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. 02
Introduction to Unit Level Data

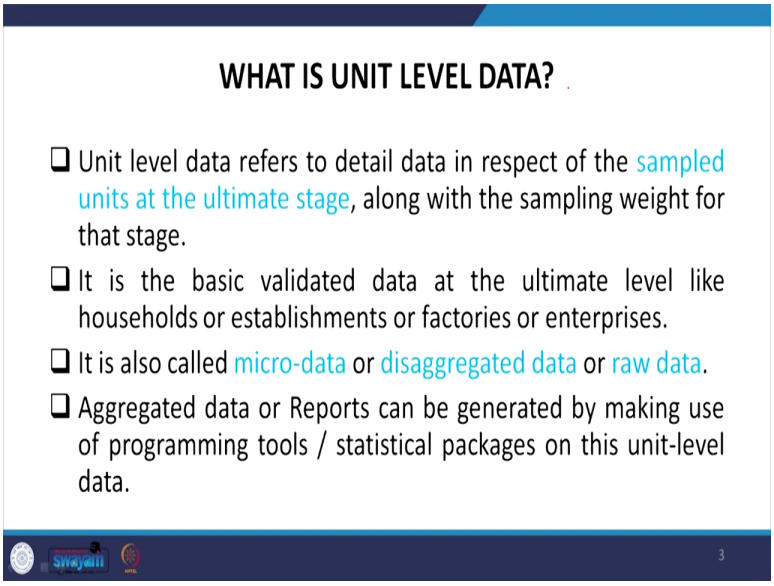
Welcome friends once again to the NPTEL MOOC module on Handling Large Scale Unit Level Data using STATA. This is our lecture number 2. Myself Doctor Pratap Mohanty, a faculty member in the Department of Humanities and Social Sciences, IIT Roorkee.

(Refer Slide Time: 00:46)

I have already handled this kind of module in different platforms and so I am trying my best to make it more interesting so far as the current gaps of different platforms handling unit level data is concerned. My attempts here is largely to help the stakeholders like PhD students and professionals and policy makers, and I hope this is going to be a direct help to the students as well as professionals.

The second lecture largely focuses on the unit level data. The first one was on understanding data. Here we are specifying on unit level data. You can see from this picture that which sections we are actually dealing throughout our lecture. As I have already mentioned in the last lecture that this is on unit level data. We will handle NSS, IHDS, NFHS, extraction, append, case of merging, pooling, panel, various aspects are there.

(Refer Slide Time: 02:00)



How to understand unit level data and what is unit level data? Unit level data refers to the detailed data in respect of the sampled units at its ultimate stage and along with their weights for that particular stage. The sample units, how they have been defined in different formats; we will be at this moment emphasizing on NSS format because this is considered to be the standard dataset in the international level on unit level information. and how sample units are defined, we will actually talk in our next slides.

Now let me talk about unit level data. It is the basic validated data at the ultimate level like households or establishment or enterprise or factories. So, usually in the unit level data like NSS, we have the unit household. Whereas, if the survey is directed purely for the understanding of enterprise and their relationship, their characteristics the unit is actually enterprise or factories. This is also called micro data or disaggregated data, and as I already mentioned this is called raw data.

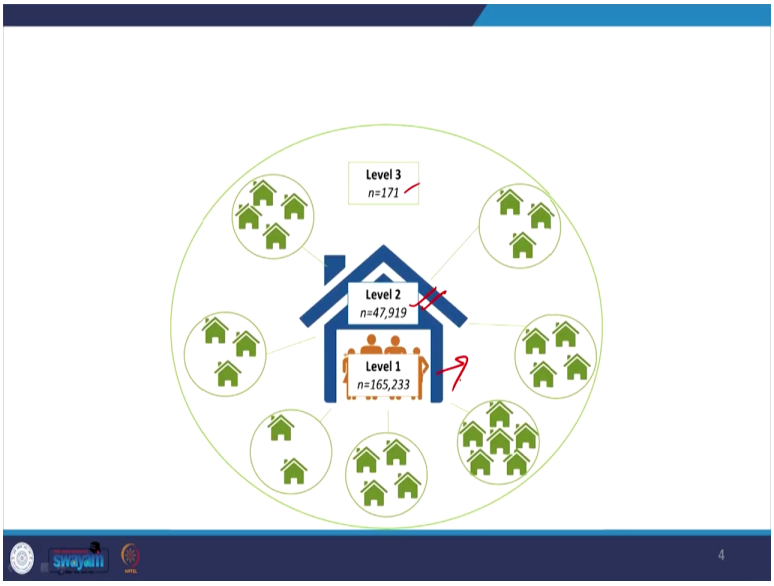
Raw data, if it is processed it is not just called a raw data. There are large number of confusion, various forms of confusion by the researcher is: whether I interpret this data as primary data or secondary data. This is a clear confusion among the researcher. Let me make it very clear before you that, what do we mean by primary? Primary, usually when the researcher directly observe and get the data.

If I am working in a project and my team is working for that project, my team, one of the member has visited the field and collected information. For that person it is direct but for the team entirely it may not be direct even if it is the first source but it is not direct. what it matters whether it is raw or not. If it is in raw form various interpretations can be derived out of it. Various possibilities are derived out of it.

So usually, for me NSS, for many researchers NSS data or these type of large scale unit data is called secondary but so far as its length and breadth is concerned and the kind of information it has covered, is actually not less than a primary data. The NSS data or this type of large scale data is also called primary data in official sense because the primary information has not been actually distorted. The same information in the raw form will be passed on to the researcher.

So, the researcher has the privilege to process it accordingly, recode it, mine it and interpret it with different techniques. So, nowhere the data has been actually distorted, so for me, if I handle the NSS data, is not that indirect, okay. It may also be called direct dataset. So that is the beauty of working with NSS or IHDS or NFHS data. Aggregated data or reports can be generated by making use of programming tools as I just mentioned or statistical packages on this unit level data.

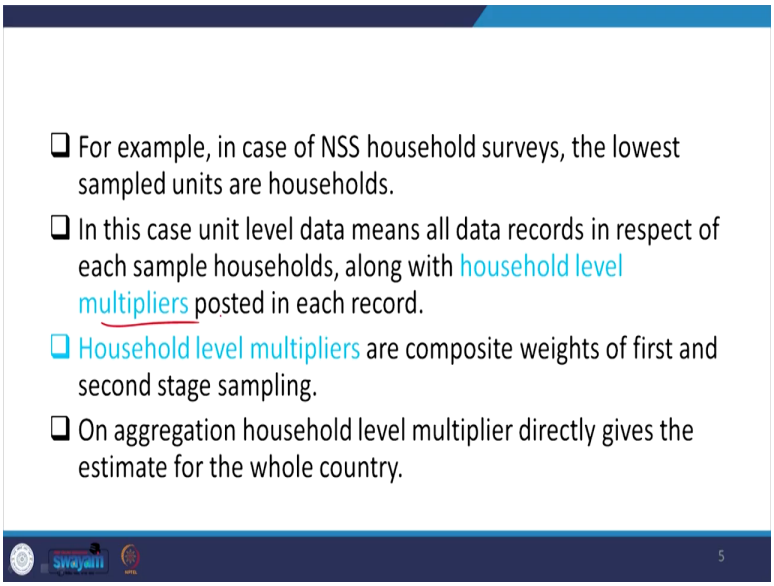
(Refer Slide Time: 06:02)



In this picture, there are different layers, hierarchical orders followed to understand the unit level. What do you mean by unit? unit here mentioned as n equal to 165,233 Unit, if it is household then certainly the number will be higher. Above to that level is, may be in a particular state, maybe a block maybe in any form and above to that, then above to that, it is a state. There are various layers defined in NSS. So, various forms are actually defined.

So, accordingly the end level actually reduced. Initially it is 171. This is higher, then higher so unit actually, the number of units are actually much higher. So, the final unit is actually much higher.

(Refer Slide Time: 7:11)

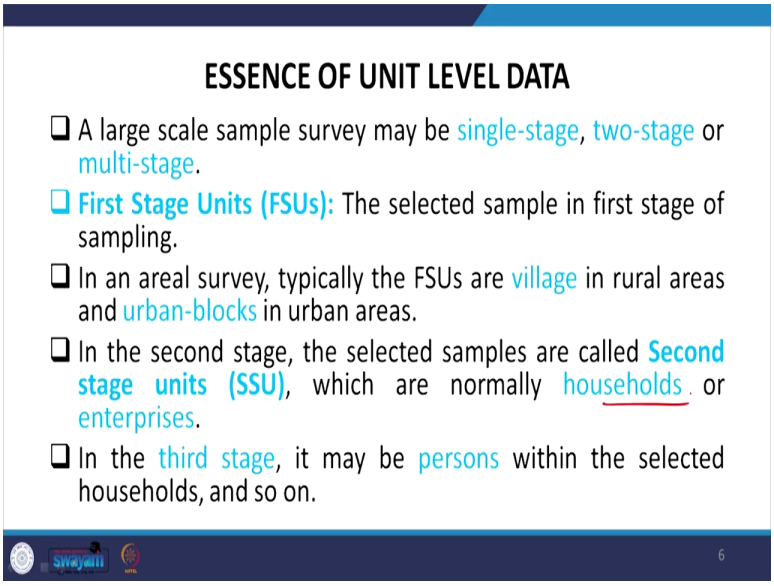


And so, it is defined as level 1. And in case of NSS as I just mentioned relating to or understanding the household information, these are called household based serveys. The lowest sampled units are as I mentioned called households. In this case, unit level data means all data records in respect of each sampled households along with household level multipliers posted in each record. This is very important to note. We will show you in our next slide, next to next slide on what do we mean by multiplier or layout files, how these are defined.

So far as household level multipliers are concerned, these are defined as composite weights of first and second stage sampling. So, obviously I will have a question on what do you mean by first stage and second stage sampling and accordingly how the multiplier is defined? On aggregate household level multipliers directly gives the estimate for the whole country whereas,

on the disaggregate level you will stick to a particular unit. So, what is the essence of this unit level data?

(Refer Slide Time: 08:22)



The large scale sample survey maybe of single stage, maybe of two stage, or of multi-stage. what do you mean by FSUs? If you look at the reports of NSS it clearly mentions these FSUs (first stage units). What do you mean by first stage? These are the selected sample in the first stage of the sampling process. So, usually FSUs are villages in rural areas and urban blocks in urban areas. Then how they go for it?

In villages, the standard format of referring to the FSUs are through Census to define the FSUs. Census is the standard reference. Whereas from the Census, we have two categories. One is called village level census, another is called, census enumeration blocks. Census enumeration blocks is usually referred to urban blocks. Urban blocks are actually filtered and clubbed differently in NSS approach.

The second stage of the selected samples are called second stage unit samples or also called second stage units, SSU, the FSU and SSU. SSU is called second one which are normally the households or enterprise we have already discussed. So if you again disaggregate to the third stage then certainly we will refer to the particular person within the household. A number of persons are actually covered.

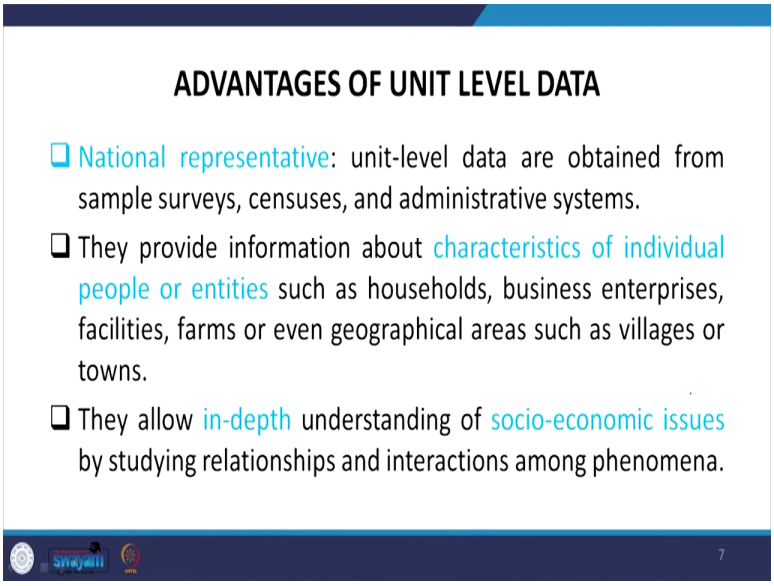
You will be surprised to know that those who are not acquainted with NSS and have interest to work, NSS covers so many persons. In the latest round 24 lakhs information are there which are actually sampled. You imagine what quality of information you can derived out of it. So, over the third stage as I just referred in the first bullet point called multi-stage sampling method, this is not exactly multi-stage sampling but the stage at which sampling units have been concerned and so we are referring as multi-stage.

Here our third stage is the person. the person exceeds 24 lakhs which is in fact a very good large number and interpretation will be very interesting and these are in our module to be explained in our successive lectures. What kind of advantages are there in dealing with the unit level data? We have national representative datasets and NSS are also defined as national representative due to certain reasons.

Suppose I do survey as a researcher. I simply collect based on my budget, I simply collect certain information. Those information may not be actually representative enough. Even if I add weights to it, weights then maximum I can go for 500, 1000 or even 10000, 20000 in particular area based on the context. That does not mean it is representative to the wider variety or a different spaces, because out of the total population in the country if I simply go by randomized approach the number is actually very miniscule, so cannot be actually represented. But if it is a particular context that 500 units might be also representative enough to that particular area or to the context.

It depends upon what exactly we are searching and how we are searching. Such examples we will certainly cite while we will deal with the statistical inferences. So, the unit level we are referring here is going to be very representative enough because it follows multi-stage approach, multi-stage sampling process. It goes by referring to the census villages or census blocks. It also follows a systematic and multi-stage sampling procedure with representative weights. Weights are given. We will discuss. We have a dedicated module on weights, and, so we will identify those aspects very clearly.

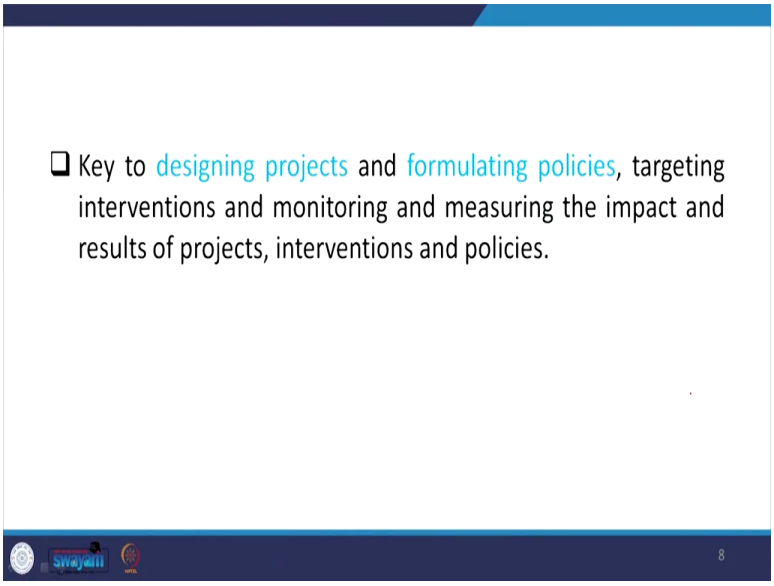
(Refer Slide Time: 13:06)



In addition to do that to understand the advantages, we also have certain characteristics of individual people or entities. So, the characteristics such as household, business, enterprises, facilities, farms or even geographical areas such as villages or towns. why this is important? As I referred in the previous lecture that some companies or maybe some marketing units or agencies wanted to deal with that particular area. So, they may collect certain features from those. Surveyed information and since survey is relatively less expensive and if survey is multiplied with different variables and information, survey gives larger qualitative information than that of census

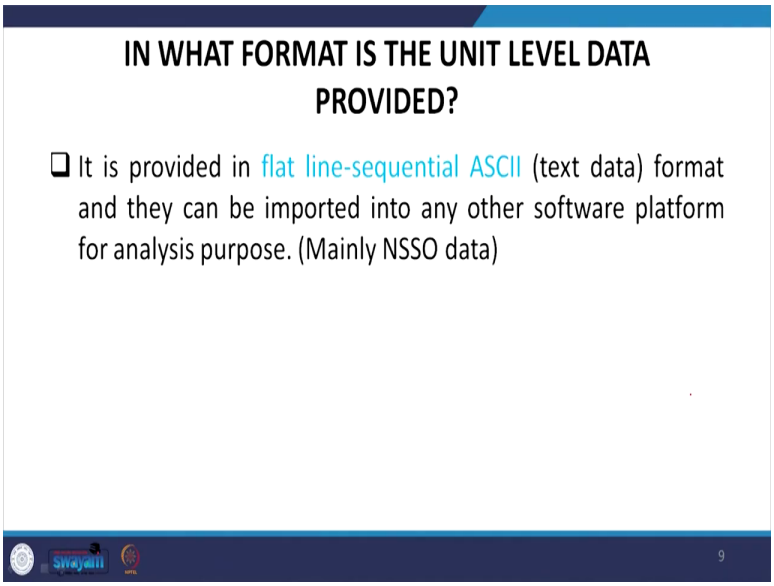
So, the third most important aspect is, it is in-depth. As I just say it is actually in-depth because of its coverage. It covers socio-economic issues by studying their relationship interaction among various phenomena.

(Refer Slide Time: 14:18)



The key to designing projects and formulating policies and targeting interventions, monitoring, measuring the impact, results of projects, interventions and policies etc. are very important so far as NSS or unit level data is concerned. We are emphasizing NSS because it is very systematically presented. Then, what format is the unit level data provided?

(Refer Slide Time: 14:41)



What are the format? It is very interesting to note because to read individually without spending sufficient time is very difficult. So , this module is actually targeted to simplify to get a

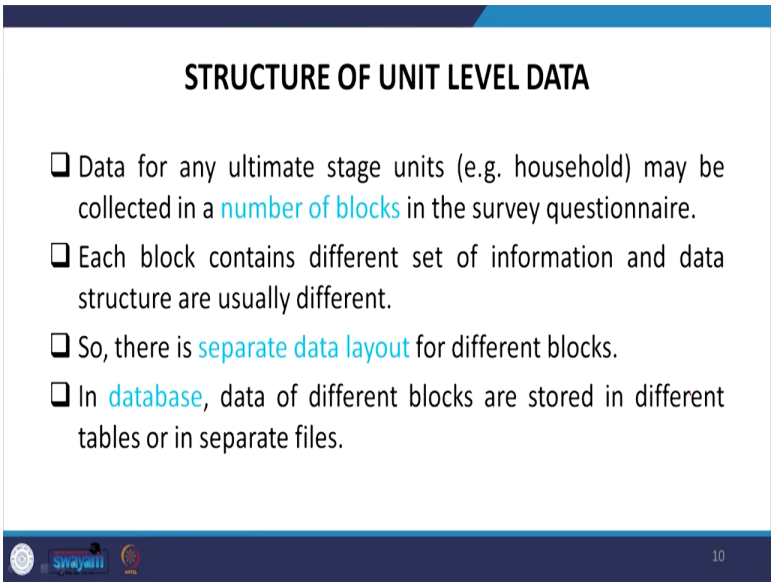
structured idea of looking at unit level data. So, it is generally interpreted or provided in flat line sequential ASCII format. ASCII format are also generally discussed as text data or text format.

So, in the Notepad file we have already shown in the last lecture that informations are in byte space, character space are given. So, this is also called flat line sequential ASCII format. They can be imported to any other software platform for analysis purpose, mainly NSS data, imported very systematically.

Let us understand the structure of those unit level data through the original document of Government of India and those large scale data. So, data for any ultimate stage units may be collected in a number of blocks in the survey questionnaire. For example, in the latest periodic labor force data we have different blocks of information. Some blocks might identify the demographic behavior of the particular labor being surveyed. Another block may identify the nature of work they deal with.

Some other block may deal with their skill related information. Some other block, the last block of that particular round, the latest round deals with the daily affairs, per day, even the per day and their remuneration paid, the per day within the weekly status is very clearly identified.

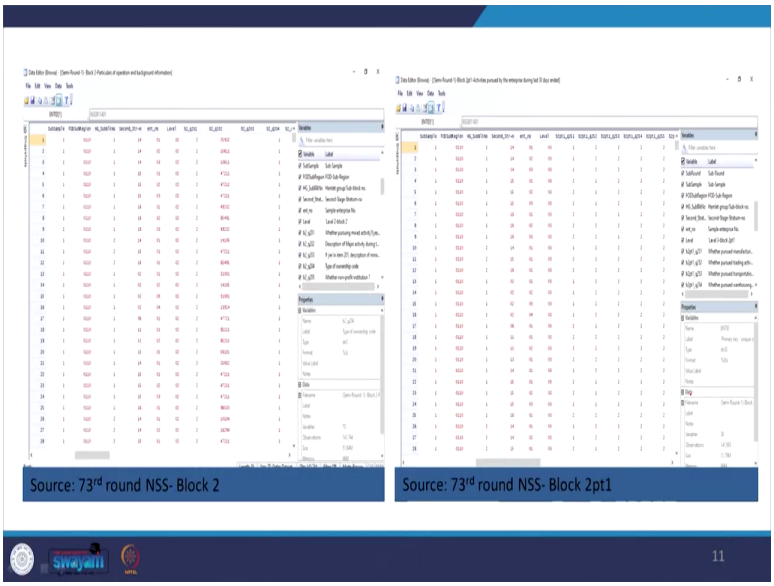
(Refer Slide Time: 16:55)



Each block contains different set of information. As I just, said data structure are usually different then. So, there is a separate data layout for different blocks. So just by merging different

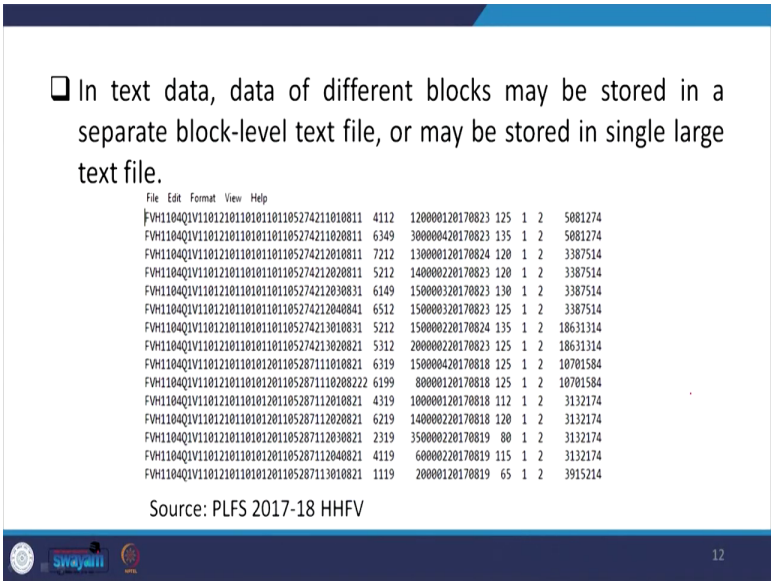
blocks without giving the correct specification and identifying the unique characteristics of, or unit, unique idea of it is actually completely meaningless. It will give you a nonsensical direction. So, we have to structure it carefully. Just blocking is actually not suggested; just merging is not suggested at all. So in dataset, data of different blocks are stored in different tables or in separate files.

(Refer Slide Time: 17:38)



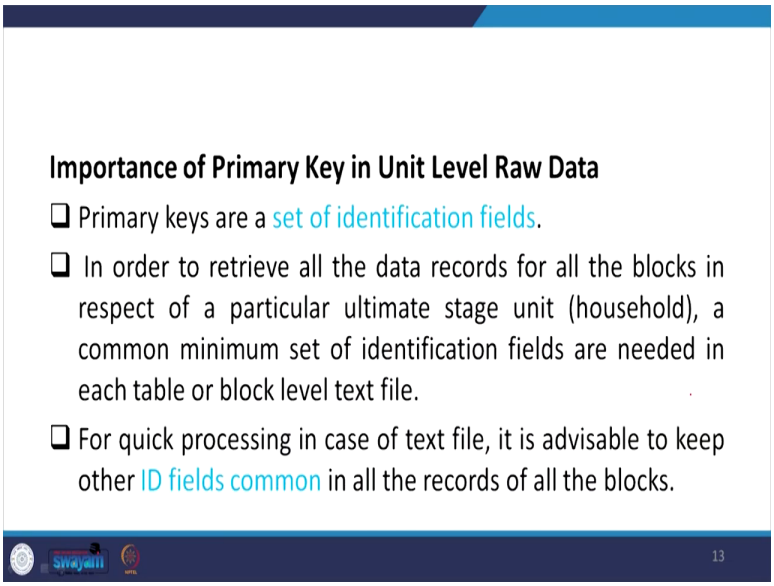
This is the layout, the snapshot from our STATA software and we have the 2 blocks information given here, and we will explain it in detail, since it is not visible, just for your knowledge I am keeping it. If you open the dataset or by block, you will have this kind of features. I am not explaining because the font is very small and we will actually decode very correctly in our respective session.

(Refer Slide Time: 18:08)



I have already shown Notepad file, Notepad file earlier in the last lecture. So, in the text data as I said, this is also called ASCII data and usually mentioned in Notepad file. The text data, data of different blocks may be stored in a separate block level text file or may be stored in a single large text file.

(Refer Slide Time: 18:36)

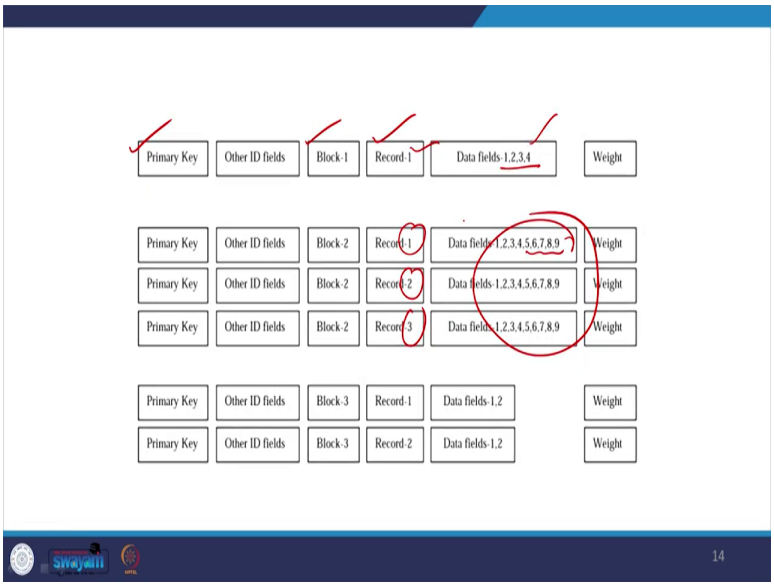


And the importance of primary key in the unit level data is very very important starting point for data understanding and also for extraction or interpretation. Even in my knowledge after having

some years of experience in unit level data I do understand that at least 20 to 30 percent of knowledge, even more than that, is derived just by understanding background of those datasets, the structure of the dataset. Number of interpretation can be made without any errors.

So, the primary keys are very important. Primary keys are set of identification fields and those identification fields are very useful for extraction. In order to retrieve all the data records, all the blocks in respect of a particular or ultimate stage unit, that is household, a common minimum set of identification fields are needed in each table or block level text file. For a quick processing in case of text file it is always suggested that to keep your other ID fields common, identifiers common, variables in all the records of all the blocks, without that it is not possible. Though it is merging but it is actually having a meaningless information.

(Refer Slide Time: 20:04)



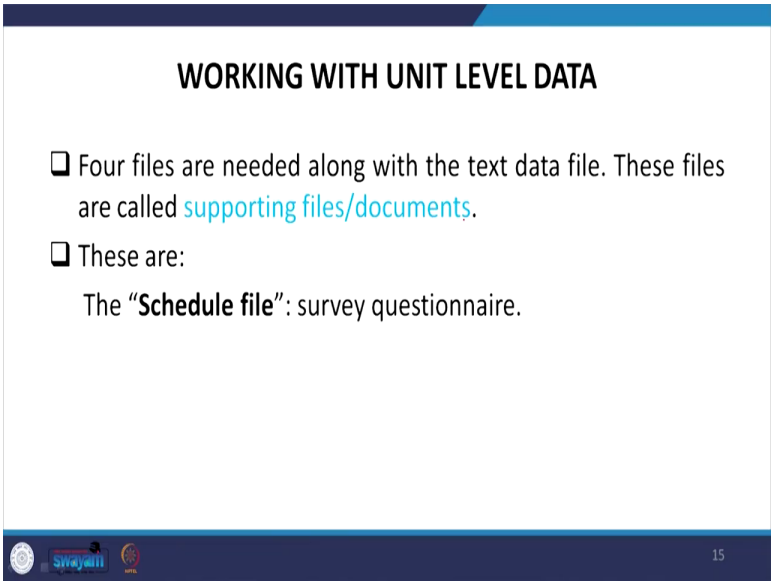
This is the structure. If you look at the screen very carefully, the informations are given like this we are referring to the latest NSS dataset. The initial informations are given as primary key in the respective file, in the page, primary key, then with their space defined, location, the other ID fields that I just mentioned a couple of minutes back. This is also important for merging and followed by the block information, in which block this information has been collected and in which block this has been captured like each block is different than that of other block because it gives a detailed information of that particular information.

A record refers to which round the study is conducted, whether it is in the first round or second round. It is not necessarily the fact that all the units, the final units of the study, that is the third unit we have discussed in our lecture here, that the third unit may be individual or even the second unit, household. The second unit is actually studied repeatedly over time or not repeatedly or a separate complete separate household is studied.

If it is a complete separate household, if you are simply merging all the information you are actually getting not a correct picture. why am I saying, I will clarify all those details in the respective NSS data. So, there are different rounds. Record 1, record 2, record 3, again in another file I will find the same differently. In addition to that, there are certain fields, data fields and data may contain, like in the first record you have only 4 information. In the second, there may be more information. In the third there may be even more information.

Usually in the latest round of NSS, in the second, third and fourth repeated visit, it is called repeated because the urban is actually the common, urban household is the common one, and the same information has been collected.

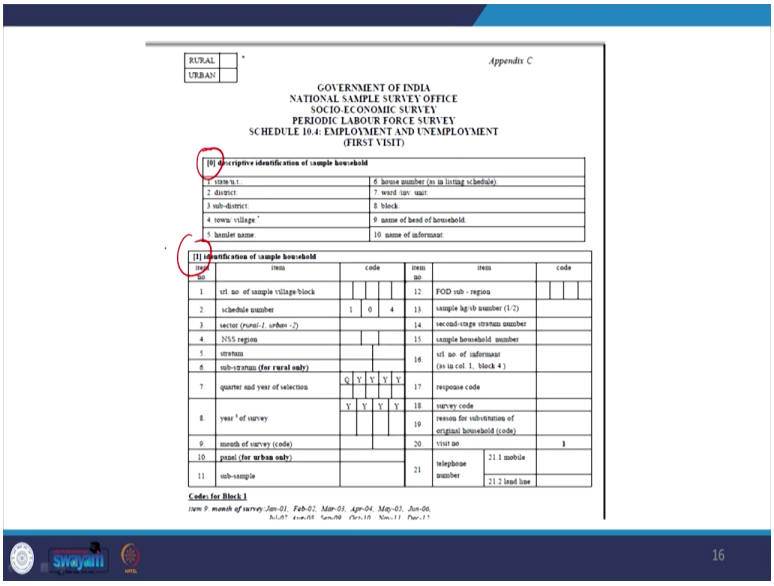
(Refer Slide Time: 22:47)



And last information is added with weight of the particular unit and weight is given to make the data more representative. Now let us work a bit on the unit level data. So, broadly there are 4 files, very very essential, to work or as a starting point for extraction and merging. These files

are also called supporting files or documents. The first important supporting file or the document to start working is none other than schedule file. The schedule file is also called a schedule, also called survey questionnaire.

(Refer Slide Time: 23:08)



RURAL
URBAN

Appendix C

GOVERNMENT OF INDIA
NATIONAL SAMPLE SURVEY OFFICE
SOCIO-ECONOMIC SURVEY
PERIODIC LABOUR FORCE SURVEY
SCHEDULE 10.4: EMPLOYMENT AND UNEMPLOYMENT
(FIRST VISIT)

[0] descriptive identification of sample household

1. state/u.t.	6. house number (as in listing schedule)
2. district	7. ward /urv. unit
3. sub-district	8. block
4. town/ village*	9. name of head of household
5. hamlet name	10. name of informant

[1] identification of sample household

item no.	item	code	item no.	item	code
1.	srl. no. of sample village/block		12.	FOD sub - region	
2.	schedule number	1 0 4	13.	sample hg/vb number (1/2)	
3.	sector (rural-1, urban -2)		14.	second-stage stratum number	
4.	NSS region		15.	sample household number	
5.	stratum		16.	srl. no. of informant (as in col. 1, block 4)	
6.	sub-stratum (for rural only)				
7.	quarter and year of selection	Q Y Y Y Y	17.	response code	
8.	year[1] of survey	Y Y Y Y	18.	survey code	
			19.	reason for substitution of original household (code)	
9.	month of survey (code)		20.	visit no	1
10.	panel (for urban only)		21.	telephone number	21.1 mobile
11.	sub-sample				21.2 land line

Code: for Block 1
item 9: month of survey: Jan-01, Feb-02, Mar-03, Apr-04, May-05, Jun-06,

16

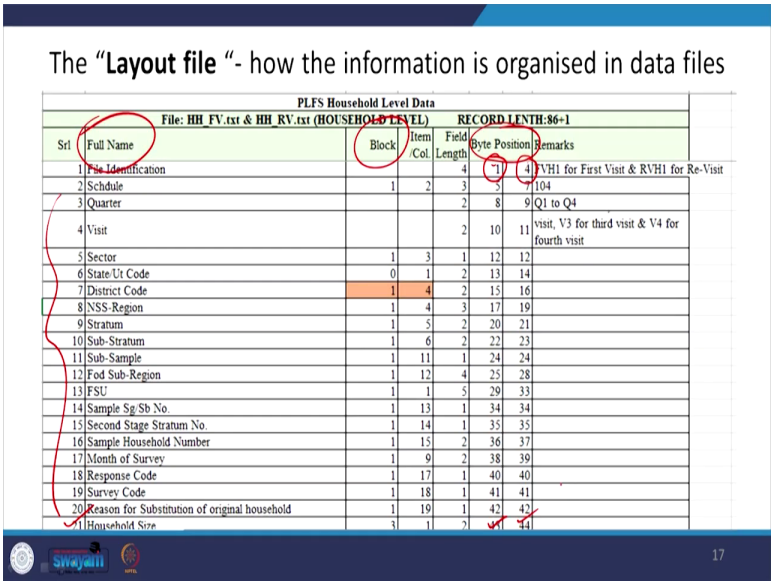
It looks like this. The survey questionnaire, the snapshot we are giving it. You can go through the PLFS, you can type MOSPI and you search PLFS. It gives the information in all those details. Even in IHDS also, if you simply type IHDS, India Human Development Survey and their schedule it also gives its appropriate link. We will provide all those links also, links and details in our respective lecture for IHDS, NSS, NFHS. And this is the schedule.

I just wanted to mention as I already started that this is a block 0 and 1. Usually they are the very starting point of information to understand the unique identifiers. And the other blocks after that actually very important for understanding our variables, information, quality information. Last one would certainly add the weights but here we have only given the snapshot of the first, 0 and 1, and 0 gives the information like state ID, district ID, sub-district, town, village.

Interestingly town and village may not be actually repeated. So, we cannot consider that in different rounds of NSS. These are same. At maximum, till state are all same. If those who are interested for working with panel, it is not that easy to convert panel because NSS is not at all a

panel data. Some possible pseudo format can be developed. We will discuss those in our last week module.

(Refer Slide Time: 25:02)



The "**Layout file** "- how the information is organised in data files

PLFS Household Level Data

File: HH_FV.txt & HH_RV.txt (HOUSEHOLD LEVEL) RECORD LENTH:86+1

Srl	Full Name	Block	Item /Col.	Field Length	Byte Position		Remarks
1	File Identification			4	1	4	FVH1 for First Visit & RVH1 for Re-Visit
2	Schdule	1	2	3	5	7	104
3	Quarter			2	8	9	Q1 to Q4
4	Visit			2	10	11	visit, V3 for third visit & V4 for fourth visit
5	Sector	1	3	1	12	12	
6	State/Ut Code	0	1	2	13	14	
7	District Code	1	4	2	15	16	
8	NSS-Region	1	4	3	17	19	
9	Stratum	1	5	2	20	21	
10	Sub-Stratum	1	6	2	22	23	
11	Sub-Sample	1	11	1	24	24	
12	Fod Sub-Region	1	12	4	25	28	
13	FSU	1	1	5	29	33	
14	Sample Sg/Sb No.	1	13	1	34	34	
15	Second Stage Stratum No.	1	14	1	35	35	
16	Sample Household Number	1	15	2	36	37	
17	Month of Survey	1	9	2	38	39	
18	Response Code	1	17	1	40	40	
19	Survey Code	1	18	1	41	41	
20	Reason for Substitution of original household	1	19	1	42	42	
21	Household Size	3	1	2	[illegible]	44	

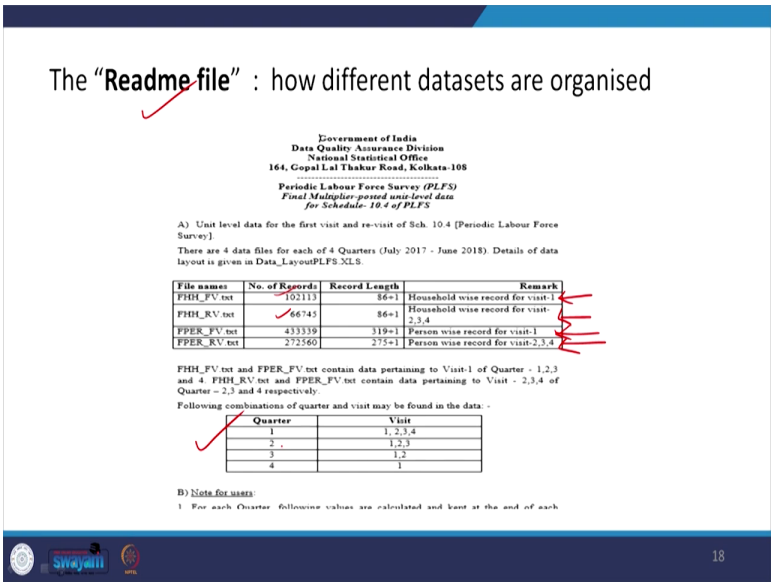
The second important file is called layout file. Layout file is very very important to extract because it gives position of the particular character or the variable position. there are 3 important information required within the layout file. One is name, which names are given. Basically names for us are the variables and another byte position and block information.

So, block information as I mentioned, different blocks are there, so let me also refer to the example of our dataset on enterprises called Economic Census. Economic Census is the dataset for the enterprises, all forms of enterprises, informal as well as formal. And it keeps the blocks as all the states at the blocks. So, during that time it was of 28 states were there.

So , if there will be any, economic census, the latest round of that economic census was sixth, sixth economic census. we do not have more rounds yet. If it is published you can check accordingly. let me also understand some of the information for defining the common characteristics. It varies from NSS round to different rounds. So, for the common identifier, usually we go for state ID, we go for household ID. We go for other detail like sample ID even. So, there are district code as well.

So, we will discuss these in detail later and there are block information, byte position. I need to add one line on it that byte position defines the starting position till the end position in the ASCII format. If you start with the first position we check till fourth position contain file identification. let me pick up, one particular information, household size at 21, byte position is 43 to 44. So, that means there are 2 space given. So, at maximum, household size can have 2 digit. It cannot be of 3 digit. 100 and more is not possible. It at maximum, household size is actually limited to 99 as we all know. So, accordingly it is defined and we will use it later.

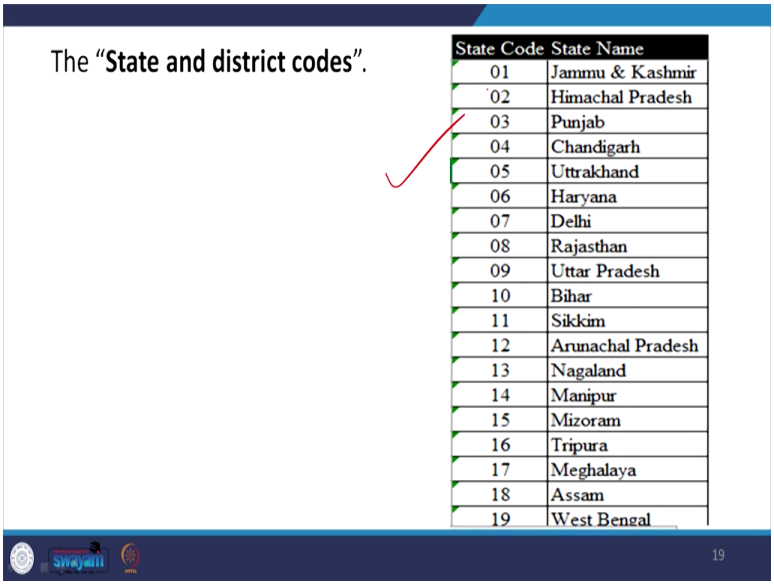
(Refer Slide Time: 27:52)



Regarding the understanding of the third most aspect is called Readme file. for you, Readme file is important because of the fact that it contains so many information of each variables, therefore are clearly described. as I mentioned here that we are referring to this Readme file of PLFS round, it is PLFS round 2017-18 studied from July 2017 to June 2018.

There are different rounds as I mentioned, household-wise record for Visit 1 is given. Then household-wise record in visit second and third and fourth, how many households, number of records is given here. And in the next one, person is also recorded, person-wise record in Visit 1, and in other 2, 3 and 4. Here, in the chart given, in which quarter those have been studied, in the first quarter, second quarter or third quarter. So, you can have those detail information in our respective lecture.

(Refer Slide Time: 29:15)



The last information in this segment, in this particular four information, 4 important files is state and district code. Because if you do not understand the district code or the state code, interpretation will be very difficult. Though merging can be done because district ID you will get it, and their byte position you will get it very clearly and with the command you will get the merging. But if you interpret further with the statistical package it will give meaningless result. So, the state code as well as the district code is very very important. And this kind of picture is there, we have taken the snapshot of the original file.

(Refer Slide Time: 30:00)

Then What is left in this particular segment, as I mentioned some unit level data will be handled in our due course of lectures and we will discuss NSSO, this is the original symbol, given in the MOSPI website. You go through that. You will see this. IHDS is very very important so far as recent database is concerned. Though the last round study was 2011, it is by University of Maryland, there are topmost team involved and University of Maryland and our NCAER, they are two topmost institutions in the world, observe the data and the data is very qualitative.

And most importantly this database is panel and very useful for analysis because panel data usually gives policy based suggestions.You can track the particular household and their improvement or detoriation over time. So, NFHS data (National Family Health Survey data) that latest round is 2015-16, that is NFHS 4. and the last round NFHS 5 is almost studied and they are about to publish, may be in 6 months time but not exactly defined with the time.

So it is by IIPS (International Institute of Population Sciences). So, these are all the information in our module for lecture number 2. We will carry forward rest of the details in our successive classes. With this thank you so much.