Handling large Scale Unit Level Data Using STATA

Professor. Pratap C. Mohanty
Department of Humanities and Social Sciences
Indian Institute of Technology, Roorkee

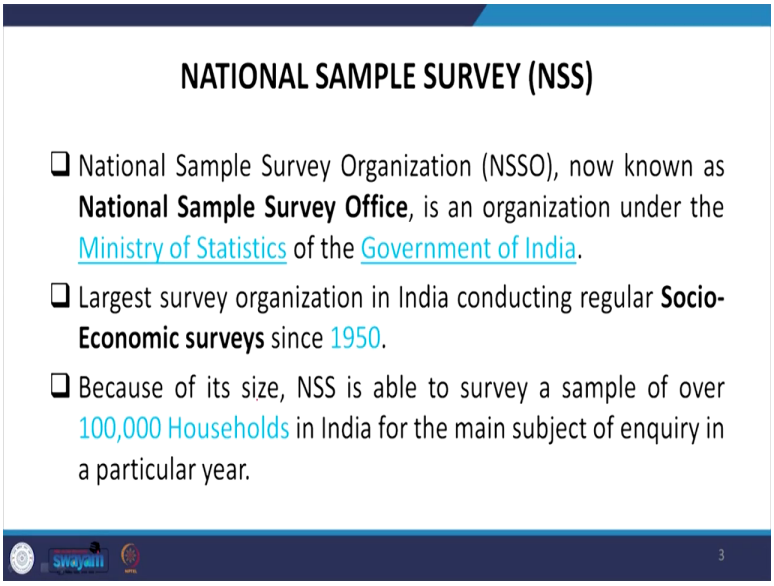
Lecture No. 03
Understanding Unit Level Data: NSSO

Welcome friends once again to the MOOC module on handling large scale unit level data with STATA. This is our week number 1 where we have been trying to understand the unit level data. So, the module is named as familiarization with unit level data. And in fact this is our third lecture. In the first two lectures we have already introduced you regarding the data and its history behind data and also we discussed the different forms of unit level data in the very introduction part.

Now we are unfolding with the core data sets. We are trying to explain you the large scale unit level data which are very famous in the world as well as for the researchers in Indian contexts. And so this is our lecture number third where we are trying to understand National sample survey data and as developed by NSSO. We have already introduced this symbol. There are also some changes due to the change of the government the organization name has been changed to office I will discuss right now.

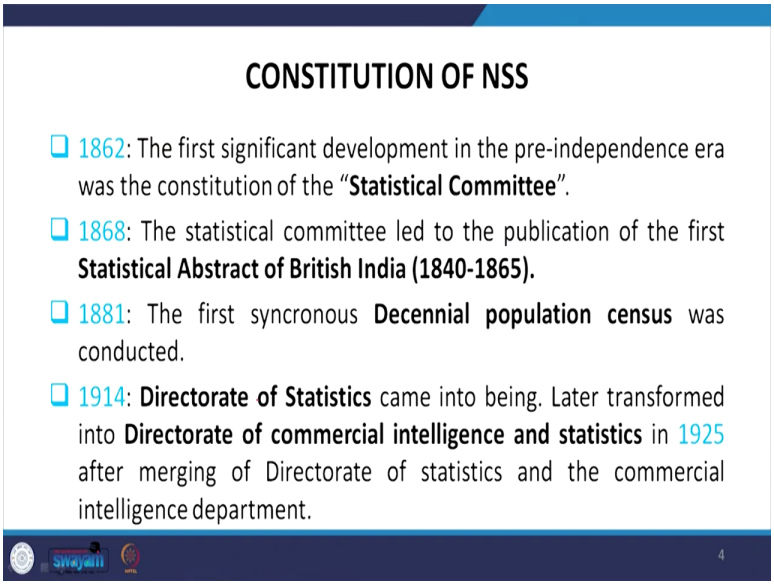
NSS as I already discussed is used to be known as NSSO but now it is renamed as National Sample Survey Office. So, this is an organization under the ministry of statistics and Government of India. This is in fact the largest survey organization as I said conducting regular socio-economic survey since 1950. So, because of its size NSS is able to actually survey the data over 1 lakh households.

(Refer Slide Time: 02:36)



In India for the main subject of enquiry in a particular year. So, imagine if the number of households are exceeding 1 lakh, this actually captures many information. So, it is scattering to the large genesis and where many interpretations could be made.

(Refer Slide Time: 2:56)



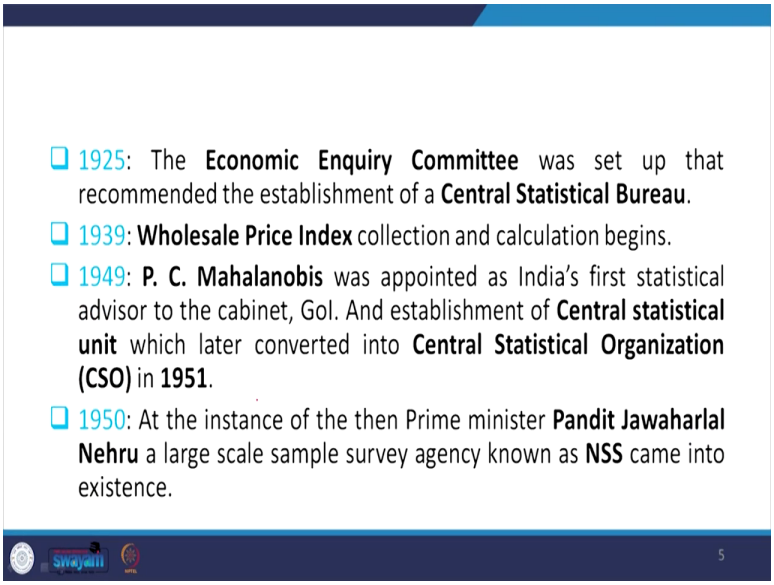
let us understand a bit on the constitution of NSS over time since it is so famous, we must be having some interest on the history behind NSS. looking back to the earlier periods especially 1862 the first significant development occurred in the pre independence era with the constitution

of statistical committee. And in 1868, that committee actually led to the publication of its first statistical abstract of the British India. So, the statistical abstract with the period cover was 1840 to 1865. In the year 1881 the first synchronous decennial population census was conducted.

So, the word synchronous matters the systematic record of the population which can be compared with other decennial population census in India. So, this timeline is very very important since then every 10 years' Indian government is conducting population census. So, counting the total record of the population till the latest population census was in the year 2011.

In 1914, the Directorate of statistics came into being. Later that was transformed into Directorate of Commercial Intelligence and Statistics in 1925 after merging of the directorate of statistics and the commercial intelligence department.

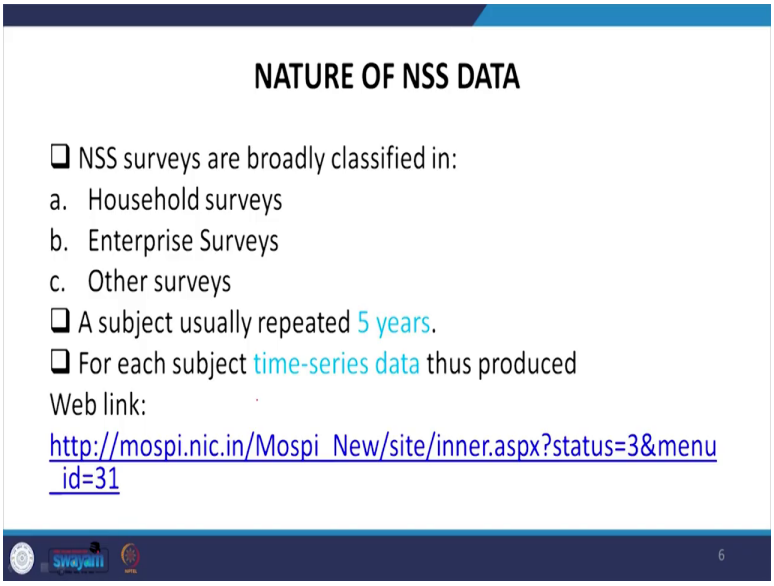
(Refer Slide Time: 04:37)



So in 1925, the Economic Enquiry Committee was set up that recommended the establishment of a Central Statistical Bureau. And especially in the year 1939, the Wholesale Price Index collection was made and the calculation for it actually began. In 1949 the famous statistician in our country none other than P. C. Mahalanobis was appointed as the first statistical advisor to that particular cabinet of Government of India and establishment of Central statistical unit (CSU) which later converted to CSO (Central Statistical Organization) in the year 1951.

CSO is very famous so far as Indian Accounting of National Income is concerned. That is in fact the most authentic source of data understanding National Income of the country. In 1950, at the instance of the then Prime Minister Pandit Jawaharlal Nehru a large scale sample survey agency was known as NSS. So, NSS actually came into existence specially from the year 1950.

(Refer Slide Time: 06:00)



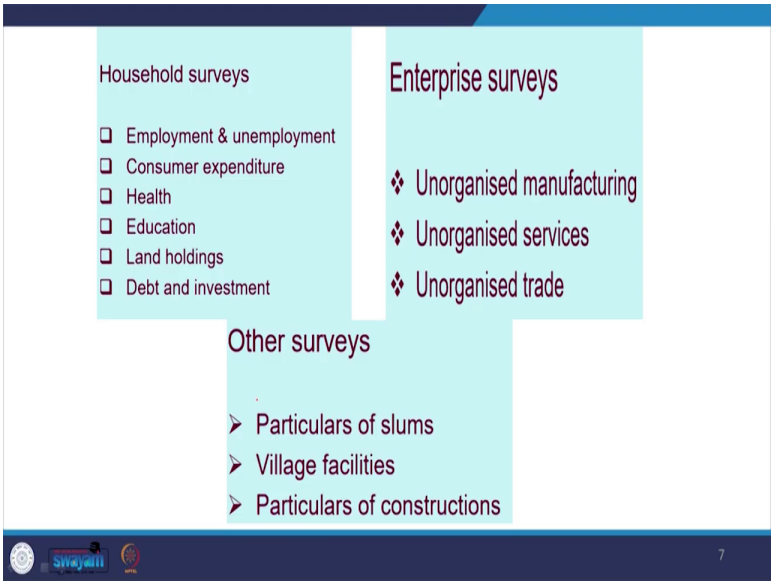
Understanding the nature of NSS data: NSS surveys are broadly classified into 3 parts. So far if you look at all rounds in the history of NSS they are broadly classified into 3. They are majorly of 3. There are others also but 3 are covering households, enterprises and other survey. A subject usually repeated in 5 years. Each subject time series data can be actually produced. Since these are actually repeated over 5 years.

So, time series we are referring to the occurrence of certain particular issue over time over different periods. In some cases, it is 5 years and in some cases it is 10 years so we will discuss that. And this is the source the web link is given here the web link is mospi dot nic dot in, you can click on this website you will be directed towards the exact source of data where you can understand further details.

Let me also mention in between that in every week we will add question papers for your practice. We will also give you some interaction module with STATA where it will be very useful for your

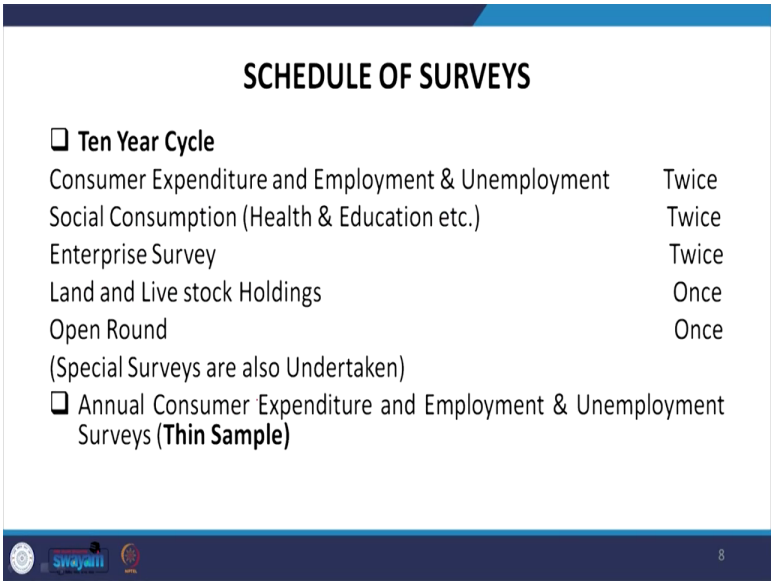
work. But at this moment let me clarify the different forms of data, you must be quite familiar with those data in order to start with using STATA.

(Refer Slide Time: 07:39)



As I just said there are broadly 3 types: household, enterprise and other surveys. Household survey covers employment and unemployment, consumer expenditure, health related information, education, land holdings, debt and investment. Whereas the enterprise survey over different period of time in a regular basis that covers organized unorganized manufacturing, unorganized services, unorganized trade. And in other survey it majorly covers slums, village facilities and particulars related to constructions.

(Refer Slide Time: 08:20)



If you check the schedules of the survey, schedules which generally refers to the questionnaire. Schedules of the survey, we find there are different time period of a survey there are frequency of its time period also like there are some 10-year cycle. If you count by 10-year cycle, in the 10-year cycle there are some double occurrence, there are some single occurrence also. So, what I wanted to mention here that the 10-year cycle, consumer expenditure or employment-unemployment surveys were conducted in India.

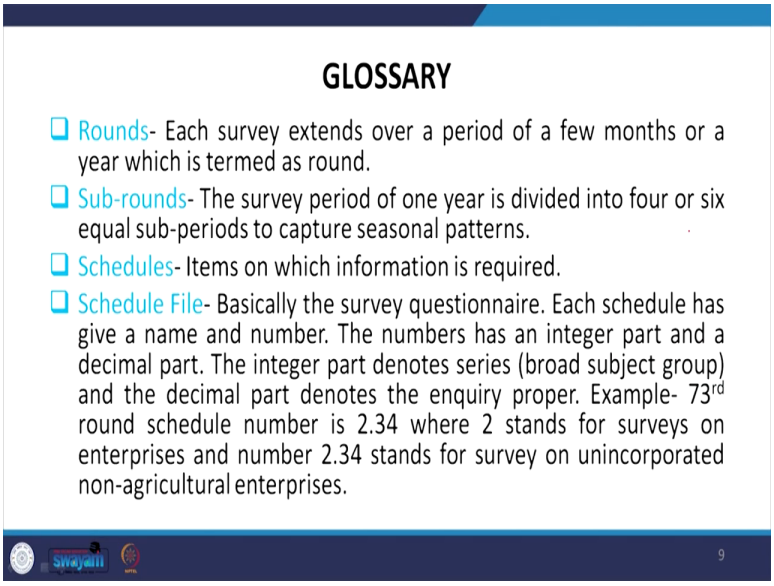
In a 10-year cycle, it occurs twice. So, 2 times survey done in Indian context in every 10-year cycle. That means roughly of every 5 years we will have one round for consumer expenditure related information and employment unemployment related information. Similarly, for social consumption which includes health and education also twice in a 10 year cycle. So, these are very important for you to note because you may have some quiz related questions on this occurrence and twice or once or thrice may make you little confuse also. So there is no question of thrice in 10-year cycle.

Also mark the line in between that we are referring to a 10-year cycle if a twice means 5 years each for a round. if you look at enterprise survey it is also twice, in the context of land and livestock holdings it is not twice it is once in 10 years. And there are also some open rounds, some special surveys are also conducted. So special survey we will discuss actually in our due course of time that the annual consumer expenditure round similarly some annual rounds for

employment and unemployment surveys these are generally called thin rounds. Thin rounds usually cover less sample.

And so there is no such cycle mentioned when we talk about open rounds. But still in some specific open rounds there are cycles. Since it is one year, so one-year cycle will be considered. But actually it is not repeated every time. So, there is no such clear cycle defined in a thick sample we have clarity related to sample related to cycle.

(Refer Slide Time: 10:55)



As a further glossary to the explanation made let us clarify what do you mean by rounds, what do you mean by sub rounds, schedules and schedule file. Without these it is very difficult to understand NSS or even any other data. In our next class, it will be on IHDS data India Human Development Data. And I am deliberately mentioning this name here to have your interest for data, unit level data those who want to work for development related issues purely Indian based development issues as a comparison to the world or especially for the developing countries context, these surveys are very very important.

How to explore, mine those data sets. These are not that easy unless you read between the lines. So, rounds, sub rounds, schedule and schedule file are important. Let me clarify each of the concept. So, round actually we meant here or else for the NSS is that each survey extends over a

period of a few months or a year which is termed as a round. If the survey started if it continues for a few months or a year to complete that surveys simply called as round.

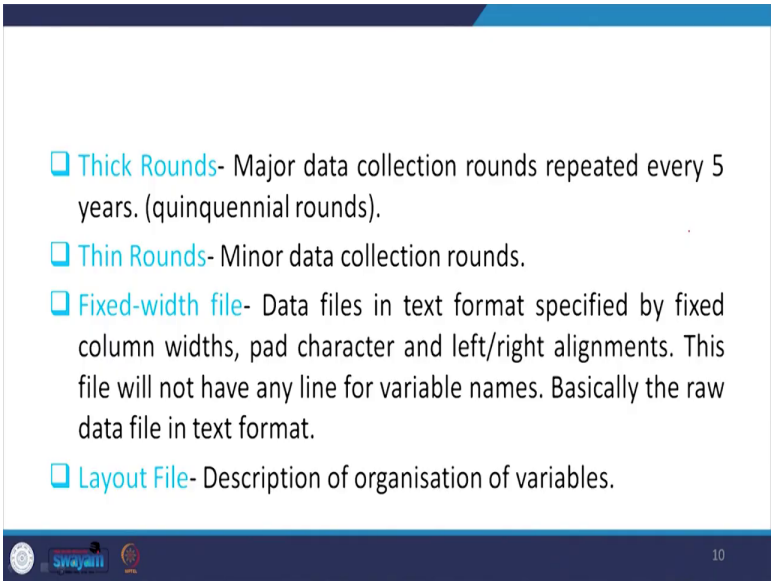
Whereas, if you check the sub rounds the survey period of one year actually classified or divided into different parts maybe 4 or 6 equal sub periods to capture the seasonal patterns. Likewise, if you check the PLFS recent data, periodic labour force data, there are number of debates you might have checked in the newspaper related to unemployment situation in India even when the database was not released.

I am just mentioning because of its importance in the news. So, periodic labour force survey finally release last year but referring to period 2017-18 which has included sub rounds like the survey are actually divided into 4 quarters. So, 4 quarters specially for the urban areas. They considered 4 quarters for rural as well as for urban but in rural the persons are different whereas in the urban the persons are same, its repeated.

And whereas in case of schedules, items on which information is required is very important and content in the schedule number. So, I will explain what do you mean by schedule. Sometimes it may 10.2. what do you mean by 10.2, can be clarified? And so last one is called schedule file. Schedule file is basically the survey questionnaire we are referring to. Each schedule has given a name and number. The number has an integer part and a decimal part. So, this is mentioned and the integer part denotes series, broad subject group is mentioned and the decimal part denotes the enquiry properly or the proper enquiry of that particular part.

Specially, if you referred to the seventy third round of the NSS which is on unorganized enterprises, a schedule number so we start with the schedule. We are actually referring to the exact schedule file schedule number if it is of 2.34 where 2 stands for surveys on enterprises whereas the 0.34 stands for survey on the particular unincorporated non-agricultural enterprises.

(Refer Slide Time: 14:44)

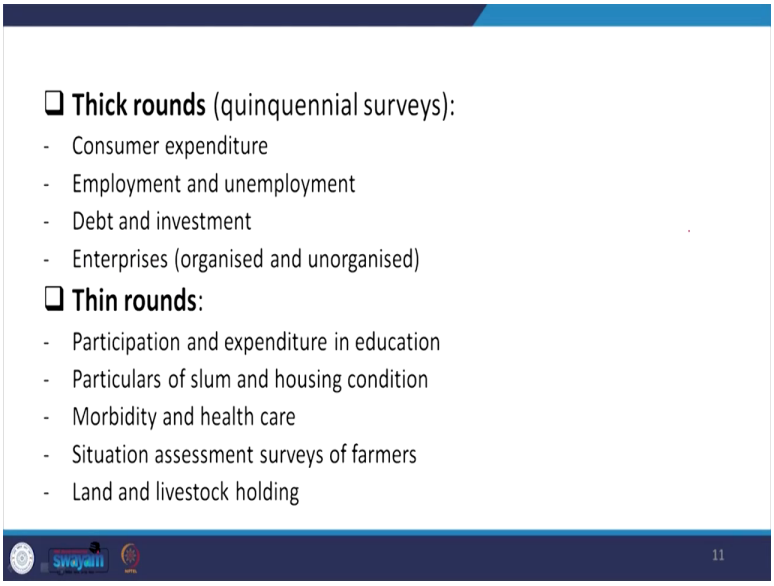


Let us clarify what do you mean by thick round and thin rounds. Thick rounds are the major data collection rounds. As I mentioned already repeated in every 5 years. Usually it is called quinquennial rounds. Whereas in case of thin rounds minor data collection rounds. Sometimes, they were also called as annual rounds even a special rounds or we call as thin rounds because coverage is either less or the coverage in terms of sample number might be less or the number of issues covered might be also limited as compared to the thick round.

So, another clarification is very very important in order to understand the data which are given in notepad file usually in NSS to start with this is called fixed width file. What do mean by fixed width file? The data files which are generally in text format specified by fixed column width, the total column width we will show it in our one example. It includes pad characters also and it includes left or right alignment also. So, this file will not have any line for variable names there should not be any specific line these are mentioned continuously. including all those aspect, column width, pad characters and as well as alignments basically the raw data file actually is in text format so far as fixed width file is concerned.

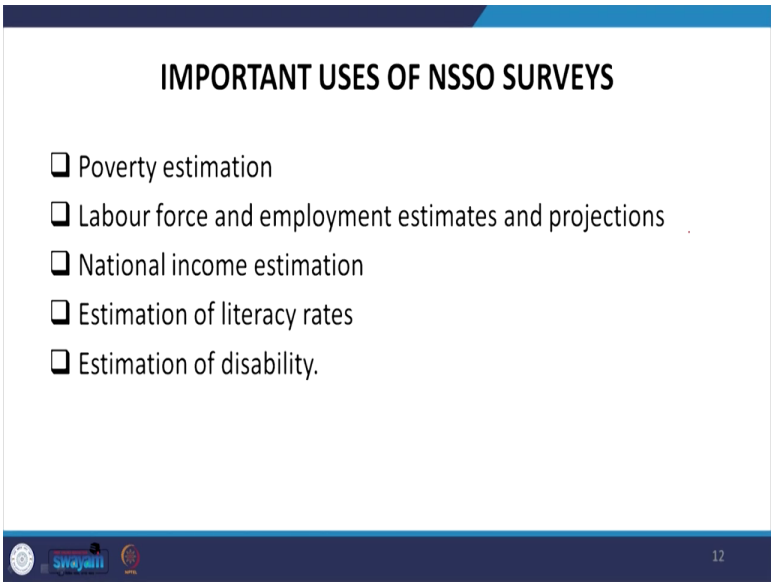
Another file is called layout file. I am taking the name of it, we will show it in the exact dataset. So, layout file gives the description of organization of the variables how variables are actually organized.

(Refer Slide Time: 16:43)



in the thick round as I mentioned it is called quinquennial survey like consumer expenditure, employment unemployment, debt and investment, enterprises. Whereas in case of thin round it is on expenditure in education, particular of slum and housing condition, morbidity and health care, situation assessment of farmers, land and livestock holding.

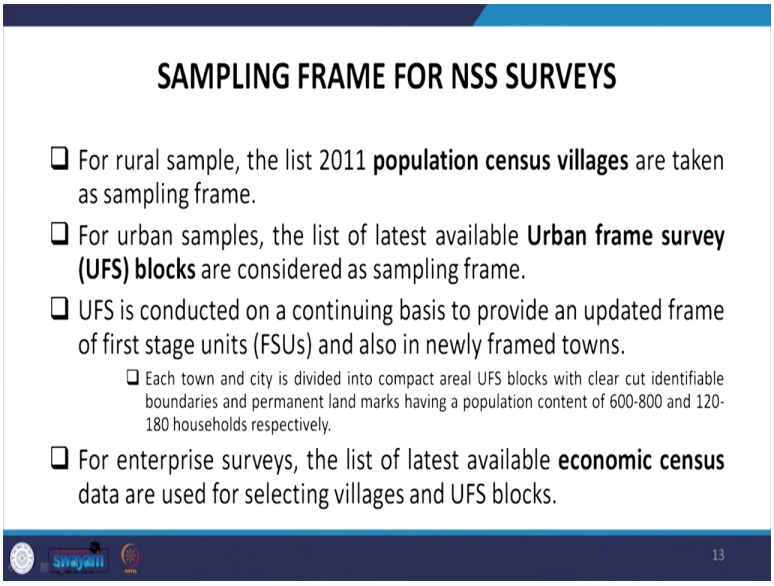
(Refer Slide Time: 17:10)



An important use of NSSO survey like where we can use this survey for our analysis. we can use for poverty estimation, we can use for understanding or estimating employment or in its

projection also, because we have a time component as well over so many years for this survey so there is possibility of projection for another 5 years or 10 years ahead. National income estimation, estimation of literacy rates, estimation of disability as well why we do it and why we do need it because of some kind of policy framework for future generation or future context.

(Refer Slide Time: 17:51)



What are the sampling frame for NSS survey? What kind of sampling frame used? This is very important if you look at, go through the report of an NSS probably it is too difficult to read between the line because there are so many information given. So, we tried our best to simplify it by collating various information from various sources.

For rural samples, start with the sampling frame of NSS for the rural sample, the list 2011 population census villages were taken as sampling frame. So, we take the base as population census as the sampling frame. Whereas, for the urban samples we consider urban frame survey that is UFS blocks. I have already discussed this in our last lecture UFS how this is considered urban frame survey UFS blocks are considered as a sample frame.

UFS is actually conducted on a continuing basis to provide an updated frame of first stage units. I think in the last lecture we discussed first stage, second stage and third stage so broader categories to smaller category again to the further smaller category usually at the third stage we

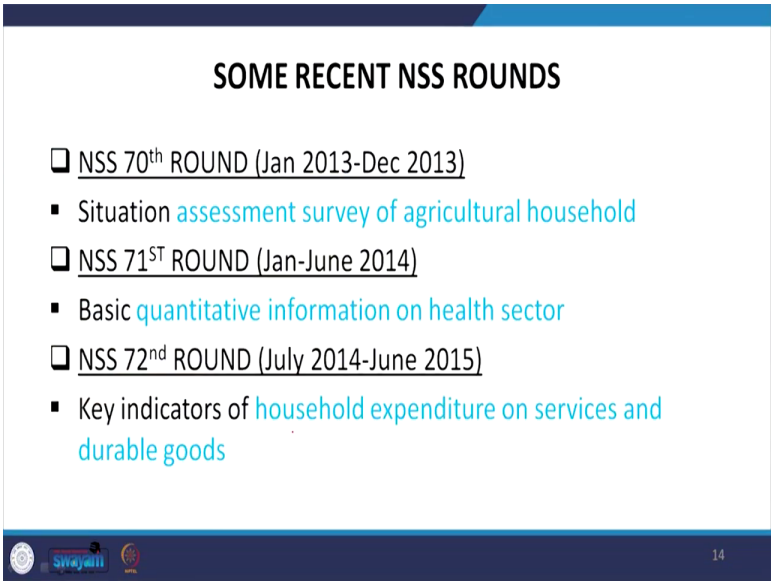
discussed the unit that is individual. Whereas in the second stage it is the household or the family.

So accordingly we define, our first stage will be the rural areas. And UFS is conducted on a continuing basis to provide an updated frame of the first stage unit and also in newly framed towns. So, let me have a clarification on it, each town and city is divided into compact areal UFS blocks by area but how these areas are defined and how these are actually classified it depends on between town and city there are the population and there are the boundaries, are defined by the population content of 600 to 800 if it is a town and 120 to 180 households if it is a city.

If these are based on these benchmark or limit city or towns are defined and the FSUs are actually defined. Basically UFS blocks are defined based on these limits. So, different different blocks based on 600 to 800 population 1 block or it 120 to 180 if it is a city of household as another block.

Similarly, for enterprise survey what could have been included? It is of course not the household it should be the unit as the enterprise. But the enterprise survey is actually based on the census. Likewise, this is based on the census so enterprise is also based on economic census not population census. Economic census considers all possible enterprise in our country. So far we have the latest economic census is called sixth economic census into 2013 and 14. So, data used for selecting villages and this is used for selecting our sampling base for villages as well as for UFS blocks.

(Refer Slide Time: 21:16)



These are some of the recent NSS rounds conducted in India. We are not discussing the previous round you can go through that website and track all previous rounds. From seventieth round that is 2013 January to December round, it covers situation assessment survey of agricultural household. Whereas in seventy first round, quantitative information related to health sector was given. In seventy second round it was on household expenditure on services and durable goods.

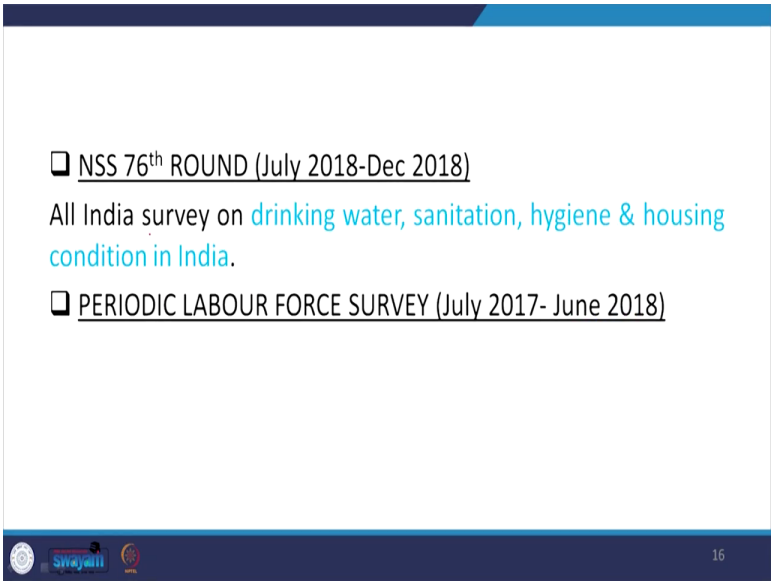
(Refer Slide Time: 21:52)

Seventy third that is 2015-16 July to June the enterprises which were actually covered. The title of that survey was unincorporated non-agricultural enterprises excluding construction. So, while

we consider this as our base for analysis as the data for analysis we must be very careful which data it is not considering. What are the limitation of the data? Even while communicating research papers we need to mention that this is the limitation of our data. Data does not cover all the enterprises it excludes some of the other enterprises. Even the agriculture is not covered.

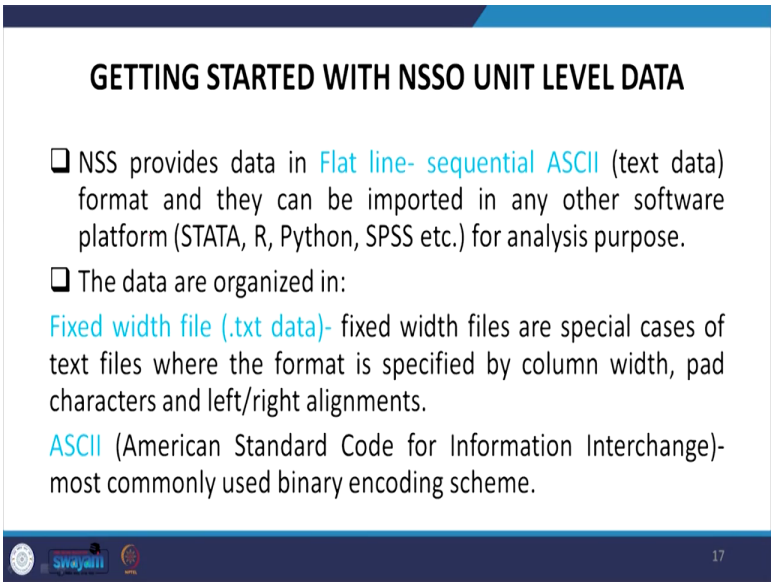
Looking at the seventy fourth round of NSS that was conducted in July to June 2016-17 it was on service sector enterprises so this is in fact the latest enterprise information available in Indian context. Seventy fifth is on social consumption on education in India.

(Refer Slide Time: 22:57)



Seventy sixth is on drinking water, sanitation, hygiene and housing condition. The most famous and debated data that is periodic labour force the title of the survey is not as like the earlier rounds NSS seventy sixth or seventh they have actually taken the name of it is called periodic labour force survey PLFS July to June 2017-18.

(Refer Slide Time: 23:26)



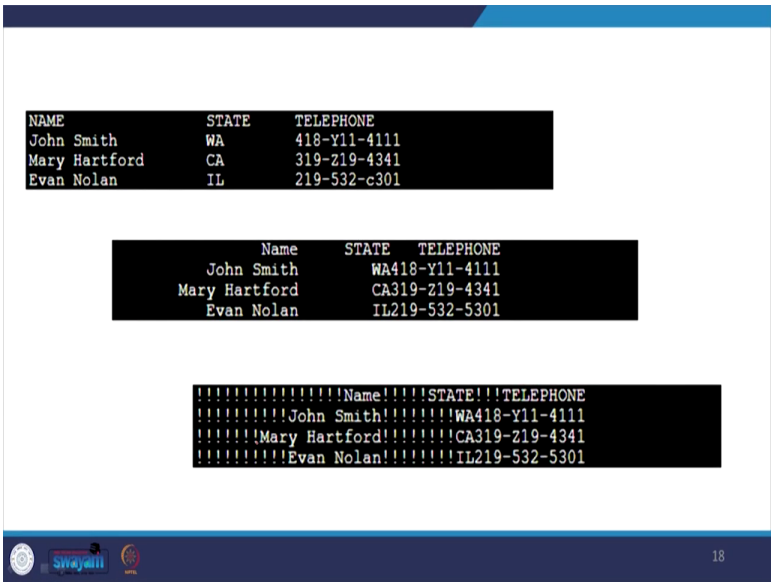
Let us start with using NSS data a bit. How we can use it in our STATA module. In order to use it we need to go through the exact file so the file is actually as I mentioned originally given in

ASCII format those are also called text data and they can be imported to any other software like STATA, R, PYTHON, SPSS etc. For analysis purpose. The data are organized in fixed width format in NSS data notepad file or ASCII format is maintained with fixed width format.

So, fixed width as I already said that total column space are same and it contains pad characters then the notations for the pad characters must be same throughout and here also alignment given in the notepad file or the data must have been also same either it will be left aligned data or right aligned throughout.

We will show it in our example so fixed file in the next slide I have the picture for it. So, I have already said that it is specified by column width, pad characters, left and right alignments. So, ASCII is basically American Standard code for Information Interchange most commonly used with binary encoding scheme.

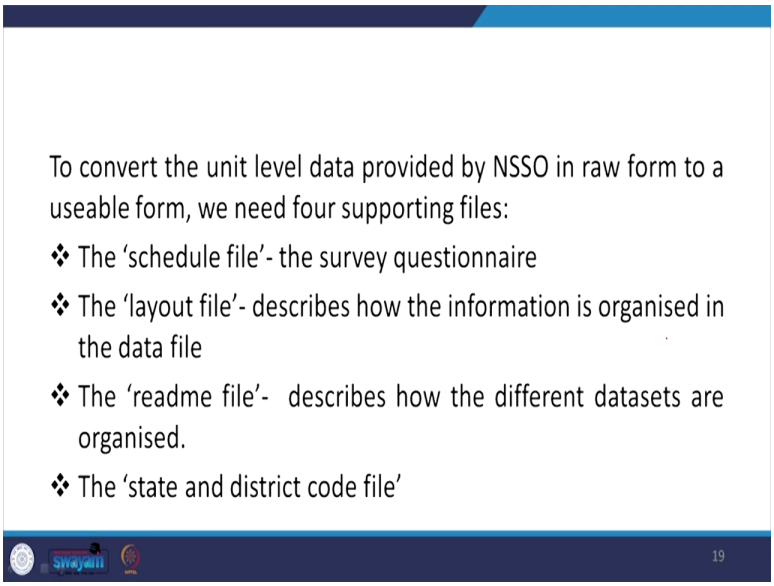
(Refer Slide Time: 25:03)



So, these are the examples as I said. Look at the data extracted from a notepad file. After extraction we get certain figures. This is name, state name, telephone number for example if it is there, these are actually given in left aligned format. look at the second one is in right aligned format, the whole dataset is actually interpreted in right aligned format. if you try to check with the pad character, here actually given with this symbol. This symbol every time. Either we go by space in between space would be instead of this symbol space may be considered.

Or any other symbol is considered but that should be similar throughout the dataset. What do you mean by this? This is similar to right aligned dataset but if blanks are there, these blanks are filled up with a pad character with this symbol and this symbol must be similar everywhere. Why we are taking it? Because this is going to be very useful for our interpretation and explanation using any software, maybe STATA. So, you please take a note on pad character the meaning of pad character.

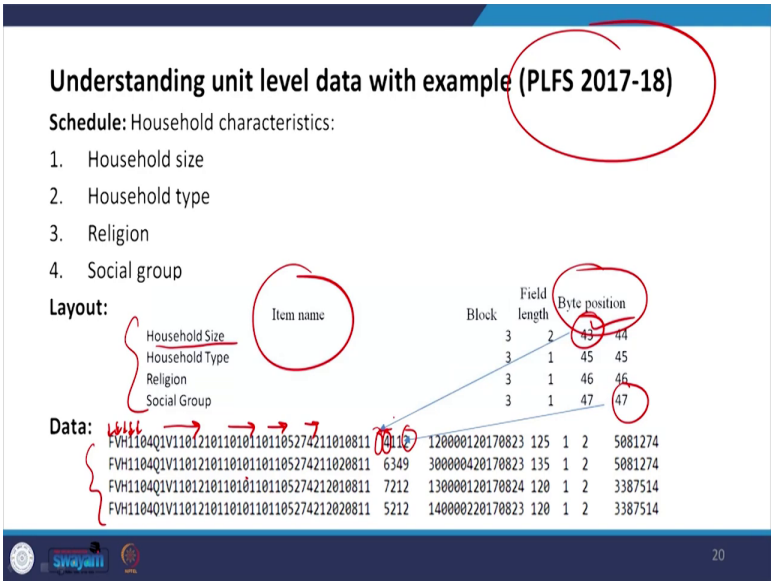
(Refer Slide Time: 26:33)



To convert the unit level data provided by NSSO in raw form to a useable form, we need 4 important files. As I already discussed in our last lecture that it requires schedule file, layout file, readme file and state or district code. So, schedule file contains survey questionnaire. Layout file contains information regarding the data, how the data is organized and the Readme file describes how the different datasets are organized.

Whereas, the state and district code file gives the code of the state and district taken. Without this code it is not possible to interpret the data. Otherwise data remains only as data.

(Refer Slide Time: 27:26)



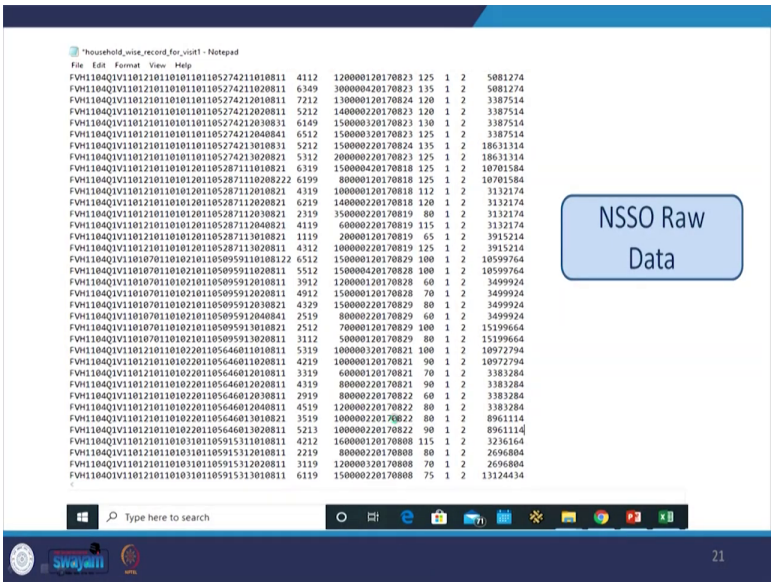
We are taking the snapshot from the original data from some of the extracted information from the data and we are presenting it for your clarity. And this is with respect to the latest PLFS (periodic labour force) dataset of 2017-18. It contains schedule of household characteristic, household size, household type, religion and social group and there are other information also.

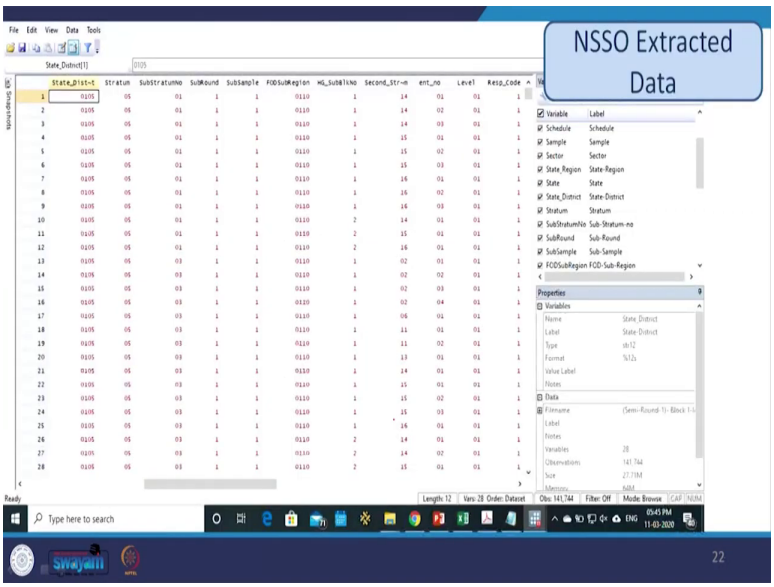
So, let me understand a bit on how these are given in our notepad file. Notepad file contains with fixed width as I already mentioned with pad characters and alignment. So, here it is not visible the pad character and the alignment is not visible once we extract we will get to know whether that is in left or right, but after extraction it is clearly understood but at this moment I just wanted to understand how we should read the byte position. And this byte position information is given in the layout file. So, layout file is as I just said usually given in excel format in NSS and sometimes in PDF format also.

But I have already shown you those excel layout file in our last lecture go through that layout file the samples of that layout file and try to look at. Even this is the sample here given as the layout file, it contains this information. From the layout file we should know this and byte position. Byte position is very very important and also we need to understand in which block those are considered.

Now suppose I wanted to read what is the total byte position here, it is from 43 to 47. 43 if I start counting 1, 2, 3, 4, 5 like this if we keep on counting the position here actually the position stands at 43 and this is our 47 position. So, in a STATA command we are supposed to give the exact byte position for correct extraction. Suppose we wanted to extract with the household size we have to give 43 and 44 these two are quite important for our better extraction. So, let me proceed we will explain in our exact lecture.

(Refer Slide Time: 30:02)





So, this is the raw data format we already said number of times. The extracted format will look like this and from here you can find out whether these are left aligned or right aligned after extraction it seems it is clearly right aligned one, not left aligned and pad characters as I said these are actually in blank space everywhere blank space is somewhere mentioned. But it is not visible if you go by particular character you can also assign a character but probably it is not useful at this moment. So, we will discuss in the exact lecture.

So, these are all in sum we have given you many handy information, very very handy information in understanding NSS data and with the right command, right understanding of the format the files you can able to extract and enjoy the data. So, thank you learners.