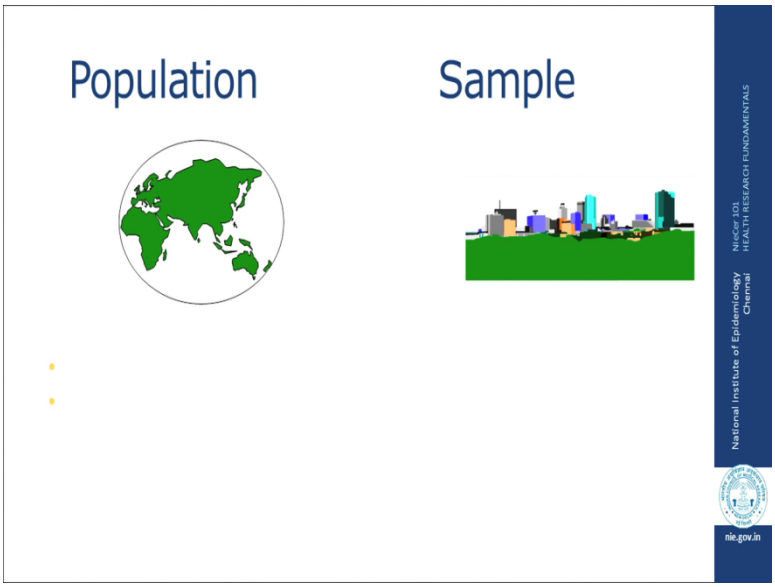
See now: let us introduce you to some concepts or terminologies that are often used in the sampling parlance. What do you mean by Sampling Unit? Sometimes it is called basic sampling unit, BSU. These are the elementary unit that will be sampled, that could be people or health care workers or hospitals as we had seen in our early example. What do you mean by Sampling Frame? The sampling frame is a list of all sampling units in the population and what do you mean by Sampling Scheme? The sampling scheme is a method used to select sampling units from the sampling frame.

(Refer Slide Time: 03:12)



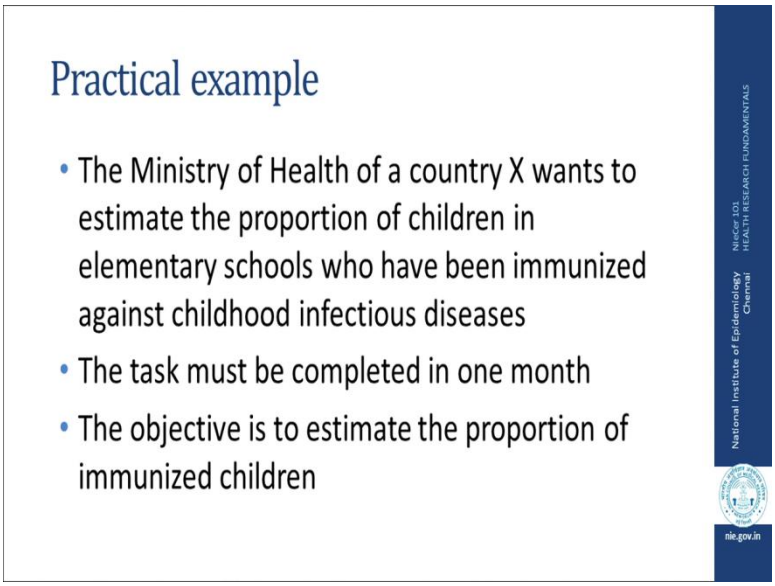
So now, there are different ways how or why we should do the sample populations? If you have enough resources, you can probably study the entire population. But still, even if you have the resources, it is not wise to study the entire population because often the population is very large and a large population when you are going to collect information one of the major constraint could be the time. You may require lot of time to collect information and you may say I will employ lot of people to do that, but what would rather probably happen is if you have lots of people collecting information there could be lot of inter observer variations which could rather add on to a tremendous amount of error and unfortunately you cannot measure the amount of such errors. So, it often rather happens that by doing a sample survey you often get accurate information. The information that you get from sample surveys are more accurate than the information you do on a large scale population studies.

(Refer Slide Time: 04:36)



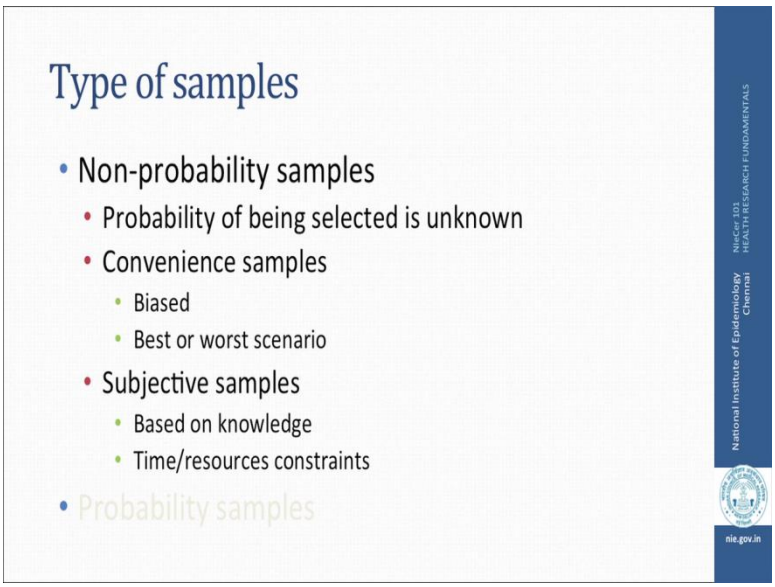
So, a population could be an entire universe, whereas a sample could be as selected a small regions.

(Refer Slide Time: 04:44)



Let us look at a practical example. Suppose, the ministry of health of a country X wants to estimate the proportion of children in elementary schools, who have been immunized against childhood infectious diseases. You could just imagine, you know the proportion of children of all elementary schools, who have been immunized against childhood infections of a country. So that is a task but one of the conditions that he has put is the task must be completed in one month. So, the objective is to estimate the proportion of immunized children and you want the results in a month's time.

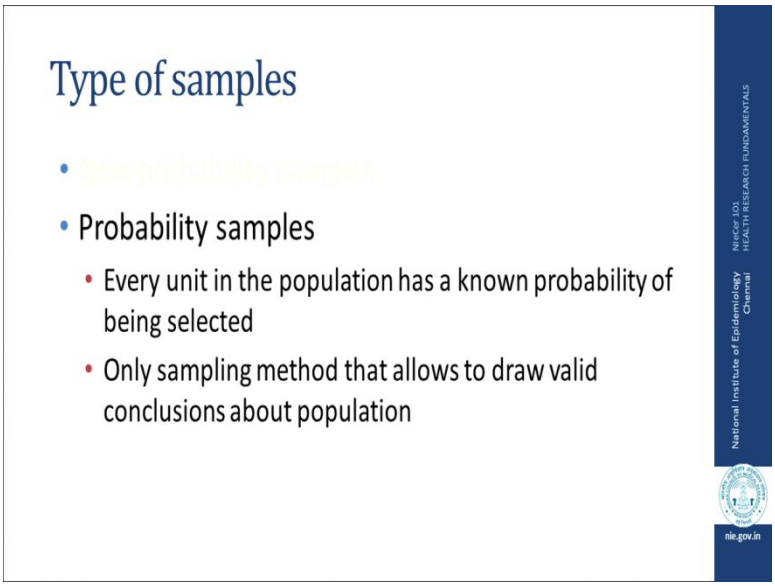
(Refer Slide Time: 05:24)



Now, let us rather look at the different ways; how you can rather get this information? Or in other words what are the different types of samples that could rather group. See, broadly speaking the sample could be a Non-probability Sample or a Probability Sample. What do you mean by a Non-probability Sample? Non-probability sample is the probability of being selected that is a sample, the probability of being selected for your study is not known. It could be a convenient sample or purposing sample, you just rather convenient whatever the region that is convenient to you, close by to your place, you can rather go and rather see first 100 people that you come across that could be a convenient sample.

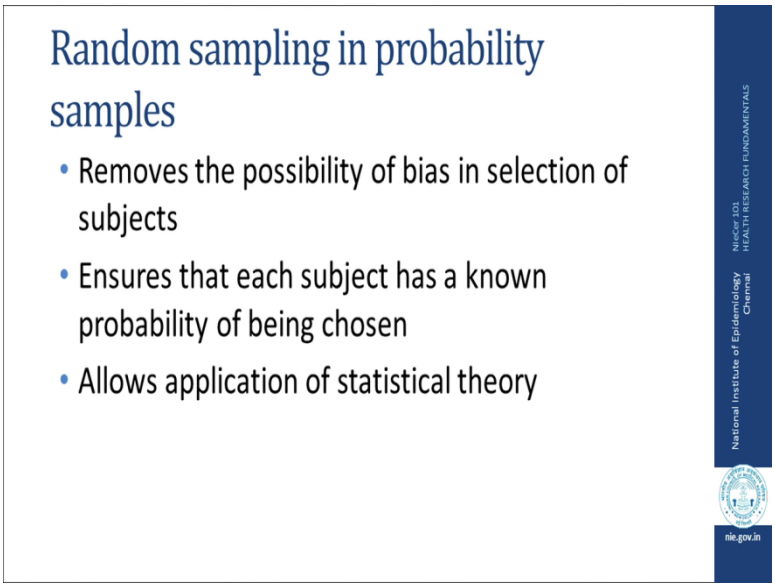
What could rather happen? That sample could be biased or it can rather give either a best or a worst scenario, people you know it is a convenient location. You may get rather the results very different from a location, which is not very convenient or which is very remote and difficult to approach. And also, some of these are all very subjective samples and to derive some objective criteria from a subjective sample is always difficult. But, nevertheless these non-probability sampling methods still are useful and that is being extensively used mainly to generate hypothesis or to prepare for more systematic probability samples. Now let us look at, what do you mean by Probability Samples?

(Refer Slide Time: 07:07)



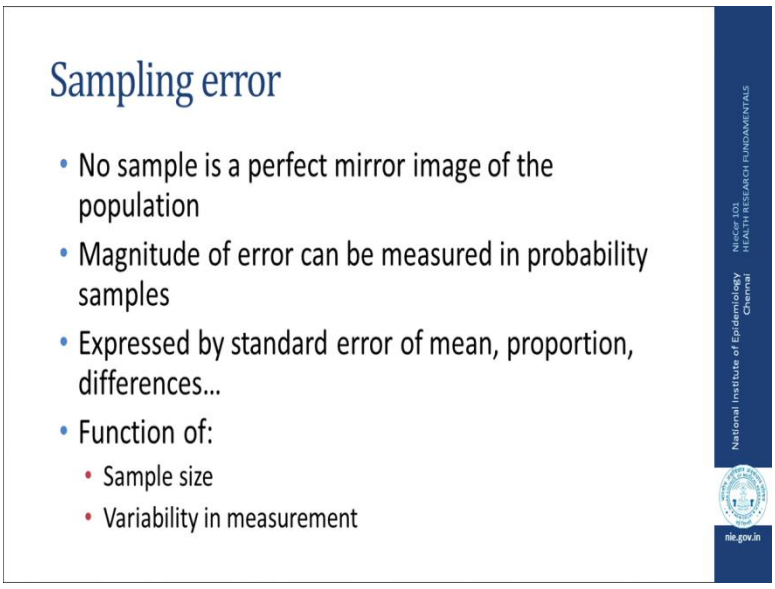
In a probability sample, every unit in the population has a known probability of being selected. What is the advantage? This is only sampling method that allows to draw valid conclusions about the population.

(Refer Slide Time: 07:30)



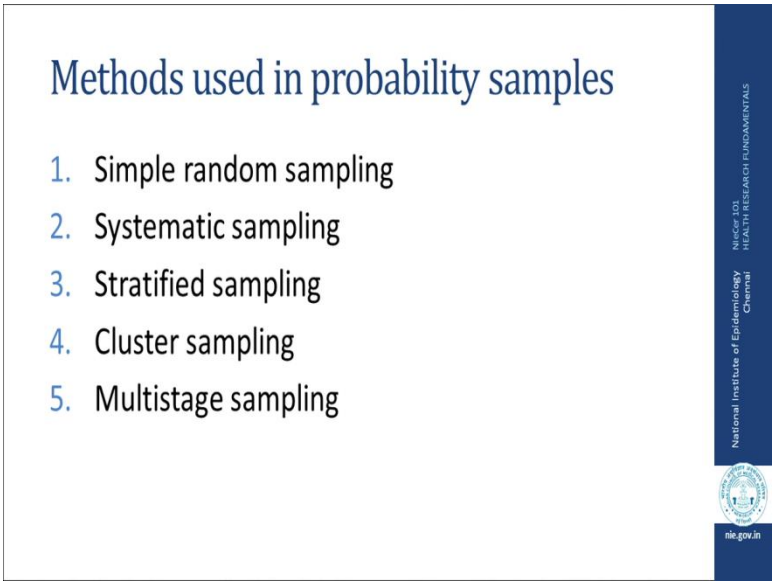
It removes the possibility of bias in selection of subjects and also ensures that each subject has a known probability of being chosen. It allows application of statistical theory because many of the statistical text that you do it insist on a random sampling and these tests are valid only if the samples are a random sample.

(Refer Slide Time: 07:54)



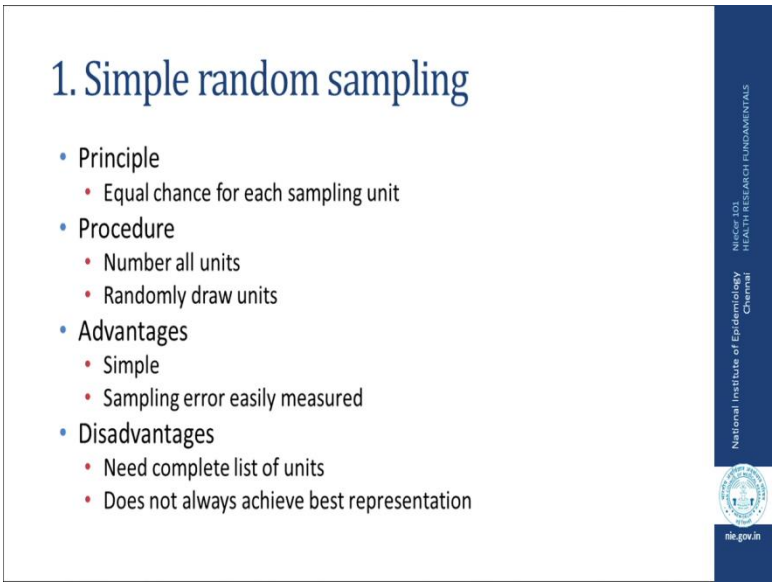
I would like to rather introduce you to the concept call Sampling Error. No sample is a perfect mirror image of the population. Always you know when you pick a sample from a population and when you look at the results, it may not be exactly the same as the results in the population. But, fortunately the magnitude of error could be measured in terms of probability in the case of probability samples. This is expressed by standard error of mean or proportion or differences and that is a function of the sample size and then the variability in the measurement. So, sampling error is a very important component in sampling theory, which helps us in identifying the sample size and things so on.

(Refer Slide Time: 08:56)



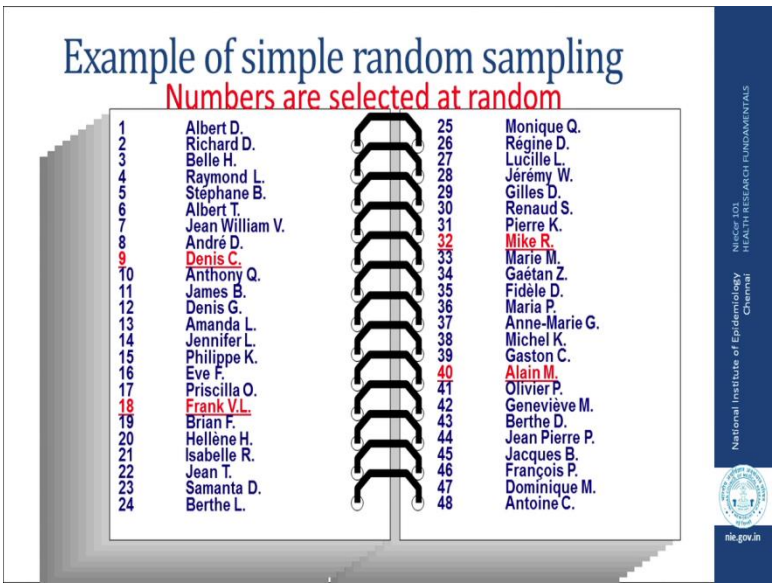
Now, let us look at some of the popular sampling methodologies that are employed in sample service. Let us rather look at the first Simple random sampling.

(Refer Slide Time: 09:11)



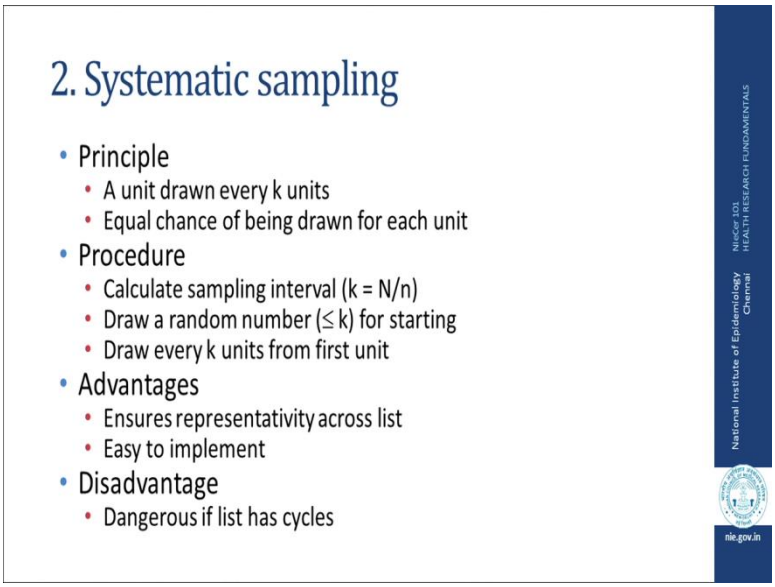
As a name suggests, it is a very simple sampling procedure, very easy to understand in which every individual sampling units have got an equal chance of being included into a sample. How do you do that? We number all the units and we randomly draw units. The advantages as I mentioned, it is very simple and sampling error is also very easily measured. Major limitation of this is, you need to have a compete list of all units. Many times it may not be available and also some times you may get a sample, which is very different from the whole population may not be very representative of the population.

(Refer Slide Time: 10:07)



See, an example of a simple random sampling could be if you have the list of all say about these 48 names, you pick a random numbers of 9, 18, 32, and 40. So these are all the names that are selected as your sample.

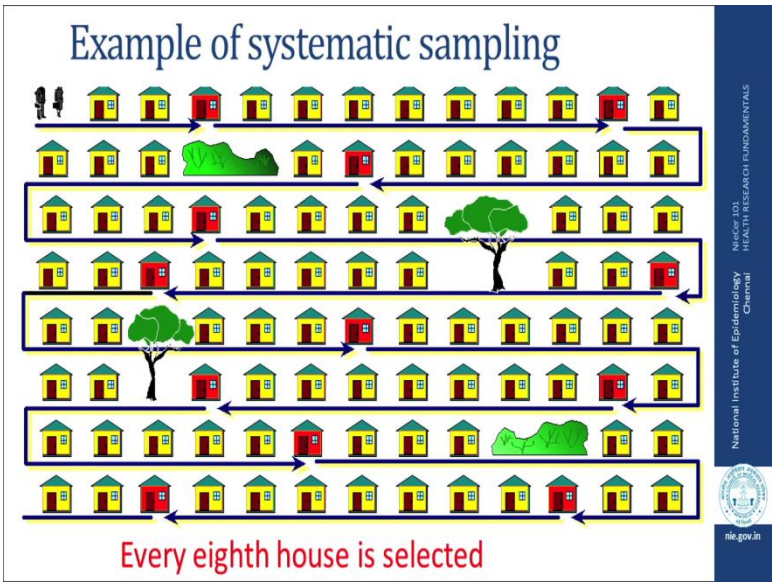
(Refer Slide Time: 10:21)



Now, the next sampling type is a Systematic Sampling. A systematic sampling. what is rather done is, initial sampling unit is picked by random and then every k^{th} unit from that from your population are examined. A unit is drawn and every k units and every equal chance of being select for each of the unit. So, you calculate the sampling interval called k, which is divided by N divided by the number of sample size that you require. And you draw a random number which is less than or equal to k for starting and draw every k units from the first unit.

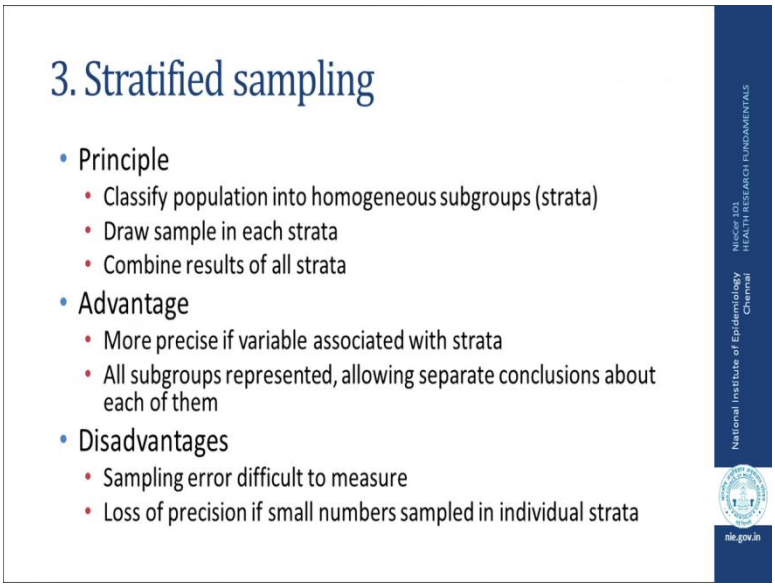
What are the advantages? It ensures representatively across the list. It is easy to implement. You gave a worker that you say, you start from this house and every 10th house you go on rather see and cover all the houses, its very easily been done. If there is some sort of a cycle of some specific characteristic that you are studying then you might probably get a sample, which is very atypical in a systematic sampling and also some of the statistical measures that you are going to compute, it is difficult when you are going to have systematic sampling, where you do not have an exact formulas, you may have to use some approximate formulas.

(Refer Slide Time: 11:53)



The example of a systematic sampling is you see, in the first, the red house is selected and then every eighth house from that is selected and all the red houses in these houses are your selected samples.

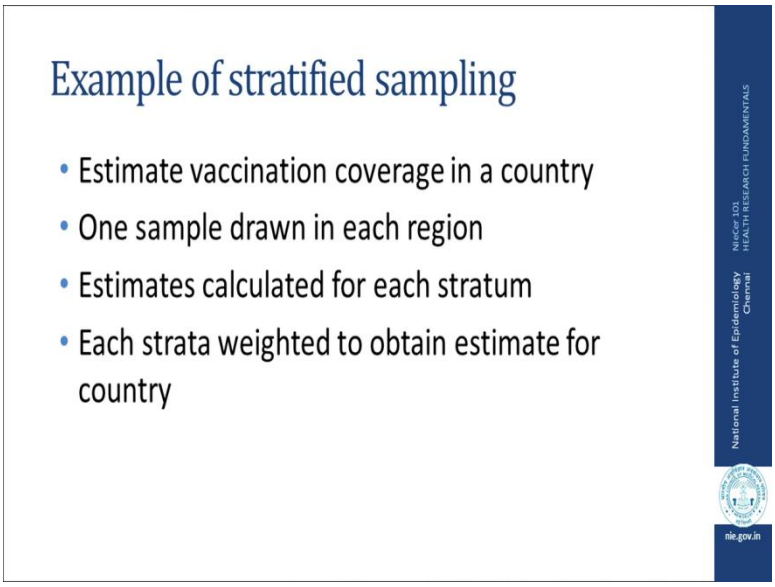
(Refer Slide Time: 12:13)



There is a sampling method called Stratified Sampling. The principle of it is, who classify population into homogeneous subgroups, which are called 'strata' and you draw sample from in each strata combine the results of all the strata to get an idea of the whole population. The advantage of it is its more precise, if variable associated with strata and

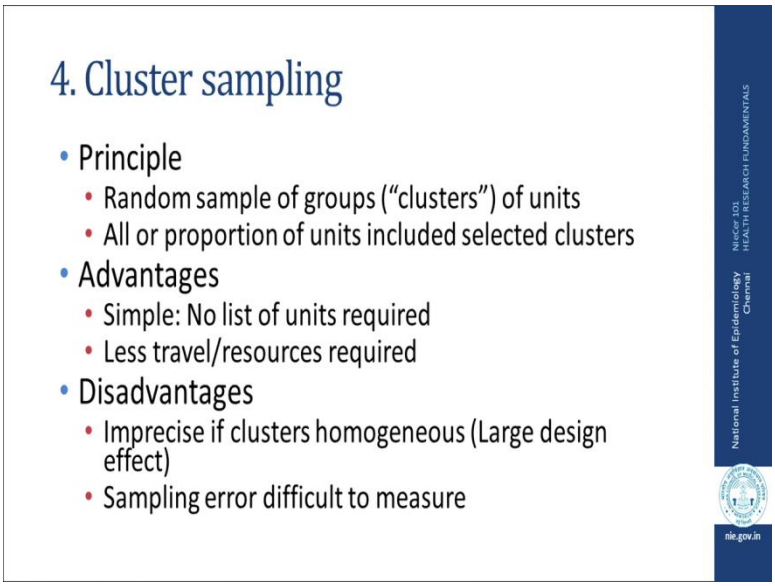
all subgroups are represented, allowing for separate conclusions about each one of them. Suppose, a natural strata could be male and female, so you have an estimate for male and you have an estimate for female and you can have an estimate for a combine male and female for the whole population. But the disadvantage is, sampling error is difficult to measure, and that could be loss of precision, if you are going to rather have a lot of strata and for each strata you have small numbers in it.

(Refer Slide Time: 13:13)



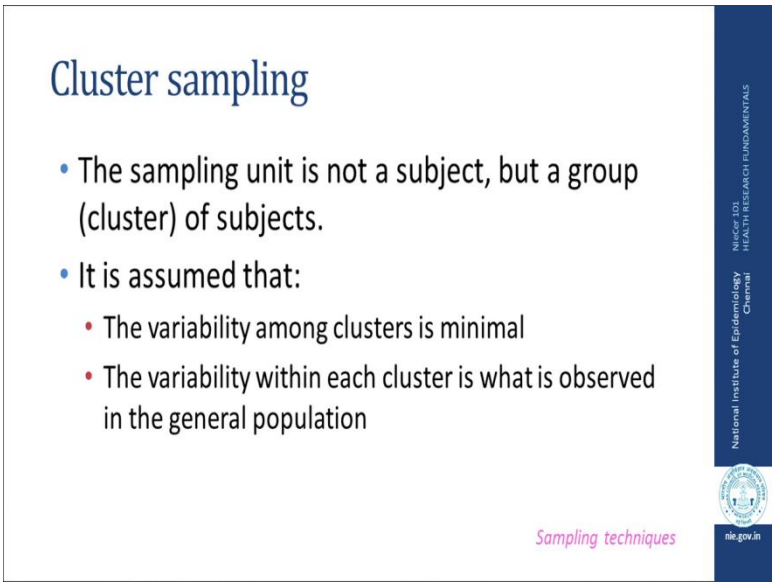
Example of a stratified sampling is, suppose if you want to estimate the vaccination coverage in your country. One sample drawn from each region north, east, south and west and the estimate calculated for each of the stratum and at the end you can weight the stratum according to the size of the regions.

(Refer Slide Time: 13:36)



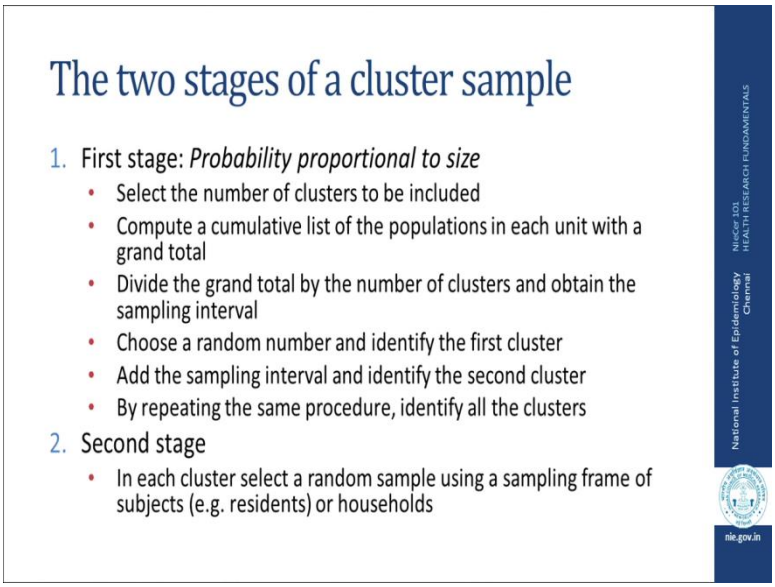
Another important type of sampling, which is very popularly used in the health surveys in research, is called Cluster Sampling. The principle of cluster sampling is that, a random sample of groups or a cluster of units, and all proportion of units are included in these selected clusters. Its advantages is, it is simple, we do not require a list of units and less of travel or resources are required because you are going to collect a cluster and you are going see only within the clusters. And the disadvantage is, if the clusters of homogeneous then it may result in a large design effect. All the people in the sample may have very homogeneous results which could result in a design effect and sampling error is difficult to measure in a cluster sampling.

(Refer Slide Time: 14:43)



The sampling unit is not a subject, but a group or a cluster of subject. The assumptions here it is, that variability among the cluster is minimal. The variability within each cluster is what is observed in the general population.

(Refer Slide Time: 15:00)

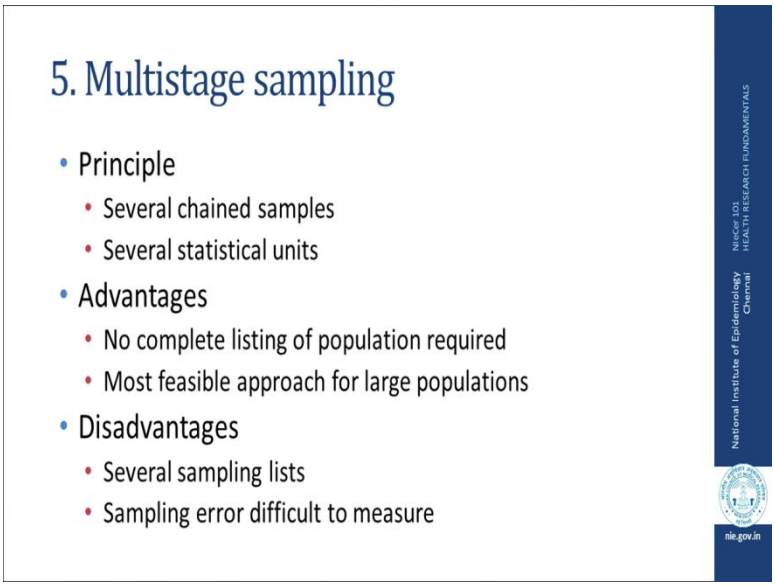


Now, how these clusters sampling is usually done. It is done as two stage approach. In the first stage, a probability proportional to size, that is select the number of clusters to be included, compute a cumulative list of all the population in each unit with a grand total, divide the grand total by the number of clusters and obtain the sampling interval,

choose a random number and identify the first cluster, add the sampling interval and identify the second cluster and so on and by repeating the same procedure, identify all the clusters.

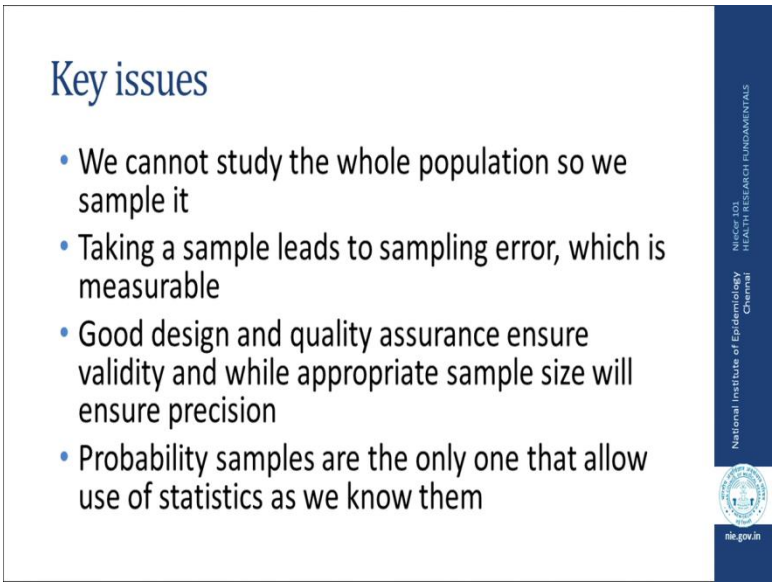
Once your clusters are identified, then in the second stage in each cluster, you select this random sample using the sampling frame because as I had mentioned you earlier on simple random sampling when you want to do a simple random sampling you need to have all the list of the your sampling frame. So, in a small cluster it is possible for you to formulate the sampling frame and you can select people from that sampling frame on a random basis.

(Refer Slide Time: 16:14)



Another important sampling methodology employed is called a Multistage Sampling. In this multistage, especially in a very large, you want some estimates for at the national level, you need to do a sampling in several chains samples and several statistical units are there. The advantage is, there is no complete listing of the population is required and it is most feasible approach for large populations. The disadvantage is, there are several sampling units and sampling error at times, it is very difficult to unless you follow certain very specific methodologies for selecting at each stage.

(Refer Slide Time: 17:02)



Some of the key issues that I would like to bring to you is we cannot study the whole population so we sample it. Whole population studying could impact a result in inaccurate results so taking sample leads to sampling error, but which is easily measurable and we do not have a measure for non-sampling error, whereas we have a measure for sampling error. Good design and quality assurance ensure validity and while appropriate sample size will ensure precision. The probability samples are the only one that allows the use of statistics as we know them and so it is always advantage to use a probability sample so that you can have a valid conclusion, a precise conclusion and also you can employ statistical test on them.

Thank you so much.