

**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. 23**  
**Combining Datasets in Stata-I**


Welcome friends once again to the NPTEL module on Handling Large-Scale Unit Level Data Using Stata. Here, we are trying to enrich you with some handholding exercise of Stata. Over the last two lectures we are continuing with handholding of unit level data with Stata. Last class particularly the handholding was on extracting and the data extraction was dealt in different context, specially on, if there are different type of data available, I discussed like infix or fix format, free format, delimited format, if those datasets are available accordingly different ways of extraction were discussed and guided to you.

Today we are trying to explain you combining of datasets. After extraction dataset should be combined. Why combining is important, what is the advantage of it, if you do not combine then what is going to happen and just combining is important or horizontal combining or vertical or combining with continuing with the time variant aspect as well. There are various ways of combining dataset. But at this moment, time aspect we are not dealing with. We will have separate lecture on it.

(Refer Slide Time: 02:09)

### INTRODUCTION

- ❑ In many empirical research projects, the raw data to be utilised are stored in separate files. For example, separate files for household and individual records, separate files for each states, separate waves of panel data, separate files for each category of information (blocks in the questionnaires) etc.
- ❑ Combining two datasets is a common data management tasks. **Merging** and **appending** are such tools that combines two datasets.

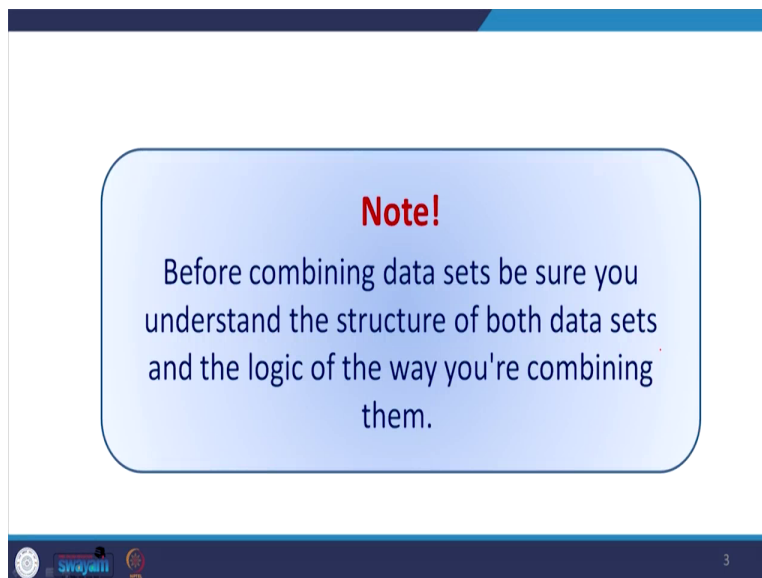
2

Let me explain to you without putting different words on combining, let me straight away clarify to you the systematic approach of understanding combining datasets. So, in many empirical research projects these days, the raw data to be utilized are stored in separate files. For example, separate files for household and individual records, separate files for each states or separate wave of panel data like as I mentioned about time, contains the same individual list studied several times over different period that is related to panel. We will discuss later.

Several files such as the category of information in different blocks is usually given in National Sample Survey data, NSS. There are different blocks maybe for consumer expenditure, different heads of expenditure are given in different blocks. Why they are giving, because first of all separating the data as per the need and that too making a structured format and that too dealing with the space of separating files. If a researcher only wants some of the blocks, not necessarily merging everyone together to occupy the space, some of them, we discussed earlier.

But now onwards we are guiding you from the perspective of combining two datasets in a common data management task. So, the two approaches are important that is merging and appending that combines two datasets.

(Refer Slide Time: 03:48)



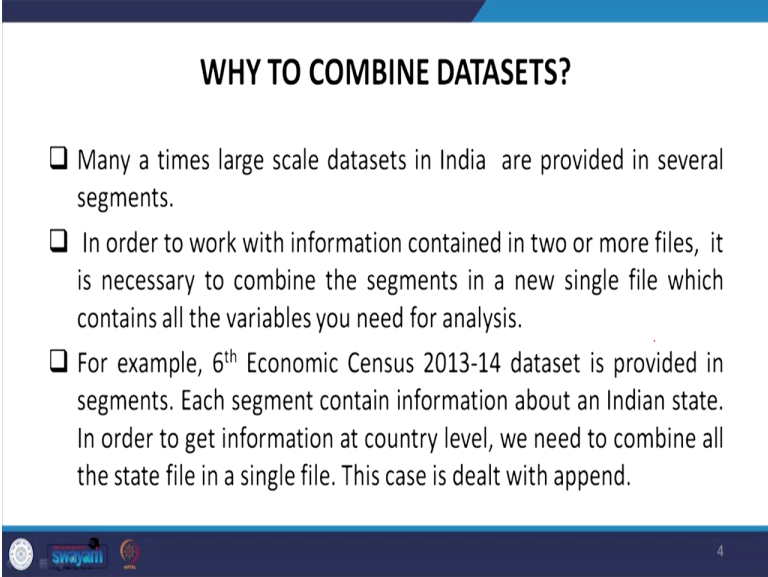
**Note!**

Before combining data sets be sure you understand the structure of both data sets and the logic of the way you're combining them.

Swayam 3

One of the note here for you before combining dataset is that you must be sure to understand the structure of both datasets and the logic of the way you are combining both of them. So, the structure and the logic of combining is very important as well.

(Refer Slide Time: 04:08)



**WHY TO COMBINE DATASETS?**

- ❑ Many a times large scale datasets in India are provided in several segments.
- ❑ In order to work with information contained in two or more files, it is necessary to combine the segments in a new single file which contains all the variables you need for analysis.
- ❑ For example, 6<sup>th</sup> Economic Census 2013-14 dataset is provided in segments. Each segment contain information about an Indian state. In order to get information at country level, we need to combine all the state file in a single file. This case is dealt with append.

4

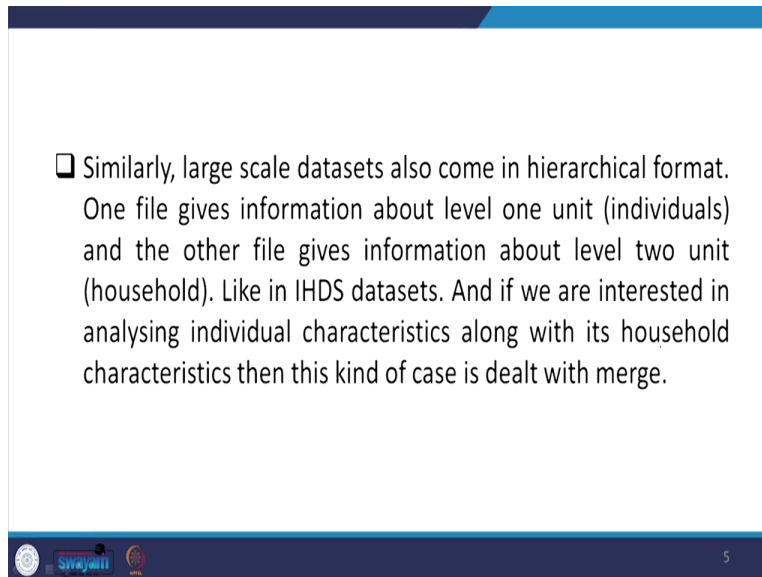
So, why to combine datasets, what are the necessity of combining datasets, I have just said, but you can take a note as per the points contain in this slide. Many a times large datasets in India provided in several segments. In order to work with the information contained in 2 or more files, it is necessary to combine a segment in a single file which contains all the variables you need for analysis.

For example, the sixth economic census that was published in 2013, provided the information in segments. Each segment contains information about an Indian state. So, in order to get information at country level, we need to combine all the states file into a single file. This case is dealt with the help of append command. Regarding sixth economic census 2013-14 dataset, if you go back to our very earlier lectures, we have modules on understanding datasets.

So, we discussed different blocks of information which are available in the sixth economic census data and the raw data is also available and we should learn because it is not just the extracted data available and we will not learn extraction that is not the right idea, because many data which are not extracted, but available, later on when you get the extracted one many

researchers might have taken the advantage out of the raw data by their own way of extraction as well as combining the datasets after extraction. So, since extraction we already discussed in the last lectures, we are helping you to understand append of this kind of information.

(Refer Slide Time: 06:23)

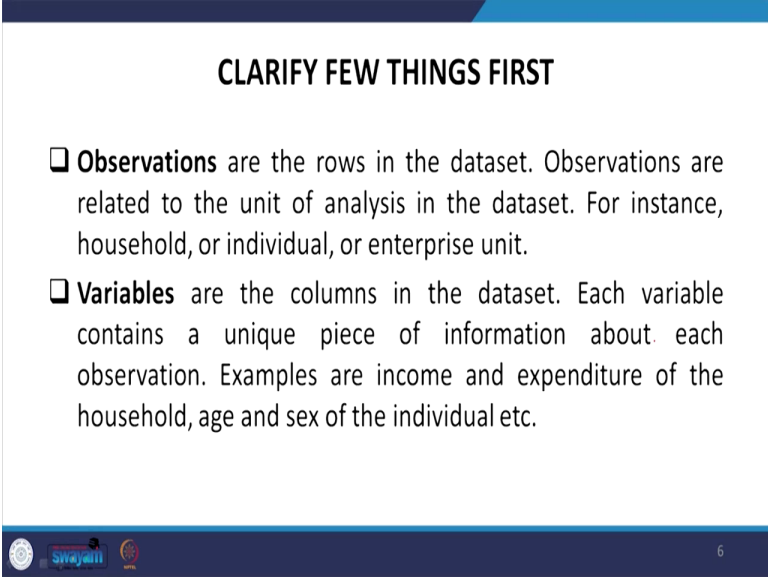


So, another important aspect of combining is that, large scale datasets also come in hierarchical format. One file gives information about one unit that is individual and the other file gives information about the level two unit that is household. This is generally available in IHDS data format, IHDS, India Human Development Survey. The latest one is 2011-12. That we discussed already. We already have had different session on different datasets which we are handling.

If we are interested in analyzing individual characteristics along with its household representatives or the information, then this kind of case is dealt with merge. So, basically either household information is merged with individual or individual information is merged with household. So, we will discuss in which approach we are supposed to merge, whether one to another or another to the first. We have some logic behind merging. Just randomization of merging is resulting you into nowhere with huge blunder and though some result you may derive but those may not be authenticated while checking with the original report. So, please hang on till the time and we are going to guide you systematically.

Now let us clarify few things first before combining datasets. First aspect of combining dataset is understanding observations. So, we have already said repeatedly that observations are generally present in the row in the datasets and those are related to unit of the analysis in the dataset.

(Refer Slide Time: 08:21)



**CLARIFY FEW THINGS FIRST**

- ❑ **Observations** are the rows in the dataset. Observations are related to the unit of analysis in the dataset. For instance, household, or individual, or enterprise unit.
- ❑ **Variables** are the columns in the dataset. Each variable contains a unique piece of information about each observation. Examples are income and expenditure of the household, age and sex of the individual etc.

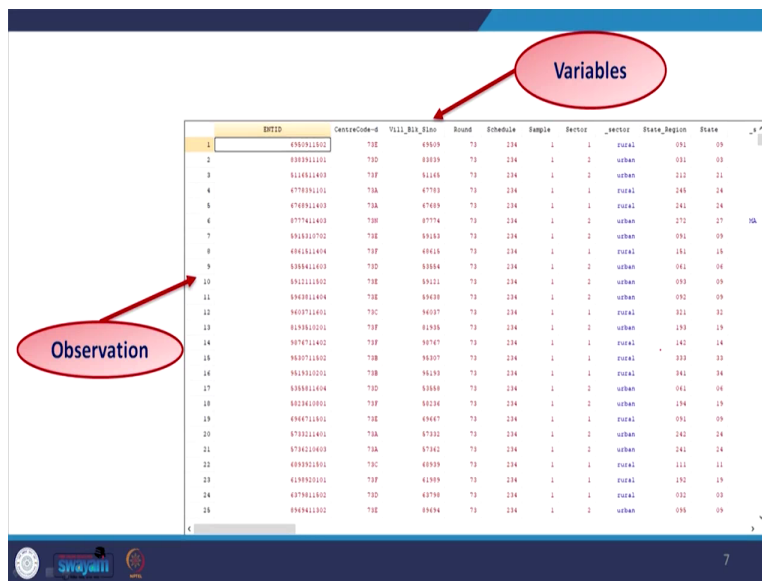
6

For instance, household or individual or the enterprise in the context of seventy third round of the NSS we discuss enterprise, Number of enterprises are the observations. And those are generally present in rows. regarding another most important aspect is called variable. Variables are the columns in the dataset. Each variable contains a piece of information about each observation. So, against to each observation there are variables on the columns. So, each column gives information about the observation. So, those columns are called variables.

So, we also try to understand what are those variables, which variables are important for me? I may take entire number of variables, but as I mentioned to you in National Family Healthy Survey, more than 5,000 variables are present, which is very difficult to remember at a go. So, you are supposed to deal with tactically or cleverly or with a systematic approach of keeping some variable or dropping some variable which we already discussed earlier, keep and drop are important. We will also have a review session at the end of this week, how to review the variables which already dealt in the previous lecture. So, we enrich you further on that direction.

For example, variables in our case are income and expenditure of the household, age or sex of the individual etcetera. Usually in the consumer socio-economic survey datasets you will have those variables very common and these are regarded as variables for us to append as well as for merging.

(Refer Slide Time: 10:20)

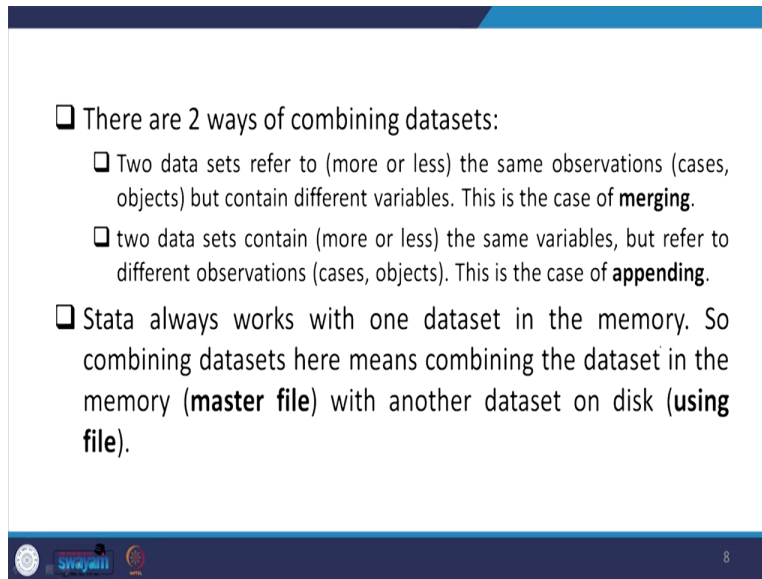


The image shows a screenshot of a data table with 18 rows and 11 columns. The columns are labeled: ID, CentreCode-H, Vill\_Bli\_Stage, Round, Schedule, Sample, Sector, \_sector, State\_Region, State, and \_id. The rows contain numerical data. Two red callout boxes are present: one labeled 'Variables' pointing to the column headers, and another labeled 'Observation' pointing to the first row of data.

ID	CentreCode-H	Vill_Bli_Stage	Round	Schedule	Sample	Sector	_sector	State_Region	State	_id
1	4593911002	708	45939	70	204	1	1	INDIA	091	09
2	4593911002	708	45939	70	204	1	2	urban	091	09
3	4593911002	708	45939	70	204	1	1	INDIA	212	21
4	4593911002	708	45939	70	204	1	1	INDIA	245	24
5	4593911002	708	45939	70	204	1	1	INDIA	241	24
6	4593911002	708	45939	70	204	1	2	urban	272	27
7	4593911002	708	45939	70	204	1	2	urban	091	09
8	4593911002	708	45939	70	204	1	1	INDIA	181	18
9	4593911002	708	45939	70	204	1	2	urban	041	04
10	4593911002	708	45939	70	204	1	2	urban	090	09
11	4593911002	708	45939	70	204	1	2	urban	092	09
12	4593911002	708	45939	70	204	1	1	INDIA	321	32
13	4593911002	708	45939	70	204	1	2	urban	190	19
14	4593911002	708	45939	70	204	1	1	INDIA	142	14
15	4593911002	708	45939	70	204	1	1	INDIA	330	33
16	4593911002	708	45939	70	204	1	1	INDIA	341	34
17	4593911002	708	45939	70	204	1	2	urban	041	04
18	4593911002	708	45939	70	204	1	1	INDIA	194	19
19	4593911002	708	45939	70	204	1	1	INDIA	091	09
20	4593911002	708	45939	70	204	1	2	urban	242	24
21	4593911002	708	45939	70	204	1	2	urban	241	24
22	4593911002	708	45939	70	204	1	1	INDIA	111	11
23	4593911002	708	45939	70	204	1	1	INDIA	192	19
24	4593911002	708	45939	70	204	1	1	INDIA	032	03
25	4593911002	708	45939	70	204	1	1	INDIA	056	05

The sample dataset we have many times shown to you. I am just once again showing to you that the upper part, that is, in the vertical portion those are the variables, the observations are mentioned in the rows. And there are different ID, unique ID or not we need to also understand in a short while.

(Refer Slide Time: 10:44)



- ❑ There are 2 ways of combining datasets:
  - ❑ Two data sets refer to (more or less) the same observations (cases, objects) but contain different variables. This is the case of **merging**.
  - ❑ two data sets contain (more or less) the same variables, but refer to different observations (cases, objects). This is the case of **appending**.
- ❑ Stata always works with one dataset in the memory. So combining datasets here means combining the dataset in the memory (**master file**) with another dataset on disk (**using file**).

There are two ways of combining datasets, two important ways. One is merging, another is appending. Two datasets refer to more or less the same observations, what do you mean by same observations, those are cases or objects, but those contain different variables and number of observations or cases we are going to take from two datasets and having different variables we need to get the different variables to be merge with the concerned observation, with the same observation so that is called merging.

So, I will tell you, there might be some difficulties at this moment to understand the concept. When we operate it and show, you can easily clarify the meaning of these things. Similarly, two datasets contain same variables, variables in the one dataset and another dataset, the number of variables are same, but referred to different observations.

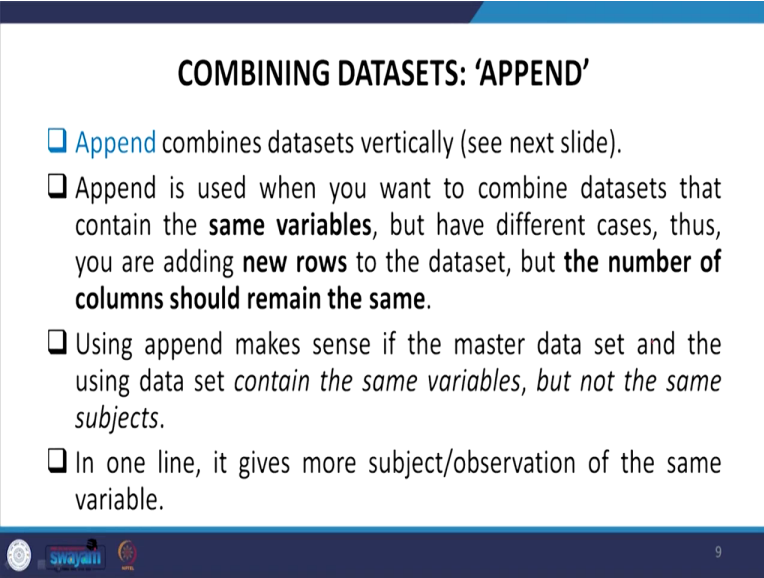
Here like state, as I told you already, there are several state information given in the sixth economic census data, so each state contains different number of enterprises and so we are supposed to add those, first state, second state, third state, fourth state like the entire state if you want to combine and to receive all India information, you are supposed to append them. In rough sense, it is called adding the observation as per the variable.

So, Stata always works with one dataset in the memory. So, combining datasets here means combining the dataset in the memory that is the master file. So, master file and that is the base

we wanted to merge with another one. So, whichever extra we are going to merge or even append, those are called using file. The another one called using file, but the first one, but why first there are some logic.

The why first, why master, first is even not the right meaning, the master, we are going to guide you on why that is called master, because it contains the base information and other information we are going to add or merge. So, there are two ways, I will do in a short while in our next couple of slides. I will be going to guide you on this direction.

(Refer Slide Time: 13:36)



**COMBINING DATASETS: 'APPEND'**

- **Append** combines datasets vertically (see next slide).
- Append is used when you want to combine datasets that contain the **same variables**, but have different cases, thus, you are adding **new rows** to the dataset, but **the number of columns should remain the same**.
- Using append makes sense if the master data set and the using data set *contain the same variables, but not the same subjects*.
- In one line, it gives more subject/observation of the same variable.

9

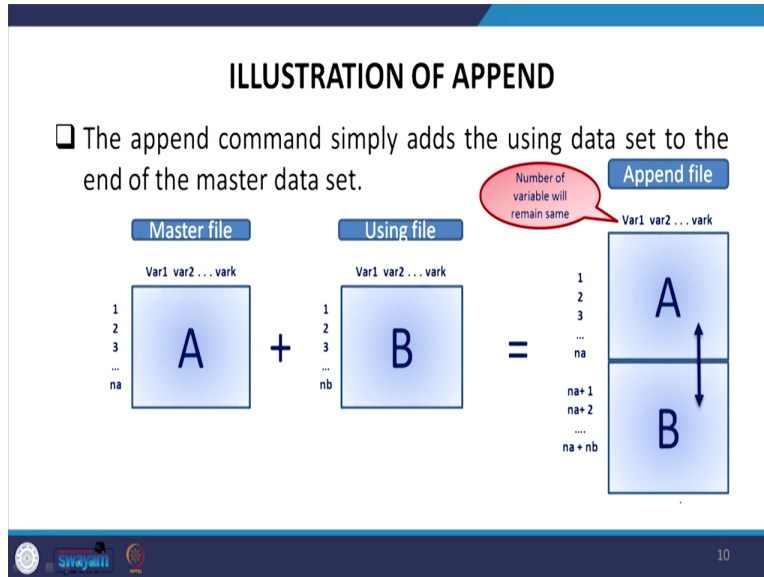
Combining datasets as I told you in short it is called append in Stata. So, append combines dataset vertically. Our next slide will guide you like this how, why it is called vertically? So, let me just give you some clarification in terms of interpretation then I will explain that diagram for you. Append is used when you want to combine datasets that contain the same variables, as I mentioned in the previous slide, but have different cases or observations, thus you are adding new rows, additional rows to the original or the master dataset or the first dataset, but the number of columns should remain the same.

So, please mark between the lines wherever we have mentioned in bold font so those are important to remember. Number of columns should remain the same, but number of observations should rise after append. Using the append make sense if the master dataset and the using dataset



contain the same variables but not the same subjects. In one line, it gives more subject or observation of the same variable which I have mentioned.

(Refer Slide Time: 15:14)



Let us understand what do you mean by vertical alignment or vertical addition or the data. The append command simply adds the using dataset to the end of the master dataset. So, for us, here let A be our master dataset having rows as a number of observation and variables in the column as the variable identified by variable 1 to variable k. Similarly, the using dataset has number of observations till nb, 1 till nb and with variable, having the same variable is important here, because then only we can able to append vertically.

A and B are vertically appended which is very clearly visible from the picture. The number of variable will remain same. You can mark this very carefully and highlighted with this picture that variable 1 variable 2, variable k is same now. We have an increased number of observation that is na plus nb in total. I hope you could guess and understand things correctly and after appending in practice you can able to understand even in more detail.

(Refer Slide Time: 16:35)

- ❑ For example, **PLFS (periodic labour force survey) 2017-18**, for rural areas, in each quarter 25 per cent of households of annual allocation were covered. Information on visit one is provided in a separate file and remaining three visits in a different file. If you wish to analyse annual employment rate, you have to combine the datasets. As you are interested in adding all observation in one file, this case is dealt with the 'append' command.

Why appending is important in the very recent examples like the latest dataset which the Government of India published very recently which has raised huge hue and cry in the country related to the extent to massive unemployment. People are even claiming for more than 6 percent unemployment in India based on the figures of PLFS, Periodic Labor Force Survey 2017-18, even another latest has also been published by Government of India.

What is intriguing here for you to note is that, this new format of data, earlier it used to be relating to labor force survey. Those who indeed deal with unemployment or labor's research let me point it out that labor survey usually come up with the round called employment and unemployment rounds of National Sample Survey. For the first time Indian government has changed to the name called Periodic Labor Force Survey. Since the name periodic, we are expecting the data to be released every year or they define a clear periodic information on it and there are some reasons behind as well.

Another information let me give it to you in this regard making you in a lighter format that those who work for labor issues that employment-unemployment round of the data usually come up with another data called expenditure round. But this time Government of India has not yet published the expenditure round along with the Periodic Labor Force Survey. So, what is becoming very difficult here is that while understanding the labor force, their nature of

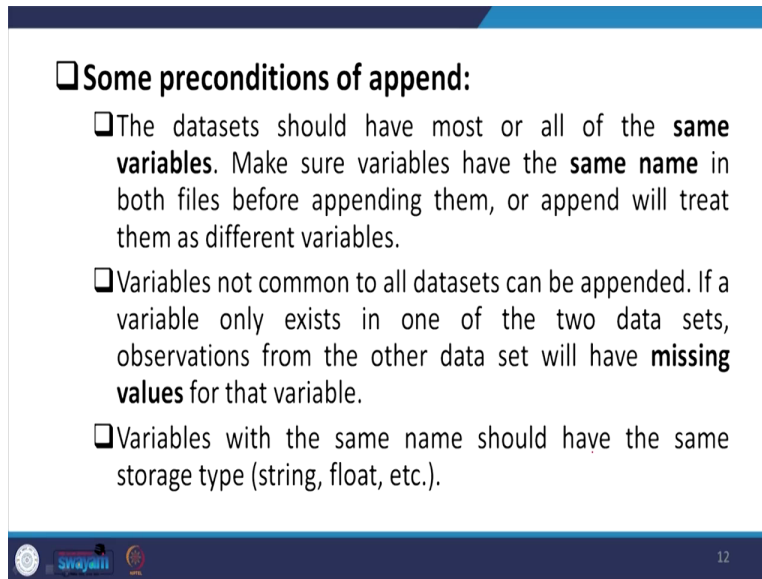
employment, but we cannot derive their expenditure pattern of the labor force, because we do not have the data yet released for the 2017-18 expenditure.

So, that is one of the problematic aspect of the current data. But let me guide you about PLFS, because then only we can able to combine the datasets correctly. This dataset has rural areas and urban areas, but in rural areas, one thing they have followed is 25 percent household at a time. In each quarter they surveyed 25 percent of the households of the annual allocations and so 25, 25, 25, 25. So, the another 25 they are different in different quarter for the rural areas only, mark that it is for rural area.

So, if you want to get the information about all India and rural employment scenario, you need to add all the quarters then only entire population which we have surveyed can be covered and entire information can be captured. But if you just go by one block, one information that gives you only 25 percent of the households of the rural area, you cannot have the entire information. For the entire information you are supposed to, in this case, what you will do, you will certainly append, because append will add other segment of the population which were surveyed or responded.

So, information on visit one is provided in a separate file and remaining 3 visits in different files. If you wish to analyze annual unemployment rate, you have to combine the dataset. As you are interested in adding all observation in one file, this case is dealt with the append command as I mentioned.

(Refer Slide Time: 20:20)



**Some preconditions of append:**

- ❑ The datasets should have most or all of the **same variables**. Make sure variables have the **same name** in both files before appending them, or append will treat them as different variables.
- ❑ Variables not common to all datasets can be appended. If a variable only exists in one of the two data sets, observations from the other data set will have **missing values** for that variable.
- ❑ Variables with the same name should have the same storage type (string, float, etc.).

12

Some preconditions of append those are quite essential for you to understand is that the dataset should have all the same variables which we are repeatedly saying that same variable if it is there, very correctly you can able to append, isn't it. So, make sure variables have the same name also in both the files before appending them. Because if your variables are different, it does it means only the entries are different, if your variable name itself is different, then certainly it will differentiate the person from one round to another round or the observation from one block to another block.

So, you have to make sure that the same variable by name is also ensured before appending. So, append will treat them as different variables if the name is different. So, variables not common to all datasets can be appended. If a variable only exists in one of the two datasets, observation from the other dataset will have missing values for that particular variable. Like variables not common to all datasets can also be appended, but what is important here to note is that if variable only exist in one of the two datasets observation from other dataset will have missing values for sure, because it does not have that particular information in another dataset which we are combining. So, missing value is generated.

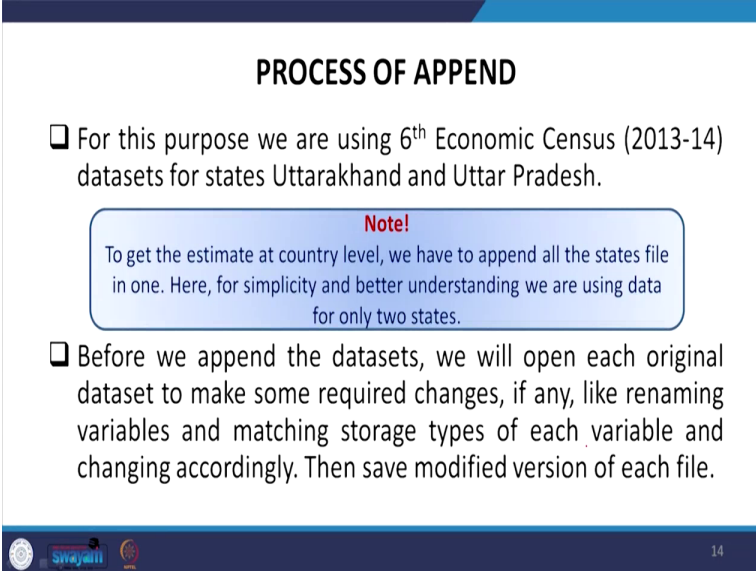
So, we even thought of discussing missing values, but let us see how we can able to deal with or not depending upon the space, limited space in this module, because there are very important aspects covered, likewise we guided you through twenty second lecture so far is on

understanding various aspects of handling Stata so let us keep some space for missing values, if it allows, we will certainly guide you.

Let me discuss with you that variables with the same name should have the same storage type also, not just the same, same storage type. Otherwise, it will read differently. Within the variable the entry will be read differently. So, same storage type either of it is string, it has to be within string, float or it has to be in float, but you can, if it is not there in that format you can do that as we already guided earlier. We will also guide you in our review lecture as well in this week.

So, the master dataset here is that the main dataset in which we want to add other observation. We want to get other observation to the master. So, that is why we are saying that is the main document or the main dataset. The appending dataset contains the observation that we are adding to the master dataset. We are adding the master dataset later on, along with the master dataset we have another dataset that will be regarded as appending dataset. Basically we are using the dataset to the master dataset. So, what are the process involved for appending, what process do we follow?

(Refer Slide Time: 23:49)



**PROCESS OF APPEND**

- ❑ For this purpose we are using 6<sup>th</sup> Economic Census (2013-14) datasets for states Uttarakhand and Uttar Pradesh.

**Note!**  
To get the estimate at country level, we have to append all the states file in one. Here, for simplicity and better understanding we are using data for only two states.

- ❑ Before we append the datasets, we will open each original dataset to make some required changes, if any, like renaming variables and matching storage types of each variable and changing accordingly. Then save modified version of each file.

14

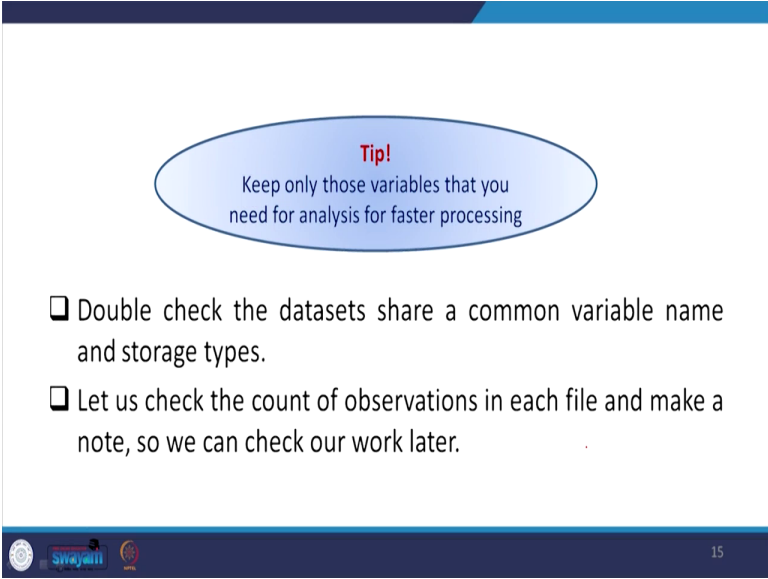
For this purpose, we are now referring to the sixth economic census that was published in 2013-14 by Government of India. For simplicity, we are only considering two blocks. For us it is

Uttarakhand and Uttar Pradesh, two states. We have so many others but for your easy understanding we are only trying to append two states information.

Please take a note that to get the estimate at country level, we have to append all the states file in one to get the entire country information as I already mentioned. Here for simplicity and better understanding, we are using data of only two states. Before we append the dataset, we will open each original dataset to make some required changes. Basically, you need to prepare the dataset as per append command or append process.

If renaming variable to make name of the variable should be same in both the datasets that is Uttarakhand and Uttar Pradesh for our example and matching storage type of the data of each variable that you can check with the color also. Also observation information is given at the right window of the Stata. Some information can be also understood. And accordingly if it is not that we have to make same. Then save modified version of both the files.

(Refer Slide Time: 25:29)



**Tip!**  
Keep only those variables that you need for analysis for faster processing

- Double check the datasets share a common variable name and storage types.
- Let us check the count of observations in each file and make a note, so we can check our work later.

Swayam 15

One of the tip for you is that, keep only those variables that you need for analysis for faster processing. Like as I told you in NFHS there are so many variables. You cannot even cross check or if you cross check it takes huge time. So, better to keep the variable, select the variables from the questionnaire or from the information short cuts or from the report file that which

information is required for you, accordingly, sort them and find out which are important and keep them for your record.

So, double check the datasets, share a common variable name and storage type. Let us check the count of observation in each file and make a note so that we can check our work later after appending. That is very very important for understanding whether appending has been successfully made or not.

(Refer Slide Time: 26:26)

The image shows two Stata Data windows. The first window, titled 'Master file', displays the following information:

Property	Value
Filename	EC6_uttarakhand_extracted.
Label	
Notes	
Variables	24
Observations	394,179
Size	21.05M

The second window, titled 'Using file', displays the following information:

Property	Value
Filename	EC6_Uttarpradesh_extracted
Label	
Notes	
Variables	24
Observations	6,683,905
Size	356.96M

The screenshot shows the Stata software interface. The command window contains the following text:

```
StataCorp.
    1995-2017 StataCorp LLC
    4905 LaSalle Avenue
    College Station, Texas 77845 USA
    800-STATA-PC      http://www.stata.com
    979-696-4600     stata@stata.com
    979-696-4602 (fax)

Special Edition

***
***      Stata lab generated command
***      Version number: 15.1
***      Licensed to: 117 Mooskee

Notes:
1. Unicode is supported; see help unicode advice.
2. Maximum number of variables is set to 5000; see help set_maxvar.

. use "G:\16th economic census\EC6_uttarakhand_extracted.dta"
```

The Variables list on the right side of the window shows the following variables:

Name	Label
state	
district	
tehsil	
block_village	
wardCodePur	
enumeration	
extended_u	
structureCode	
resident	
blockAICode	
NCD	
spatialAIC	
newenrptC	

The Properties window at the bottom right shows the following information:

Property	Value
Filename	EC6_uttarakhand.dta
Label	
Notes	
Variables	24
Observations	394,179
Size	21.05M
Memory	64M
Sorted by	

Regarding master file, in our case Uttarakhand the extracted version. It is not the raw data we are dealing with. We have to first extract the dataset. We already did that in our last lecture. So, Uttarakhand, once the extracted dataset is with you, open that and see the number of observations. So, here in our case it is 3,94,179. So, that you can also check from the data I will show you. So, it is here.

So, sixth economic census data, we have the data here, like it is Uttarakhand. I will show you now. It is going to open. In between it is taking little time, probably let me go back to our slide to guide you. So, we have to see that there are 3,94,179 observations from the original data. And for the using file in our case is Uttar Pradesh. As I told you, Uttarakhand and Uttar Pradesh we are going to append. Add them together.

So, for us it is 66,83,905 observations. So, I hope it has opened. There are some error coming may be due to byte space. I will certainly guide you in this regard. it is there. So, I will open Stata once again since some error came in between. I will try to do that. It is 64 bit will be opening. So, this is Uttarakhand. This has already been opened. As I told you, how to check the number of observation, I already told you it is visible here also 3,94,179. 3,94,179 this is there.

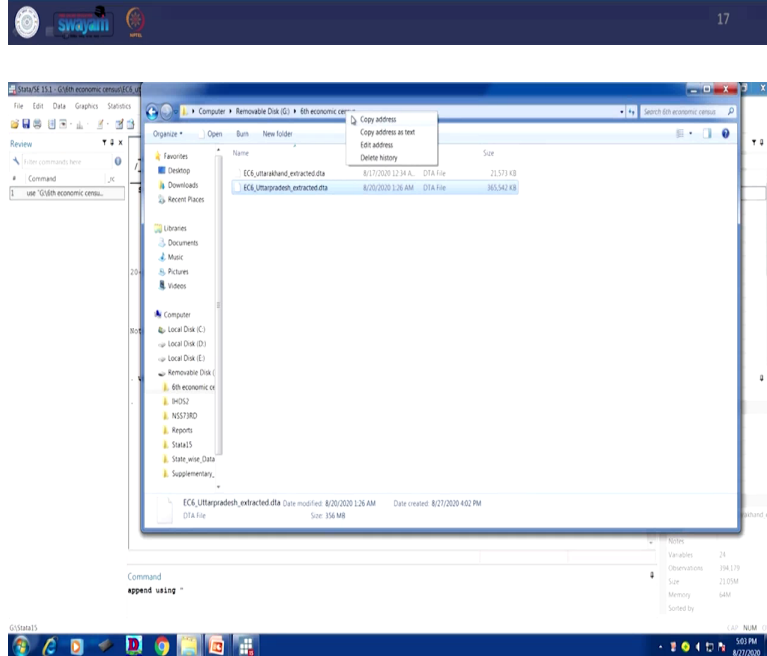
Similarly, I can open Uttar Pradesh file also and I can show you, but it is not necessary you can check, we will provide you those document for your own practice and you can double check. If you have difficulties do report to us, we will be happy to deal with, because if you have single error in between, single doubt in between, please take a note and try to put forth those discussions in a systematic structure manner so that it becomes very easy for us to also respond to you. And we will be happy to respond to you, because in between I am damn sure that you will have errors in the operation and then only you can able to learn very quickly.

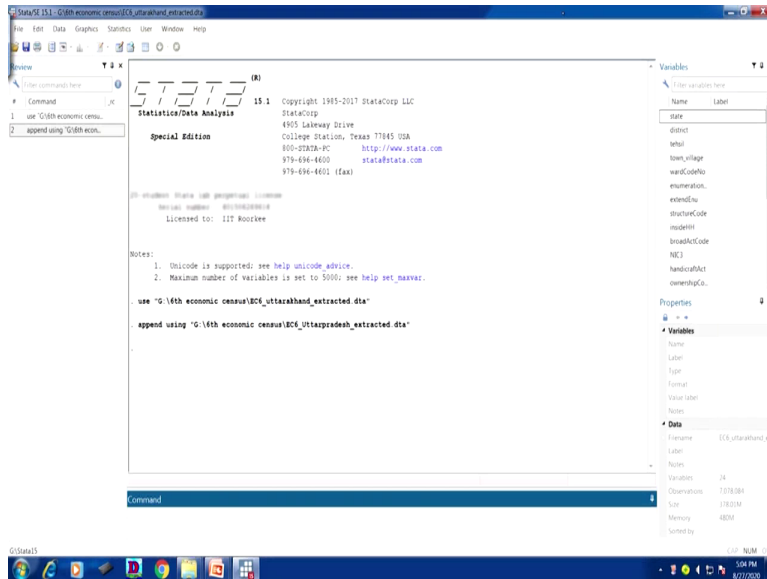
So, both the datasets have the same variable names and variable names you can check, we can also check though different approaches. I already guided you and we are ready to append them. In our dataset, both the variable names are same.



(Refer Slide Time: 28:53)

- ❑ Now that both datasets have the same variable names, we are ready to append them.
- ❑ First, open the master dataset (EC6 Uttarakhand\_extracted).
- ❑ Type in the command line:  
`append using "EC6 Uttarakhand_extracted"`
- ❑ We check that the data appended correctly by counting the observation.





First open the master dataset that is Uttarakhand. We have already opened. I am going to guide you how to append it. So, basically our command is append using inverted comma the path name of it. That is all for the operation. So, it is there, let me tell you.

Let me type here that append. We have already opened the master file. For us, master file is this one. You can also make a reverse approach. You can open Uttar Pradesh as well. Your master file will be Uttar Pradesh. But for us, for our purpose we are expecting this as the master file. So, append I will go for using, I will start with the inverted comma because the path name should be entering. how to get the path name we have to go to the file? This is the file. we need to go to the back one.

Here is the file. So, what we are using? we are using the Uttar Pradesh one. So, I will copy that first. So, right click on it, then copy address. I will go to the software then I am pasting it here. But still this is not complete. We have to add with the exact path name of that particular file that is EC6 underscore Uttar Pradesh I have to copy this. So, we are copying it. So, it has been copied and we are pasting it. closing that particular inverted comma with enter, we can show that, the append result is successfully done.

How successfully done? it is appending still you can mark. It is not complete. Still it is in the process. Look at the total number of observation 70,78,084, but earlier I am going to guide you in this regard and we will be very happy to see this and this is very interesting to understand as a

student of statistics or student of economics, students of management. This is very interesting to note. We check that the data appended correctly by counting the observation.

(Refer Slide Time: 32:27)

Data	
Filename	EC6_uttarakhand_extracted
Label	
Notes	
Variables	24
Observations	7,078,084
Size	378.01M

Now you can save the new permanent dataset with new name, here it is showing master dataset name

The number of variable will remain same as it is same in both the files.

394,179 + 6,683,905

Here it is like this. 70,78,084 as I just shown you this is there. Here it is 70,78,084, isn't it? So, we are going to explain you. So, our data has been successfully added because look at earlier the master file has 3,94,179 observations, it is added with 66,83,905 observations. So, that boils downs to 70,78,084 observations. Look at the number of observations, number of variables. So, variables were 24, it is also 24. We have to see that. Look at the number of variables before was also 24.

So, you can save the new permanent dataset that is permanent we mean we have appended it with new name, Uttarakhand underscore Uttar Pradesh or you can combine appended, whatever name convenient to you, but better to give the convenient name. Here it is showing the master dataset name because we have only opened the Uttarakhand dataset. And since we have already combined but if you carry with the same name, it will be very confusing later on to you. So, I will suggest that you please change the name as per your convenience.

(Refer Slide Time: 33:57)

If the storage type of a variable is different in master and using file. While appending stata will throw an error.

```
. append using "E:\stata_and_data\ECONOMIC CENSUS\Economic Census_2013 -14\EC6_Ut
> tarpradesh_extracted.dta"
variable sex is byte in master but str1 in using data
  You could specify append's force option to ignore this numeric/string
  mismatch. The using variable would then be treated as if it contained
  numeric missing value.
r(106);

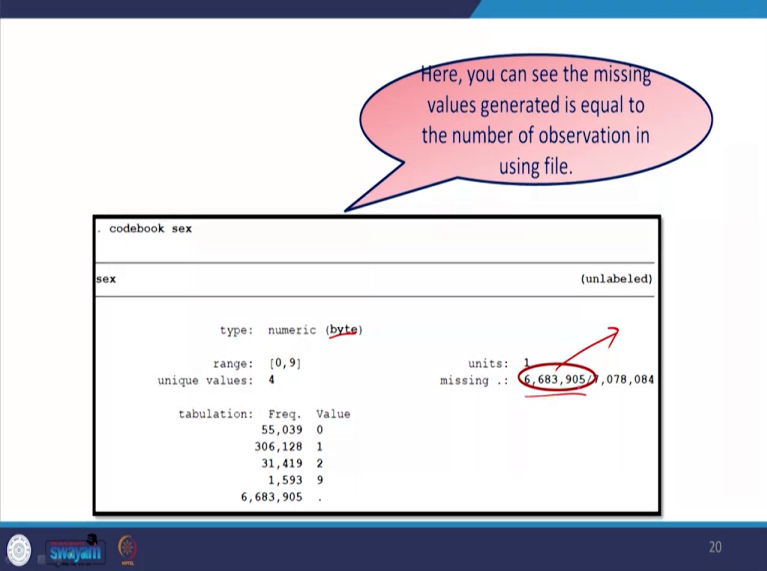
. append using "E:\stata_and_data\ECONOMIC CENSUS\Economic Census_2013 -14\EC6_Ut
> tarpradesh_extracted.dta", force
(note: variable sex was str1 in the using data, but will be byte now)
```

19

Another interesting aspect to be noted here that if the storage type of a variable is different in master and using file, while appending, Stata will throw an error if it is different. We have already guided you, it has to be the same. If it is there, there are some force way, forcefully we can combine it, but why we are doing it we are going to guide it. Append using which we have done it here variable sex is in byte. Byte in the master file, but it is string in the using data file.

You could specify append force option. Another force option in the bottom we are mentioning. To ignore this numeric or string mismatch the using variable would then be treated as if it contains numeric missing values. Like in the command in the below that it has an addition called force. You just mark this carefully. Here it is, force. Rest are same, if you just add comma and force, it will forcefully append your data. Note that variable sex was string one in the using data, but will be byte, because the master file is in byte format. So, it will be converted to byte.

(Refer Slide Time: 35:24)



Here, you can see the missing values generated is equal to the number of observation in using file.

```
. codebook sex
-----
sex                                     (unlabeled)

      type: numeric (byte)
      range: [0,9]
  unique values: 4
      units: 1
missing .. 6,683,905,078,084

      tabulation: Freq. Value
                  55,039  0
                  306,128 1
                  31,419  2
                   1,593  9
                   6,683,905 .
```

Look at, as I already guided you, it will throw an error. The number of observation in the using file was of 66,83,905, as I already told you, if that was in string, but the original was in byte, your converted one will be in byte, but this one is, as if this regarded. In our term it is called missing, considered as missing. So, here you can see the missing value is generated that is equivalent to the number of observations of the using file that is of 66,83,905. So, this is very very important to note, but what we will do.

(Refer Slide Time: 36:09)

- The picture in the last slides explains, if a variable is numeric in master data file and string in using data file. Stata will throw a warning that values of using file will be replaced with numeric storage type and converted into missing values.
- It also suggests to use force option if you agree with these changes.
- But it is not advisable to use force option as Stata will generate missing value for these particular variables from the using file and you can not alter these values once it is appended.
- So it is always suggested to make all the changes related to renaming of variables or storage type of variables before combining the files.

The bigger picture here is that, if a variable is numeric in master data file and string in using data file, Stata will throw a warning that values of using file will be replaced with numeric storage type and converted into missing values, we guided already. It also suggests to use force option if you agree with these changes. But it is not advisable to use force option as Stata will generate missing values for these particular variables from the using file and you cannot alter these values once it is appended.

After appending, what is the original value of it, it simply converted into missing that means the dot, it only gives dot. Dot cannot be changed later on and that too with a such a huge value with huge observations it is very difficult. So, it is always suggested to make all the changes related to renaming of the variables as storage type of the variables before combining the files.

So, in the next class, so far we have guided you about appending, vertical addition and we have shown you what resulted because of appending. In the next class we will be happy to show you this particular lecture onwards that is on merging. So, merging we will continue for the next class. Thank you very much.