

Handling Large-Scale Unit Level Data Using STATA
Professor Pratap C. Mohanty
Department of Humanities and Social Sciences
Indian Institute of Technology, Roorkee
Lecture 07
Review of Sample Techniques-I

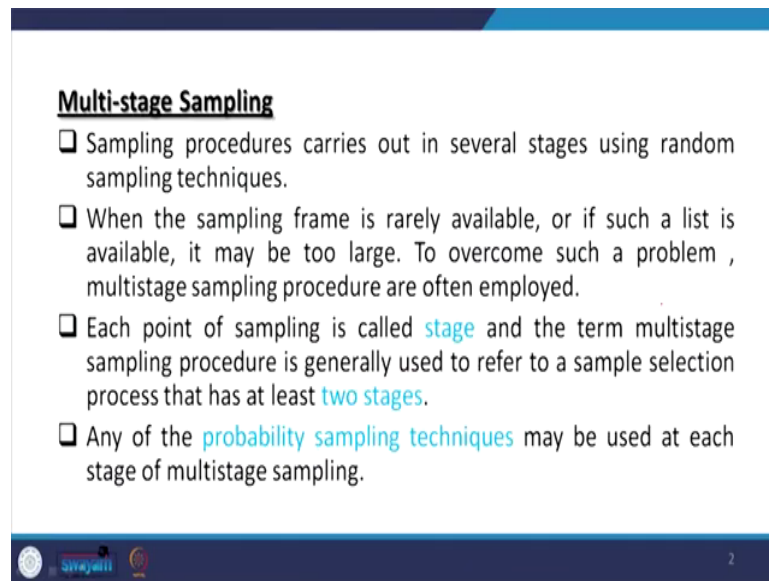
Welcome learners once again to the NPTEL module on Handling Large-Scale Unit Level Data with STATA. We are at the verge of explaining the unit-level data and its backgrounds. So, we are discussing in the first week though we have discussed all available unit-level data in India. We are discussing or approaching towards using STATA though in the next week, but we are operating through understanding the sampling design because if you do not know sampling, it is completely incorrect to start with STATA. So, this week, week number 2 is designed in such a manner that you can easily understand sampling techniques correctly.

Because for a researcher or for a Ph.D. scholar especially or even policymaker, it is very important to minimize the sample error. Always reduction in the sample error or less the sample error always gives better results for interpretation. So, we are reviewing the sampling techniques. In the last lecture particularly, I took you to the discussion on types of sampling.

Especially, I have already started with the discussion of probability sampling. In the probability sampling, we have discussed simple random sampling, then we also discussed stratified sampling, systematic sampling, and also we try to understand what is called cluster sampling.

So, these four techniques we have covered already. Some little parts of the probability sampling is still left which is highly utilized in the NSS survey or even other large-scale survey data. So, that is none other than multistage sampling. So, these days' multi-stage sampling is very very important. And let me start with the multi-stage sampling without delay further because this is going to be very useful.

(Refer Slide Time: 02:51)



Multi-stage Sampling

- ❑ Sampling procedures carries out in several stages using random sampling techniques.
- ❑ When the sampling frame is rarely available, or if such a list is available, it may be too large. To overcome such a problem , multistage sampling procedure are often employed.
- ❑ Each point of sampling is called **stage** and the term multistage sampling procedure is generally used to refer to a sample selection process that has at least **two stages**.
- ❑ Any of the **probability sampling techniques** may be used at each stage of multistage sampling.

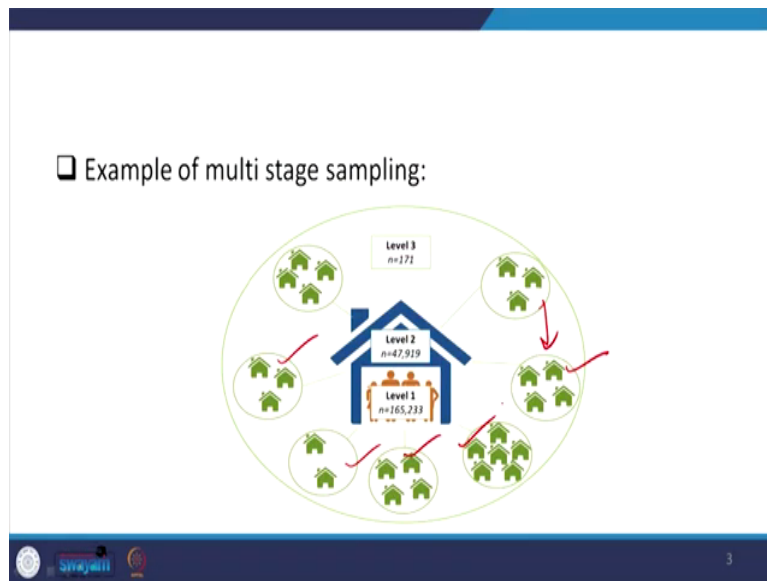
swayam 2

First point in the multi-stage sampling is, let me first give you the general understanding of multi-stage sampling, the word is very clear, this is where we are clarifying through the stages of sampling. The stages of sampling like it might include randomness, it might include stratified stage, it might include systematic roots, it might include cluster phase. If you club all the different tools of sampling in one umbrella, then generally these are called multistage sampling, but there are certain structures, a systematic structure of understanding multi-stage samplings. So, let me go through point by point and explaining the concept very clearly to you.

Sampling procedures carries out in several stages using the random sampling technique. When the sampling frame is rarely available or if such as list is available, it might be too large usually to overcome such a problem, because if it is too large in handling or proceeding through the sampling, it might be very expensive. Since it is too large, it is always suggested to go for a multi-stage sampling procedure and these are often employed in the present age. Each point of sampling is called stage.

The term multi-stage sampling procedure is generally used to refer a sample selection process that has at least 2 stages, minimum of 2 stages. Any of the probability sampling techniques may be used at each stage of multi-stage sampling as I already discussed from the beginning.

(Refer Slide Time: 04:38)



One such example of multi-stage sampling employed by NSS. National Sample Survey is through different levels, level 3, 2, and 1. So, level 3 where I have already discussed usually the sample sizes less, level 2 it is a little higher and level 1 is even much higher. What do you mean by level 3? Suppose we wanted to, especially in the NSS, we wanted to observe their sampling frame correctly. So, in level 3 they stick to the area divided in rural or urban. So, it is not just random, it is not utterly random it is called a stratified format.

So, level 3 may be referred as stratified, then at the disaggregate level they considered the household, in the level 2 it is further disaggregated. You can find out number of units here for our understanding, it is higher in the level 2 as compared to level 3, level 3 the number of samples is lesser because number of rural areas and urban areas are relatively lesser as compared to the households.

So, in the level 2 we are referring to the households, whereas the level 3 then further disaggregated level is in this context is individual. So, the individuals here, we are considering different possibilities and the number of individuals are expected to be higher because in the population itself the total number of individuals is higher.

So, the sample size accordingly also increases. So, level 3, level 2, and 1 identify different forms of or different stages of sampling. IT depends upon which exact sampling technique is adopted. Maybe at the level third, we adopted stratified, on the second, on another tier of the sampling frame, we might have adopted again on stratified technique, on the final one we may adopt any form depending upon the context we will explain in our succeeding slides.

Let us move to the understanding of a new technique which has been adopted recently by NSS and others like IHDS data set, even NFHS has already adopted probability proportional to size sampling. This is also called probability proportional to population, in the place of size NSS considered it as population is called PPP. What is the rationale behind considering probability proportional to size? First of all, this is adopted to minimize the number of samples or minimize the expenditure on sampling and there is proper reasoning based to reduce the sample size based on certain proportional indicators probability proportional indicators.

(Refer Slide Time: 08:22)

Probability Proportional to Size (PPS)

- If there are more than one subpopulation with varying size of entities each, PPS sampling ensures that the probability of an entity being selected as sample is proportional to the size of its subpopulation.
- Alternative to stratification.
- Steps:
 - List of all clusters (villages and sector/wards) is made.
 - Population of each cluster is written against them
 - Cumulative population is then written in serial order.

steps

S.N	V.A	C.P
1	n ₁	n ₁
2	n ₂	n ₁ +n ₂
3	n ₃	n ₁ +n ₂ +n ₃
4	:	:
5	:	:
:	:	:
1	:	:

Let me go through the points correctly if there are more than one subpopulation with varying size of entities, PPS sampling ensures that the probability of an entity being selected as a sample is proportional to the size of its subpopulation. So, a particular unit or entity has certain probabilities of consideration is proportional to the size of its subpopulation. So, we will explain, though it is not so clear, we will explain systematically to clarify the concept. So, this is also sometimes considered as an alternative to stratification because just stratification may not consider proportional representation.

So, the steps involved in the probability proportional to size sampling include list of all clusters. So, clusters may be villages or sector or what the way we have just discussed before and population of each cluster is written against or corresponding to that of the clusters. So, let us consider one excel sheet. In the excel sheet, we can enter the details carefully. This is serial number 1, we can write down serial number 1,2,3,4, and so on.

And we can make like the cluster or the villages, let it be villages, cluster number villages, and what do you do? we have to write down the population, population of the villages. If our cluster is defined as village then let us write down its population. Where to get the population? From the census, it is always suggested to go by the latest census 2011 those who are doing some research in this regard. So, let this is N1 this is N2 likewise these N3, likewise, if this population is there, then on the third column, we will follow a cumulative population (CP).

In the cumulative population, this will be N1. So, this is N1 plus N2 plus N3 so on, the cumulative population should be calculated. These are some steps to be followed, it is very helpful, steps to be followed to select a particular sample with a probabilistic structure.

(Refer Slide Time: 10:54)

$d = \frac{N}{10}$
 Sampling interval is calculated = total cumulative population/ no. of clusters.
 Choose the random number between 1 and SI. This is the **random start (RS)** the first cluster to be sampled .contains this cumulative population.
 Calculate the following series: RS; RS+SI; RS+2SI... RS+ (d-1)*SI
 The clusters selected are those for which the cumulative population contains one of the serial numbers.

clusters → Randomize → systematic

How to do it? Then what will we do, let me first go by the points. We have to define a sampling interval. What do you mean by sampling interval? So, once we derived the cumulative population, we will get the total N (capital N) here and N divided by the number of clusters you require. How many clusters you define? These are not clusters initially we define the villages by their population and these are the real figures collected from the census data. In the last column we will figure out which sample to be selected based on the cluster. Suppose, we want to divide the entire population into 10 clusters. So, divided by 10.

So, each cluster how many population is defined. So, this is N divided by 10. Choose the random number between one and sample interval. So, first suggestions here. The next step is to choose any number between 1 and 1 to N upon 10, if 10 is the sample interval. So, after

selecting these, how to select that random number? It might be random start or you may take the help of any tools through the computer random table, random tables are available or from the Google random number chart, it will suggest any number, you simply pick up, it is called random number start.

That is also called random start, the first cluster to be sample contains this cumulative population. Whatever is coming out of the total population, suppose, let me mention in the table itself. Suppose, it is of some value out of the total population you have got a particular figure and find out in which particular serial number it falls. Suppose, it falls in serial number 2, not in 1.

So, that is your starting point of selection. So, let me mention, that is the first starting point may be within serial number 2. Then what will you do, how to find out the exact sample? The sample will be your random sample selected from the beginning that is the first sample you have collected then you simply add random samples that is let it be within the serial number 2 you have figured 1 particular number.

And that particular number corresponding to N_2 within that N_2 a particular household is selected, if your household or is your unit of survey or here it is village is your unit of survey. So, the N_2 is selected, N_2 plus your sample interval. If you add it that will correspond it to another serial number. Let it be another serial number falls maybe in 3 not in 2 now, it in 3.

So, that one will be selected. again you will add 2 times RS plus 2 times a new sample interval RS plus 3 times a sample interval and you will keep on doing till D minus 1. D stands for the N by 10, we have already divided N by 10 is your D.

So, D is nothing but number of clusters. So, number of clusters times the sample interval if you will make it that will be the units selected for your work. This is 1, this is second unit covered, the third will be covered, and so on. And accordingly, number of samples will be collected. So, what we followed here? In this method we have followed, first of all we have followed a systematic addition of the population, a cumulative population and we divided into different clusters, why we divide into cluster we wanted to make a random.

As of each cluster is an equal chance of selection. So, randomness is attached, cluster approach is attached and we also followed a systematic sampling design. So, this method includes 3 important aspects, one is your clusters design, then randomization, then the third one is systematic approach because you keep on adding the sample interval to the first unit

selected or to the second unit selected. So, the 3 methods simultaneously included in the case of probability proportionate to size sampling.

(Refer Slide Time: 16:17)

❑ For Example,

- ❑ In PLFS 2017-18 , Urban first stage units (UFS Blocks) were selected by probability proportional to size scheme with replacement (PPSWR), size being the number of households in UFS blocks. Samples for a panel within each stratum were drawn in the form of two independent sub-samples. To implement the rotational scheme, 4 groups of sample FSUs of equal size (each multiple of size 2, half for each of sub-sample 1 and sub-sample 2) were drawn randomly
- ❑ In the rural areas, samples for a stratum/sub-stratum were drawn randomly in the form of two independent sub-samples with probability proportional to size with replacement (PPSWR) scheme, size being the population of the village and equal number of samples were allocated among the four quarters.

For example, as considered in periodic labor force data of 2017-18, the urban first stage unit we have already discussed earlier that is, UFS blocks were selected by the probability proportional to size scheme with replacement. Why with replacement is discussed? There are two approaches one is called without replacement or with replacement. since randomness is there, I already said they have made of randomization. In the random sample case, there are two approaches, one is called with replacement or without replacement.

So, in with replacement, what is going to happen? In the with replacement case your N remains same, the total N , suppose number of clusters are 10. With replacement suppose one cluster is selected how many are left 9 only 9. So, 10 minus one is left. if you are following a with replacement technique, your 10 is still left. So, your probability of selection remained constant. Whereas in case a with replacement your total N becomes N minus one, again in the next time it will be N minus 2. So, the probabilities are not same.

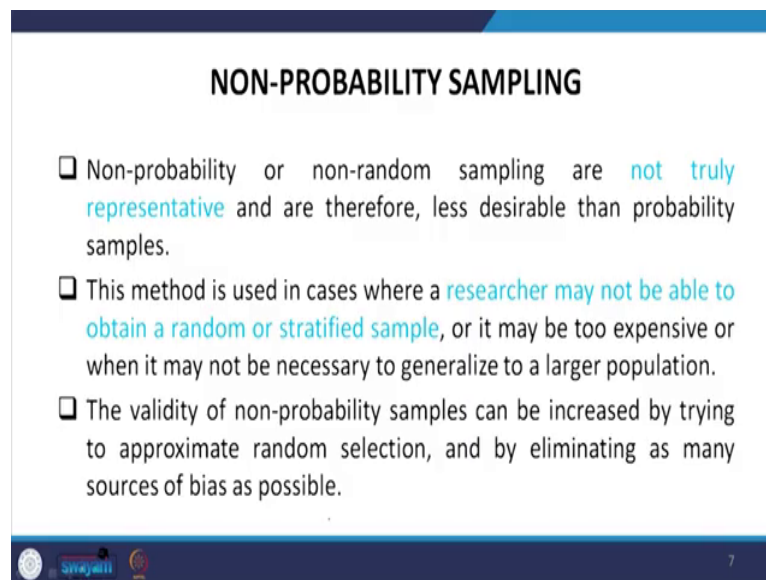
So, in the probability proportional to size sampling, with replacement is used in the periodic labor force data that is of 2017-18. And with replacements because of that, they followed urban first stage units with a probability proportional to size scheme with replacement size being the number of households in UFS blocks.

So, size we have already said, number of populations they consider as the size. Sample for a panel, within each stratum, were drawn in the form of 2 independent subsamples. To

implement the rotational schemes four subgroups of samples of FSU is of equal size, they are followed and were drawn randomly they are divided into four subgroups and they have drawn it randomly.

Similarly, in case of rural areas, sub-samples are also defined, and again they followed with a replacement, probability proportional size with replacement scheme. And size being the population of the village, an equal number of samples were allocated among the four quarters, because why four quarters are said, four quarters the periodic labor force followed four quarters and as I have already mentioned, in the rural areas, the persons are different. persons are different in successive quarters, whereas, in case of urban areas, the same persons are repeated in the successive quarters. So, accordingly follow the sampling frame.

(Refer Slide Time: 19:29)



NON-PROBABILITY SAMPLING

- ❑ Non-probability or non-random sampling are **not truly representative** and are therefore, less desirable than probability samples.
- ❑ This method is used in cases where a **researcher may not be able to obtain a random or stratified sample**, or it may be too expensive or when it may not be necessary to generalize to a larger population.
- ❑ The validity of non-probability samples can be increased by trying to approximate random selection, and by eliminating as many sources of bias as possible.

What about nonprobability sampling, we already covered probability sampling? Let us move to the understanding of nonprobability sampling. So, the nonprobability the word itself clarifies the meaning that is the probability of selection of the particular unit in the sample is not same. So, nonprobability or the non-random sample are not truly representative to the population. So, this is less desirable than the probability sampling. Those who want to make a policy design for a larger population, you have developed certain ideas certain outcomes from your experiment, but the experiment may not be representative to a larger group.

So, your sampling method you have adopted may not be representative of larger group. So, may not be feasible for policy, if it is targeted for universalization approach, but if it is targeted for a particular cohort or for a particular subsection and your experiment is meant for

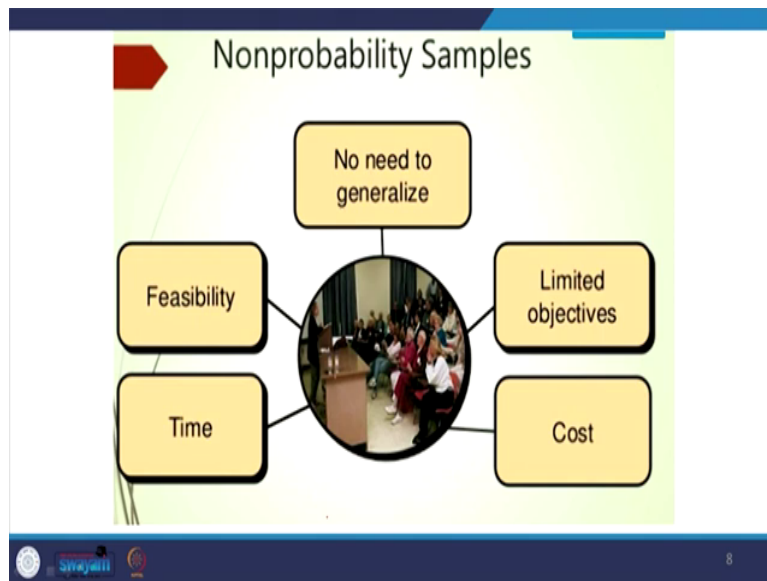
the subsection, then it is not problematic and it is perfectly fit. Let me give you the background of publication these days, the editors of different top most journals, usually expect your result must be representative enough. Except few, there are certain experiments made an experiment very expensive.

So, some experiment may not be representative because that is targeted for a particular group. But, it is observed from many papers, many journals that they expect as if your result is representative and it calls for many policies implementations. So, it is our attempt every time even if it is a nonprobability structure, the sampling frame is of nonprobability type, but there are some techniques to make it more generalizable. So, let us go through further.

So, this method is used in cases where a researcher may not be able to obtain a random or stratified sample, it may be too expensive. Usually what happens if you want to generalize every time, it is also expensive, you have to consider more sample size in order to generalize. But in this case a particular group is tested, for particular behaviour is tested, particular cases are tested.

So, we need not require large sample size. And moreover, this is more qualitative and so, the validity of the nonprobability samples can be increased by trying to approximate random selection, as I just said a couple of minutes back, and also by eliminating as many sources of bias as possible.

(Refer Slide Time: 22:34)



nonprobability samples require time, it also requires cost, it has a very limited objective as I said and we need to understand the feasibility of studying, just experiment is not enough just behavioural mapping is not enough. For example, recently we go by the health-related cases due to Coronavirus, it is declared by WHO as a pandemic, pandemic is usually declared when it is a special case usually occurring once in 100 years.

So, since it is a pandemic cases are expected to be very specific. So, people are now put in different detention center and they are being tested, even if the common public are not affected with the virus, but they are being tested, they are being captive. So, it has been reported in different newspapers that these people are under huge stress. There are also some cases where people are leaving the detention camp without intimating, they are trying to escape from the detention camp.

There are many reasons, but now point here is important to discuss because suppose one researcher wanted to understand the stress level of the people and those who have come across this kind of experiment. It is not generalizable, because all are not infected or all are not having the same probability they are not random. So, the particular cases must have been tested those who have already faced these stages.

So, so far is the present age at random samples are taken to understand the stage 3 of the Coronavirus. So, stage 3 may be generalizable because they are considering the mass infection case, whether a common public is affected due to another person or not. So, in that case, random may be fine, but so far as the stress is concerned for that particular person, you

have to go by the particular person's study and you need to observe carefully. So, let me proceed I have already given you a very recent example. Let me go by certain other methods of nonprobability sampling, one is called convenient sampling.

(Refer Slide Time: 25:06)

Convenient Sampling

- Attempts to obtain a sample of convenient elements.
- Often, respondents are selected because they happen to be in the right place at the right time.
- Sometimes known as grab or opportunity sampling or accidental or haphazard sampling.
- A sample is drawn on the basis of opportunity use- who's available.

swayam 9

Convenience Sample

select any members of the population who are conveniently and readily available



convenience sample

No purple figures in sample

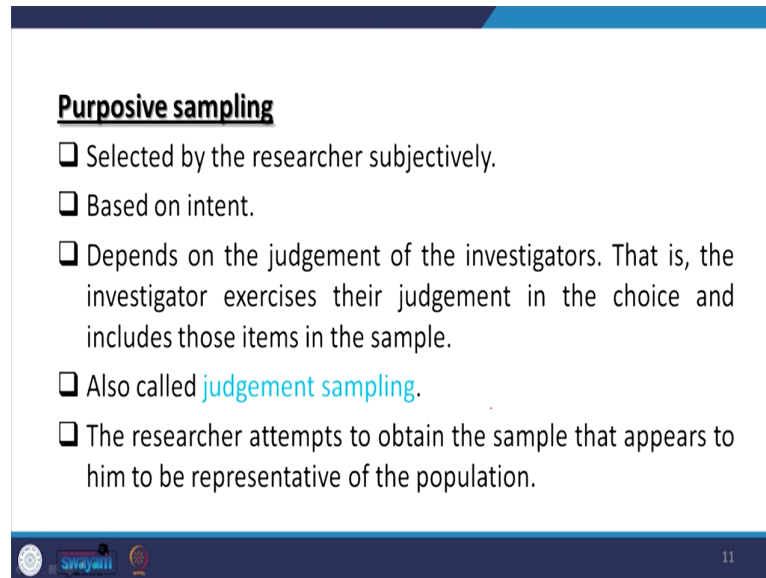
Researcher

swayam 10

So, attempts are made to obtain a sample of convenient elements often respondents are selected because they are happened to be in the right place at the right time. Sometimes known as grab or opportunity sampling or accidental or haphazard sampling. And a sample is drawn on the basis of opportunity use or we use who is available that is also called convenient sampling. Usually, you might have seen researchers consider the sample conveniently from their own proximity, that is also one form of convenient sampling.

So, here the example is very clear the picture clearly clarifies what is called convenient. The researcher wherever located maybe the nearest distance is the sampling frame, then some others may not be included.

(Refer Slide Time: 26:02)



Purposive sampling

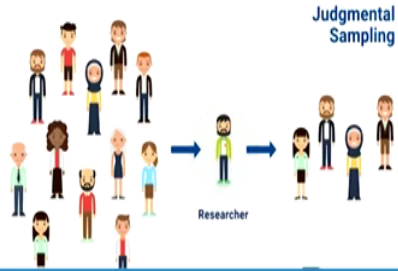
- Selected by the researcher subjectively.
- Based on intent.
- Depends on the judgement of the investigators. That is, the investigator exercises their judgement in the choice and includes those items in the sample.
- Also called **judgement sampling**.
- The researcher attempts to obtain the sample that appears to him to be representative of the population.

swajani 11

So far as purposive sampling is concerned, it is selected purposively by the researcher and is based on the intent of the research and depends on the judgment of the investigators and the investigator exercises their judgment in the choice and includes those where the judgment of the researcher perfectly fit. These are also called judgment sampling and the researcher attempts to obtain this sample that appears to him to be very representative of the population, it is related to the researcher.

(Refer Slide Time: 26:37)

For example, if the researcher wants to analyse the drinking habit of students from a class (say of 100 students), he would select 25 students (sample) who, in his opinion, are representative of the class.



12

So, here the picture is very indicative and the person researcher is located here and the researcher understood very correctly that some people are going to be representative enough in capturing the reality that is those are selected for the sample. For example, if the researcher wants to analyze the drinking habits of students from a particular class, let there are students of 100 in size, that particular researcher wanted to select 25 students a sample. So, in his opinion in the class, those who are representative in his opinion, they are selected.

(Refer Slide Time: 27:26)

Quota Sampling

The selection of the sample is made by the interviewer, who has been given quotas to fill from specified sub-groups of the population.

For example, if the researcher wants to study the enterprise performance in different type of enterprises its relationship with gender, education and socio-economic status, the researcher first identifies the subgroups (usually the characteristics and variables in the study), the researcher divides the entire enterprise into different type of enterprises (say manufacturing, trade and services), intersected with gender, socio-economic status and then he takes note of the proportions of these subgroups in the entire enterprises and then samples each subgroup accordingly.

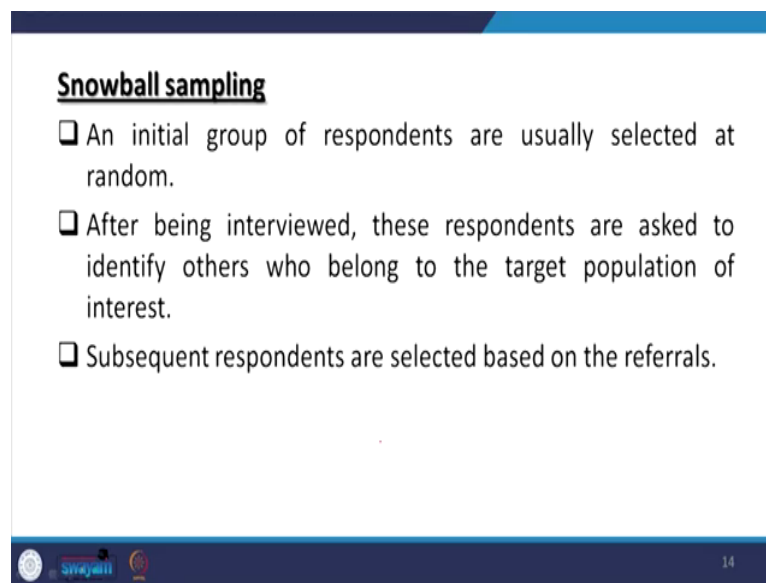
13

What about quota sampling? The quota sampling is another form of nonprobability sampling, some particular quota as specified by the interviewer is important, quota may be based on representative of the data. For example, the researcher wanted to understand socioeconomic

status, education, gender-related information, demographic, a particular feature like health indicators. So, then the researcher may divide the entire population into different groups.

So, different groups, different subgroups are divided, the subgroups may be divided on the basis of the purpose. So, the quota like socio-economic indicator, certain percentage required the researcher is going to take that off clearly, and accordingly some percentages are included.

(Refer Slide Time: 28:27)



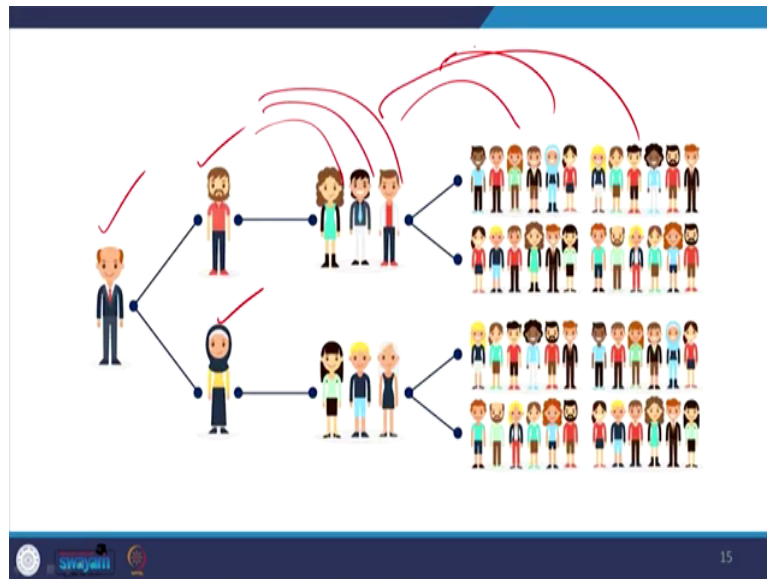
Snowball sampling

- ❑ An initial group of respondents are usually selected at random.
- ❑ After being interviewed, these respondents are asked to identify others who belong to the target population of interest.
- ❑ Subsequent respondents are selected based on the referrals.

So, what about snowball sampling? Snowball sampling is the sampling design followed as part of the nonprobability sampling where some forms of reference are considered. Since many of research or survey requires huge cost, we can get many information through our reference. Likewise, look at a case of a candidate being selected for a post. For example, in IIT pattern or even in universities faculties are recruited in the form that they call for references, letter of recommendations (LOR) forms one base of referring whether the candidate is perfectly fine or not.

So, I am testing through another reference. What we will do, what we consider? We consider another person for selection to be selecting another third person, whether the third person is good or bad, the first person is going to report to you. So similarly, if you want to increase the number of sizes, our first person might say that those persons if you select would be good for our institutions.

(Refer Slide Time: 29:44)



So, the design is clearly identifying the snowball sampling like this here, the researcher is a first-person. Referring to the three persons in our example, maybe 2 person, maybe more than that. We are suggesting the three, maybe a reviewer suggest another couple of information from your own network.

For example, even in the recent case also, some state government has already followed. In the recent case of Coronavirus, some state governments trying to tap the people from the villages, those who have some connection with international boundaries. Like, they are asking to the Panchayat and through the Anganwadis. The Anganwadi worker that you collect the households those families are having some connection with the international world and those used to travel.

So, they selectively considered those people. So, how it is possible? We selected some particular points, now those have the responsible to select more numbers. So, accordingly a better sampling design is captured. So, these are all associated aspects. We have already discussed. Accordingly, we can discuss other aspects like, once we have completed probability and non-probability sampling, we have the responsibility here also to discuss errors in the sampling design.

So, errors are very important to get a better sampling design. So, what we will do, we will discuss the sampling error and its associated factors, which can define a better sample size in our next class. Our next class will be sample size determination with understanding sampling

error. So, these we are going to discuss in the next class, I am keeping these slides for the next class. So, with this, let me stop here.

Thank you.