Exploring Survey Data on Health Care

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

Lecture - 21
Basic Understanding of SPSS and Data Filtration in STATA

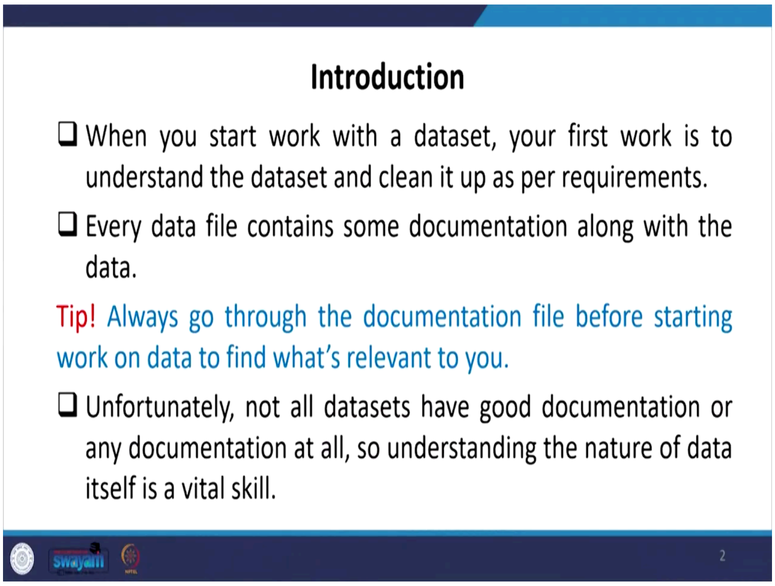
Welcome participants for this NPTEL MOOC program on handling health care data. We are on the 5th week. So far, we have covered the basics of data, availability of data, basics of handling STATA, also we have introduced about how to deal with health care data specifically. In this week, we have primarily targeted to understand how the data can be filtered, can be filtered or shaped off, can be structured, can be organized.

In the process we will also discuss extraction and merging. Therefore, we have kept the title as Data Filtration using STATA and SPSS. Myself Pratap Chandra Mohanty, I am presently associated with IIT, Roorkee as a faculty member in the Department of Humanities and Social Sciences, I teach Economics. This module I have been developing for over long time and this is assisted by our team Mr. Milind and Mr. Kamal, related to this I have been teaching research methodology over last 6 years.

Therefore, you know any sort of difficulties you might be facing, I will be most happy to address it. Now, regarding survey data analysis, this is the name of the week within that we have kept certain aspects like how to process the data, how to filter the data, how to extract the data. And we will also introduce about STATA operation, though on the last week we discuss about basic understanding of STATA.

This week we are going to extract data in STATA, also on the way we will be introducing SPSS package as well. Though we are not exclusively targeting SPSS, but some of you might have been comfortable with SPSS, I will guide you the very basic direction of it and rest I am 100 percent sure you can take off very correctly. So far as data filtration is concerned, let me just go by giving you an introduction to it like where you start work with a data set. Your first work is to understand the data set and clean it off as per requirement.

(Refer Slide Time: 03:06)



Introduction

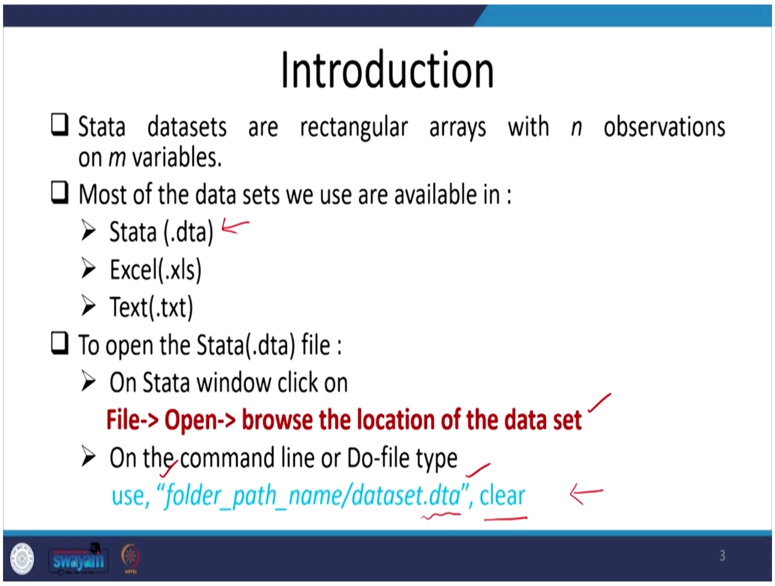
- ❑ When you start work with a dataset, your first work is to understand the dataset and clean it up as per requirements.
- ❑ Every data file contains some documentation along with the data.

Tip! Always go through the documentation file before starting work on data to find what's relevant to you.

- ❑ Unfortunately, not all datasets have good documentation or any documentation at all, so understanding the nature of data itself is a vital skill.

Every data file contains some documentation along with the data. One of the tips we have provided here that is, always go through the documentation file before starting to work on a database, there you are supposed to find out what is in fact relevant for you to use. Unfortunately, not all data sets have good documentation or any documentation at all. So, understanding the nature of the data itself is a vital skill.

(Refer Slide Time: 03:35)



Introduction

- ❑ Stata datasets are rectangular arrays with *n* observations on *m* variables.
- ❑ Most of the data sets we use are available in :
  - ➢ Stata (.dta)
  - ➢ Excel(.xls)
  - ➢ Text(.txt)
- ❑ To open the Stata(.dta) file :
  - ➢ On Stata window click on **File-> Open-> browse the location of the data set**
  - ➢ On the command line or Do-file type use, *"folder_path_name/dataset.dta"*, clear

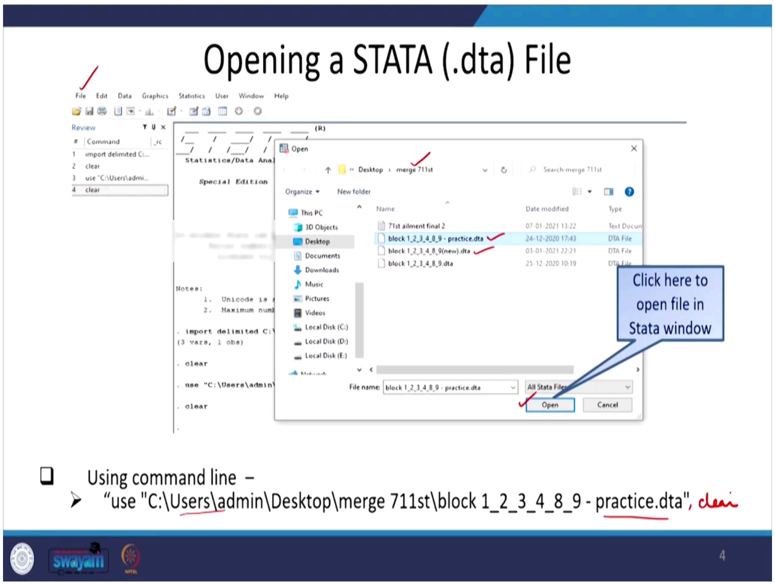
Stata datasets are rectangular arrays with you know 'n' observations on 'm' variables. So, n by m you know matrix is formed in its data set. Most of the data sets are used and data sets we are going to use are available either in Stata format or excel format or text format or also in a SPSS format as well, .sav format as well. To open Stata, the Stata file is in ‘.dta’, ‘.dta’ if that format you have just seen the extension is .dta at the end of the file.

You will see ‘.dta’ as the extension that means, you should make sure that the data is fitted in Stata format. On Stata window simply click on the data like we have already shown to you, you go to file on the very left top menu then clear click on open. Once you have opened the box then it will redirect you to browse the location where you have kept your data in ‘.dta’ format.

Once you have done this process, you can open your data and that is for your use. Otherwise in the command box on the Stata, I am not going to show it through the Stata window, because we were already guided earlier, this is simply a repetition. Here you simply write in the command ‘use,’, then is folder name folder name should be inverted this should be double inverted comma.

So, the path name should have attached with ‘.dta’ and at the end there must be a comma then clear. That is basically clearing the earlier stories on the memory of Stata and it will give you the space for the new data you wanted to load. So, here once again we are guiding how to open it the snapshot of the Stata we are just keeping it for you.

(Refer Slide Time: 06:05)



Opening a STATA (.dta) File

#	Command	_rc
1	import delimited C:...	
2	clear	
3	use "C:\Users\admi...	
4	clear	

```
Statistics/Data Ana
Special Edition

Notes:
      1.  Unicode is
      2.  Maximum num

. import delimited C:\
(3 vars, 1 obs)

. clear

. use "C:\Users\admin\

. clear

.
```

Name	Date modified	Type
71st ailment final 2	07-01-2021 13:22	Text Docum
block 1_2_3_4_8_9 - practice.dta	24-12-2020 17:43	DTA File
block 1_2_3_4_8_9(new).dta	03-01-2021 22:21	DTA File
block 1_2_3_4_8_9.dta	25-12-2020 10:19	DTA File

File name: block 1_2_3_4_8_9 - practice.dta

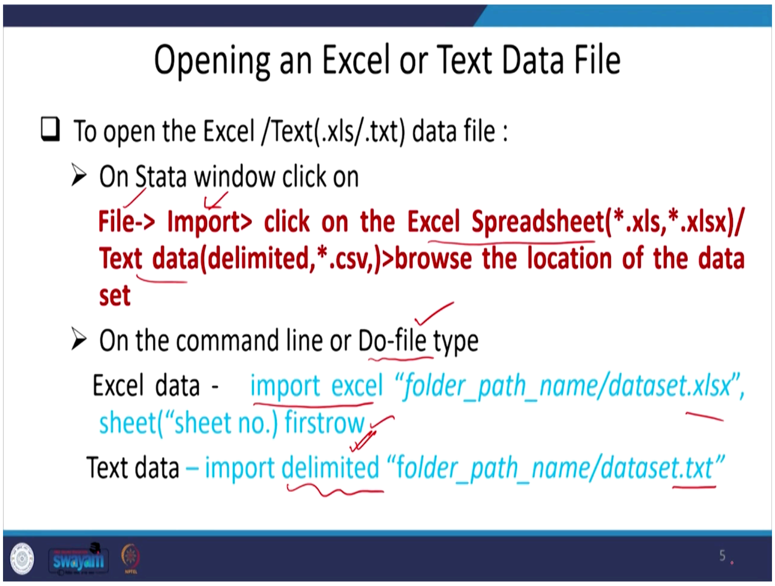
All Stata Files

Click here to open file in Stata window

- ❑ Using command line –
  - “use "C:\Users\admin\Desktop\merge 711st\block 1_2_3_4_8_9 - practice.dta", clear

It is here, like you have gone through file, then open window. File, then open, once you open it will ask you where your data is saved. We have saved it with our file here, you can open it accordingly then you just click open your concerned data. Once this is what you are working, it will be opened on your screen. Else since we kept it in this folder, we can just put these command, at the end you can add the one thing that is comma clear you can audit that will be giving you the right direction.

(Refer Slide Time: 06:48)



Opening an Excel or Text Data File

❑ To open the Excel /Text(.xls/.txt) data file :

➢ On Stata window click on

File-> Import> click on the Excel Spreadsheet(*.xls,*.xlsx)/ Text data(delimited,*.csv,)>browse the location of the data set

➢ On the command line or Do-file type

Excel data - import excel *"folder_path_name/dataset.xlsx"*, sheet("sheet no.) firstrow

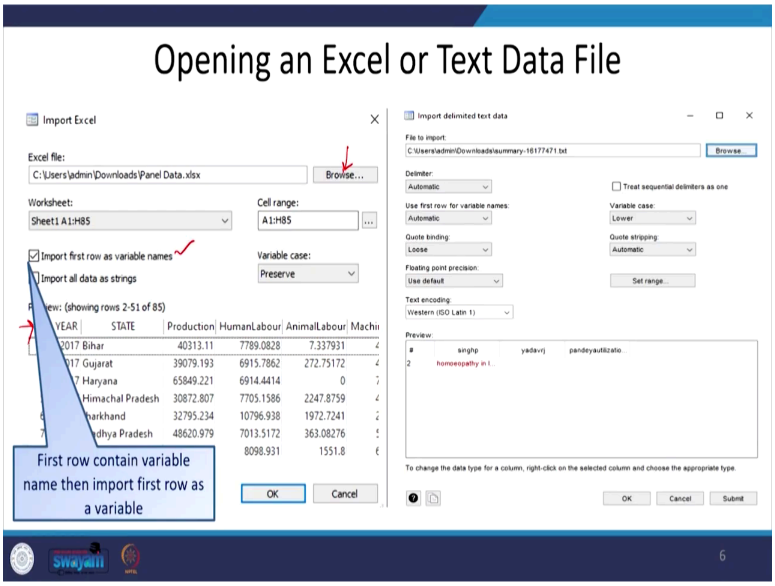
Text data – import delimited *"folder_path_name/dataset.txt"*

If you simply want to open the data through your excel file, you just wanted to open your excel file ok, that is '.xls' or '.txt' data on Stata windows; simply click on file, then import, then file, then import is there, import is once you on the top left file is there then import then click on the excel spread sheet. If you click this or that text data that is delimited text data; regarding delimited and all we will also guide you on our next lectures.

About what do you mean by delimited data, what do you mean by free forma data, what do you mean by fixed format data, we will discuss about it. Then, you browse the location of the data and click on this you will open it, on the command line or you can also open a Do-file, this is what we have already guided in the last week.

Now, on the command import should be written; import then excel, then how you wanted to fix it or keep the data you can specify it. So, based on that it will open the data. Next, if it is text data, it is a delimited data, what do you mean by delimited data we will discuss about it in our next lectures, then accordingly with that '.txt' data, you will find out the data on Stata.

(Refer Slide Time: 08:28)



Opening an Excel or Text Data File

Import Excel

Excel file: C:\Users\admin\Downloads\Panel Data.xlsx

Worksheet: Sheet1 A1:H85

Cell range: A1:H85

☑ Import first row as variable names

□ Import all data as strings

Variable case: Preserve

P[illegible]iew: (showing rows 2-51 of 85)

	YEAR	STATE	Production	HumanLabour	AnimalLabour	Machi
	2017	Bihar	40313.11	7789.0828	7.337931	4
	017	Gujarat	39079.193	6915.7862	272.75172	4
	7	Haryana	65849.221	6914.4414	0	7
		Himachal Pradesh	30872.807	7705.1586	2247.8759	4
6		harkhand	32795.234	10796.938	1972.7241	2
7		dhya Pradesh	48620.979	7013.5172	363.08276	5
				8098.931	1551.8	6

First row contain variable name then import first row as a variable

Import delimited text data

File to import: C:\Users\admin\Downloads\summary-16177471.txt

Delimiter: Automatic

□ Treat sequential delimiters as one

Use first row for variable names: Automatic

Variable case: Lower

Quote binding: Loose

Quote stripping: Automatic

Floating point precision: Use default

Text encoding: Western (ISO Latin 1)

Preview:

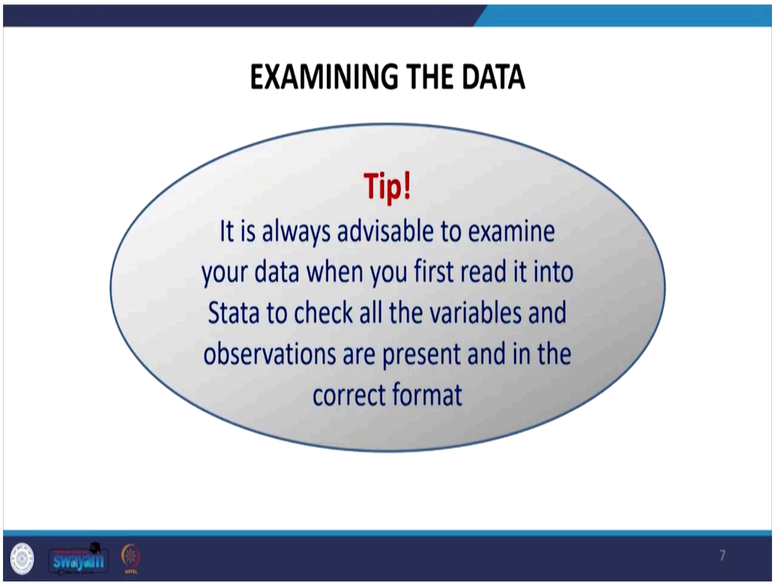
#	singhp	yadavrj	pandeyautilizatio...
2	homoeopathy in I...		

To change the data type for a column, right-click on the selected column and choose the appropriate type.

How to open through the click-based approach, here are the snapshot of in the window, like in the once you open import excel. So, excel file you have to browse that, here you need to browse, and you need to load where is your data. So, then the cell range worksheet etc. can also be defined, but here import first row has variable names that has to be specified. First row contains variable name then import first row as a variable.

So, that has to be specified, you have to put a tick mark then it will specify like this. Then regarding delimited text data, text data you have to again come up with this delimited automatic you can just select and wherever your delimited data is available you can also click on and submit it. These are the way.

(Refer Slide Time: 09:36)



EXAMINING THE DATA

Tip!

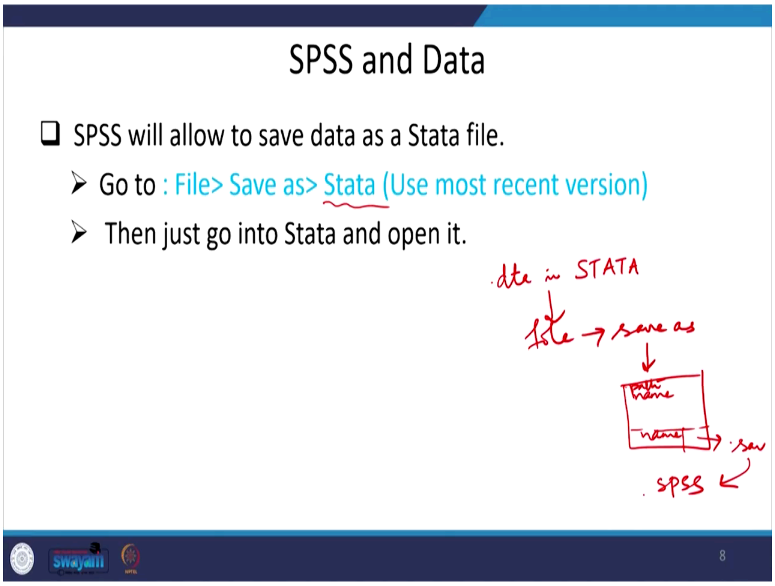
It is always advisable to examine your data when you first read it into Stata to check all the variables and observations are present and in the correct format

How to examine the data it is always advisable to examine your data, when you first read it into Stata, to check all the variables and observations are present and in the correct format. SPSS and in data we are just giving certain overview at this moment.

After maybe 20 minutes, I am going to again discuss about SPSS in detail. SPSS will allow you to save data as a Stata file as well. Nowadays, you have inbuilt SPSS to Stata transfer and Stata to SPSS transfer data available. Earlier we used to some install a patch that could help us to convert a SPSS file to Stata and Stata to SPSS.

Nowadays, these are all coming. So, once it is in a SPSS file, if you have if your data in SPSS file you want to work with Stata you can simply click on save as file, then save as.

(Refer Slide Time: 10:52)



SPSS and Data

❑ SPSS will allow to save data as a Stata file.

- Go to : File> Save as> Stata (Use most recent version)
- Then just go into Stata and open it.

.dte in STATA
file → save as
path name
name → .sav
.spss
8

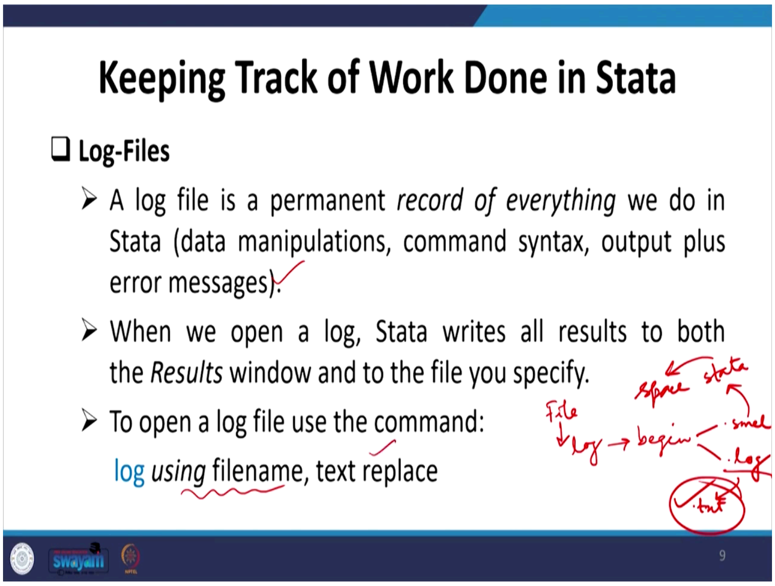
Then options will given to you at the bottom so whether it is in .dta or in .sav format, .xls format or anything you can submit it. And then your data will be saved in dta version, another one from the Stata to SPSS is also you can do it.

Once you have opened the data .dta file in Stata, then you simply go to file, then save as then you simply open. Once a box will be opened and it will ask you like in which format you need to save. '.sav' format you should have to open. Once you open it, your data is now stored with the help of by giving a proper name here, you should give a correct name or name should be here.

Then its path name, path name should be given correctly. So, with that you can be able to store your data in SPSS format. If you are comfortable in SPSS you can also operate through SPSS as well. So, let us review our previous lectures, if anything was missing in the previous lecture you can now make a track, like log files how to have the log file a log what do you mean by log file I have already told you in the previous week as well.

Whenever you open your Stata to start operating the Stata or start analyzing the data it is always suggested that you should open your log file.

(Refer Slide Time: 12:45)



Keeping Track of Work Done in Stata

❑ **Log-Files**

- A log file is a permanent *record of everything* we do in Stata (data manipulations, command syntax, output plus error messages).
- When we open a log, Stata writes all results to both the *Results* window and to the file you specify.
- To open a log file use the command:

  log *using* filename, text replace

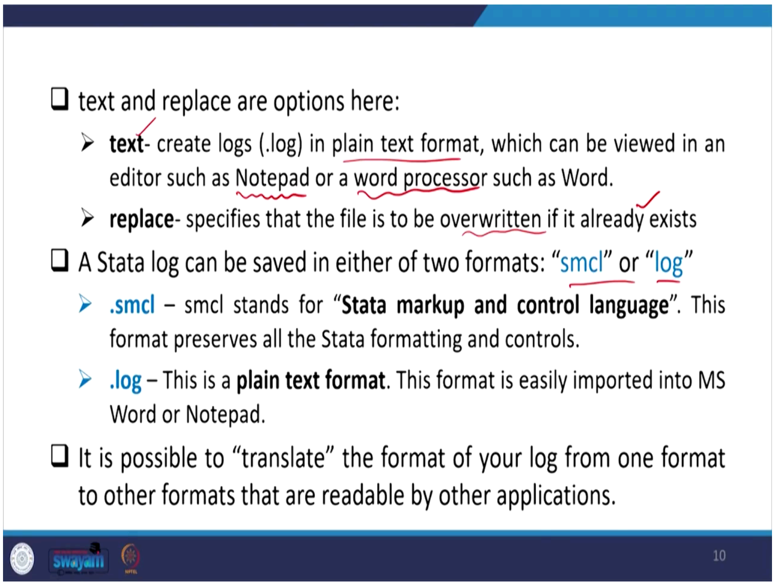
space
stata
file
log → begin
.smcl
.log
.txt
9

Then it is available on the top-down menu file, then file then you will see log, log begin. If you click on begin it gives you two options then, .smcl or .log. So, we are now suggesting you to go for .log, .log will save the data in .txt format, like your notepad file, this one is going to save you in Stata format.

But mostly it is consuming more space. So, it is suggested to go for the txt format, log file is a permanent record of everything we do in Stata. Like you know once you saved start working with .txt or .log, in future or in after certain time if you get busy at that moment you can open this data or are continuation of your output, after some time as well. So, some result you can easily find it out through this log file.

So, a log file is in fact, a permanent record of everything we do in Stata, that is data manipulation, command syntax output plus error messages etc., all details are actually visible. When you open a log Stata rights, all results to both the results window and to the file we have specified. To open a log file, we need to use the command is log using a file names you will you may give this command, then text replace it will actually if it is in text then it will actually save it in text format.

(Refer Slide Time: 14:41)



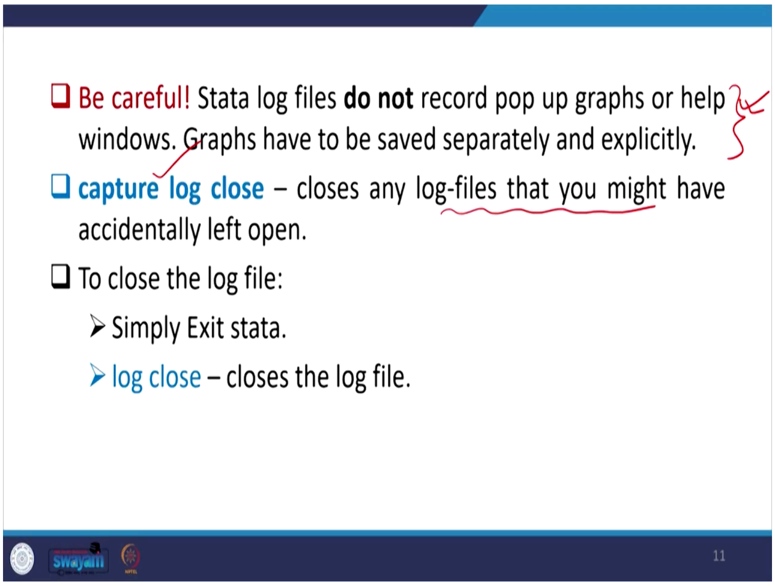
- ❑ text and replace are options here:
  - ➢ **text**- create logs (.log) in plain text format, which can be viewed in an editor such as Notepad or a word processor such as Word.
  - ➢ **replace**- specifies that the file is to be overwritten if it already exists
- ❑ A Stata log can be saved in either of two formats: "smcl" or "log"
  - ➢ .smcl – smcl stands for "**Stata markup and control language**". This format preserves all the Stata formatting and controls.
  - ➢ .log – This is a **plain text format**. This format is easily imported into MS Word or Notepad.
- ❑ It is possible to "translate" the format of your log from one format to other formats that are readable by other applications.

So, text replace is important to create the .log file, this will save in plain text as I already said. So, this will be saved and can be edited later, since it will be saved in notepad file or word processor, replace command basically is important because it is going to replace the earlier log file. If you do not do it, then it will actually overwrite it on the existing log file.

So, better to retake the replace command. If you do not do the replace command it is going to overwrite on the same log file, you might be confused that what sort of file, what sort of output I operated today. So, if it is mixed with the previous log file, then you would be highly confused. So, a Stata log can be saved in either of two format, that is as I told you .smcl file or .log, .smcl stands for Stata markup and control language, this format preserves all the Stata formatting and in control.

.log is in fact, a plain text format, this format is easily imported into MS word or notepad. It is possible to translate, the format of your log from one format to other format that are readable by other applications as well, that is also one of the interesting aspects.

(Refer Slide Time: 16:16)



- ❑ Be careful! Stata log files **do not** record pop up graphs or help windows. Graphs have to be saved separately and explicitly.
- ❑ **capture log close** – closes any log-files that you might have accidentally left open.
- ❑ To close the log file:
  - ➢ Simply Exit stata.
  - ➢ log close – closes the log file.

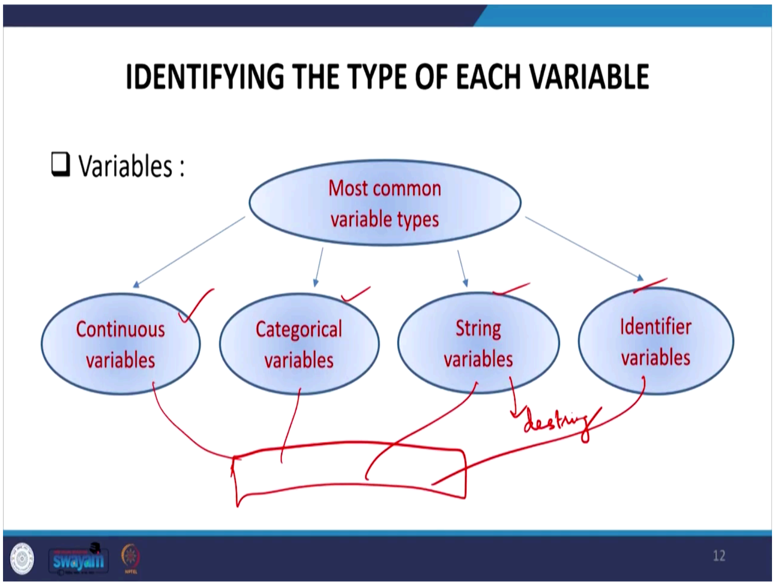
So, we need to be very careful about log file because of the fact that these do not record pop up graphs, that this is quite important. Like suppose we have given 'br' command, (browse command that is not going to be displayed on the same window.

It is going to create another, it gives another window and suppose we take help command, help. Help command is going to be opened in another pop-up bin window or suppose we wanted to do some graphs, it is opened in another pop ups window. So, that window is not saved in log file.

But if you give C help command, c help is actually going to give you the output on the same window and that is going to be saved in log file. So, graphs if any you have derived through the log, through your operation through your work with Stata, you have to save those graphs separately. You need to be careful about this aspect. Capture log close has to be made closes any log file that you might have accidentally left open, capturing log close is important.

To close the log file if you have you might have accidentally left open, you can give the command capture log close, it will actually capture which one was actually open. But to close log file it is also always important, this in fact, once you give a log close it will close it and at the next time you can open it with that particular file.

(Refer Slide Time: 18:09)



IDENTIFYING THE TYPE OF EACH VARIABLE

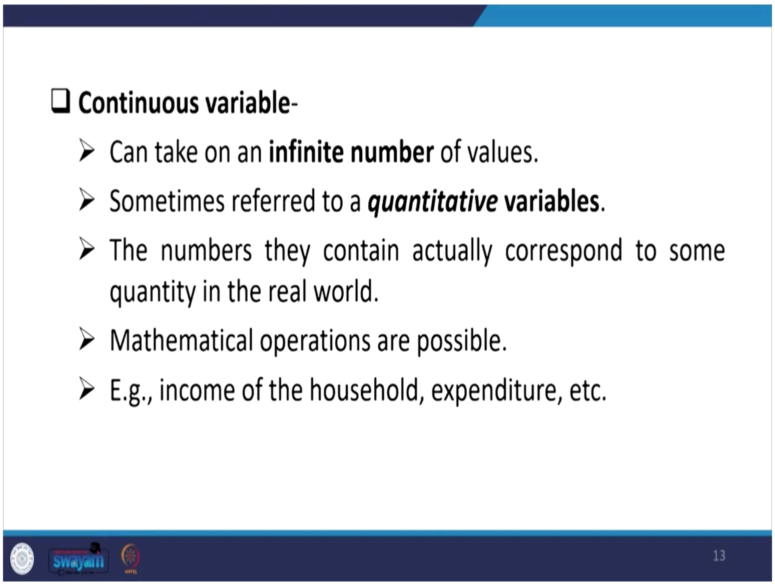
❑ Variables :

So, identifying the type of its variable those are going to be useful in our command, in our Stata operation. Most common variables that we are going to encounter, going to face is of either continuous variable, either categorical variables string variables or identifier variables.

Regarding string and all those variables and its meaning I have already discussed earlier, but here we are going to use it, how we are going to use it? We are going to use it in our analysis. How are we going to use it in analysis? Like, if the data is continuous then what model we are supposed to give.

If the data is categorical, then what econometrics model, statistical model we are supposed to apply. If the data is in string then what should I do, should I destring the data or not, those things we are guiding here. Or if it is an identifier variable then how it is going to be helpful for merging, for one database with another database. Those four things we are going to use most often now onwards.

(Refer Slide Time: 19:25)



❑ **Continuous variable-**

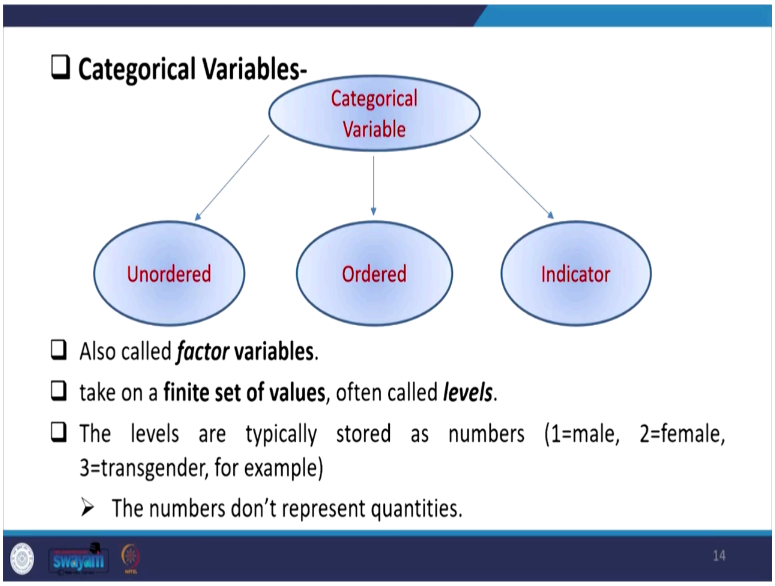
- Can take on an **infinite number** of values.
- Sometimes referred to a ***quantitative* variables**.
- The numbers they contain actually correspond to some quantity in the real world.
- Mathematical operations are possible.
- E.g., income of the household, expenditure, etc.

If it is continuous variable as we have already defined like ratio data, we have set out of four data. Usually, we find continuous characters of the ratio data, if it is continuous this is in fact an infinite number of values possible. These are sometimes referred to as quantitative variables, the numbers they contain actually correspond to some quantity in real world. So, mathematical operations are possible in this data.

And like income of the household expenditure, age of the person data the weight of the person, height, etc. are all called quantitative variables. Qualitative variables we have already clarified, but still since we are going to use further, we are clarifying once again. Categorical could be your ordered variable, could be unordered variable, could be indicator variable.

Unordered variable though it is categorical, like male and female though that is not called order. Order variable like education of the person, standard of education of the person that is clearly an order variable.

(Refer Slide Time: 20:43)



❑ Categorical Variables-

Categorical Variable
Unordered
Ordered
Indicator

❑ Also called ***factor* variables**.

❑ take on a **finite set of values**, often called ***levels***.

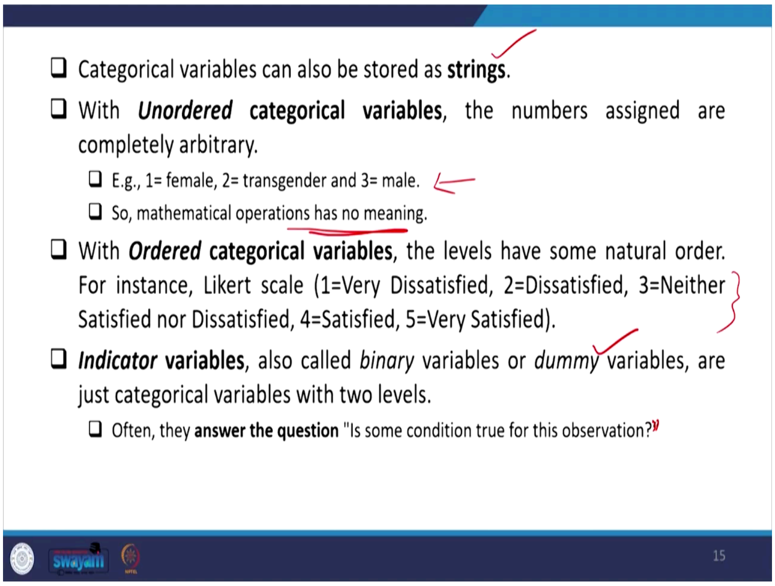
❑ The levels are typically stored as numbers (1=male, 2=female, 3=transgender, for example)

➢ The numbers don't represent quantities.

14

So, here these are also called when we say factor variable, these are also categorical variable, we also say those as factor variables. They could be finite set of values and often, since they are of having finite set of values, they are often called levels. The levels are typically stored as numbers, with 1 may be male, 2 may be female, 3 may be transgender or vice versa. The numbers actually do not represent quantities, rather it only give certain levels, certain identification.

(Refer Slide Time: 21:27)



❑ Categorical variables can also be stored as **strings**.

❑ With ***Unordered* categorical variables**, the numbers assigned are completely arbitrary.

❑ E.g., 1= female, 2= transgender and 3= male.

❑ So, mathematical operations has no meaning.

❑ With ***Ordered* categorical variables**, the levels have some natural order. For instance, Likert scale (1=Very Dissatisfied, 2=Dissatisfied, 3=Neither Satisfied nor Dissatisfied, 4=Satisfied, 5=Very Satisfied).

❑ ***Indicator* variables**, also called *binary* variables or *dummy* variables, are just categorical variables with two levels.

❑ Often, they **answer the question** "Is some condition true for this observation?"

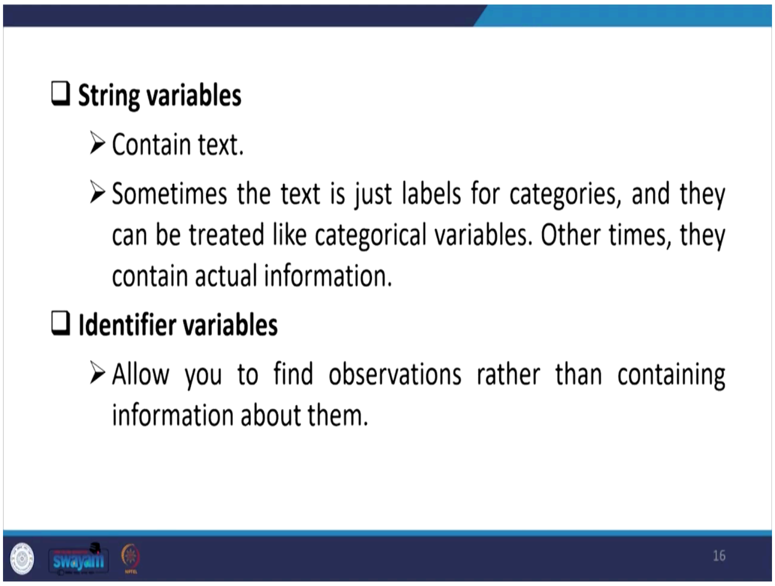
15

Categorical variables can also be stored as strings, can be stored as strings with unordered categorical variables, the numbers assigned are completely arbitrary. So, this is what we have already said, this could be completely arbitrary. If these are completely arbitrary then mathematical operation has no meaning, with ordered categorical variables the levels have some natural order.

So, like Likert scale-based identification of certain indicators, have certain magnitude of their values. Indicator variables like it is only a binary variable 0 and 1 or have certain dummies, one does not have any value, rather it is dummied by another one, it is explained by another category.

So, the often we ask a question, like we answer the question like each some condition true for this observation, what will do our answer could be yes or no ok. So, yes and no is actually a kind of dummy, it is simply indicating certain indicators, it is not having any magnitude.

(Refer Slide Time: 22:55)



❑ **String variables**

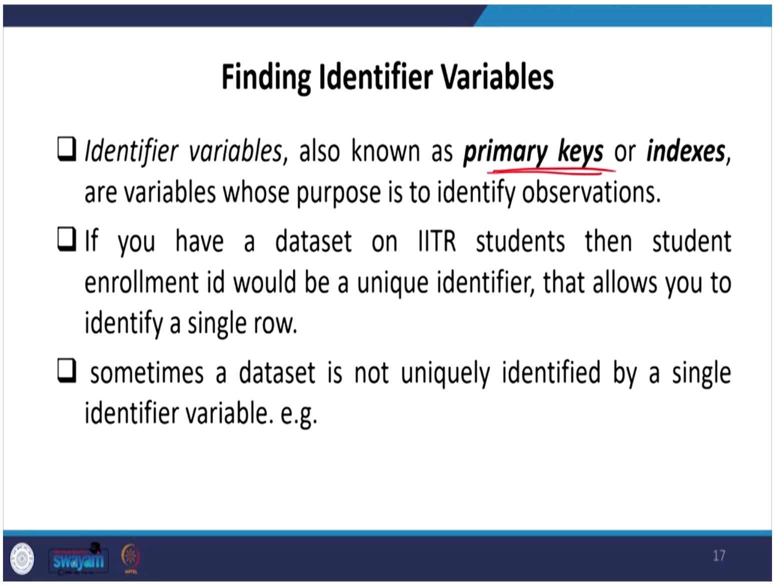
- Contain text.
- Sometimes the text is just labels for categories, and they can be treated like categorical variables. Other times, they contain actual information.

❑ **Identifier variables**

- Allow you to find observations rather than containing information about them.

String variables we have already said it may contain mix of things characters, with numbers, some identifiers, it’s like they can be treated like categorical variables as well, as we already said identifier variables I have already explained. This simply allows you to find observation rather than containing information about it.

(Refer Slide Time: 23:20)



Finding Identifier Variables

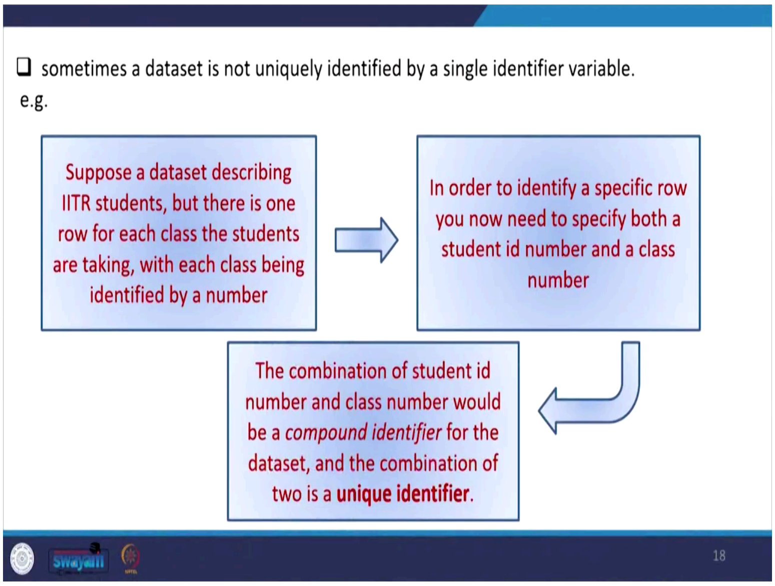
- ❑ *Identifier variables*, also known as ***primary keys*** or ***indexes***, are variables whose purpose is to identify observations.
- ❑ If you have a dataset on IITR students then student enrollment id would be a unique identifier, that allows you to identify a single row.
- ❑ sometimes a dataset is not uniquely identified by a single identifier variable. e.g.

What do you mean by identifier variables? Again, identifier variables maybe our as I already mentioned earlier maybe our primary keys or indexes. So, primary keys we will discuss about in detail at the time of extraction in our next lecture, next to next lecture as well.

And our merging where will be merging some of the databases we do require a primary keys and identifier. If you have a database on IITR students, then students' enrolment ID would be a unique identifier that allows you to identify a single particular row.

Sometimes a database is not uniquely identified by a single identifier variable as well. If a database is not uniquely identified by a single identifier variable, we have to mix multiple identifier; may multiple we need to compound we need to know, mix multiple information to define an identifier variables. You can just read it between the line, and I am sure you will understand.

(Refer Slide Time: 24:27)



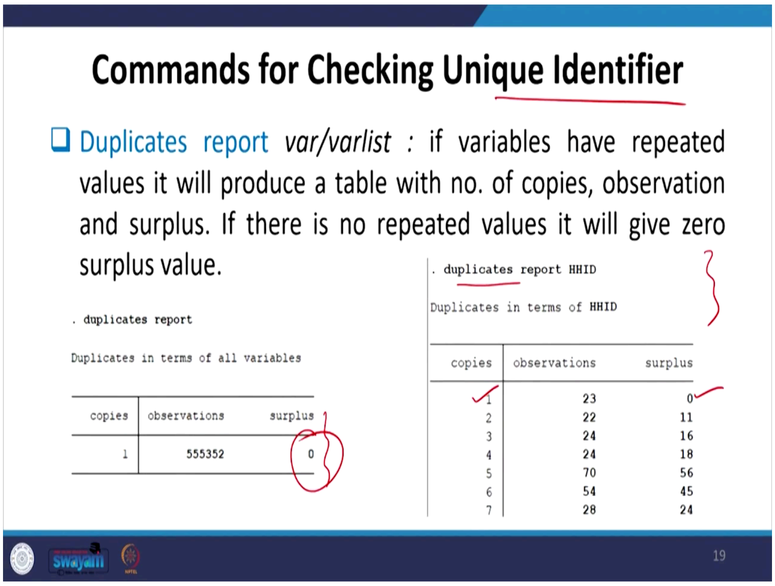
❑ sometimes a dataset is not uniquely identified by a single identifier variable.
e.g.

Suppose a dataset describing IITR students, but there is one row for each class the students are taking, with each class being identified by a number

In order to identify a specific row you now need to specify both a student id number and a class number

The combination of student id number and class number would be a *compound identifier* for the dataset, and the combination of two is a **unique identifier**.

(Refer Slide Time: 24:29)



Commands for Checking Unique Identifier

❑ Duplicates report *var/varlist* : if variables have repeated values it will produce a table with no. of copies, observation and surplus. If there is no repeated values it will give zero surplus value.

```
. duplicates report

Duplicates in terms of all variables

--------------------------------------
   copies | observations       surplus
----------+---------------------------
        1 |       555352             0
--------------------------------------
```

```
. duplicates report HHID

Duplicates in terms of HHID

--------------------------------------
   copies | observations       surplus
----------+---------------------------
        1 |           23             0
        2 |           22            11
        3 |           24            16
        4 |           24            18
        5 |           70            56
        6 |           54            45
        7 |           28            24
```

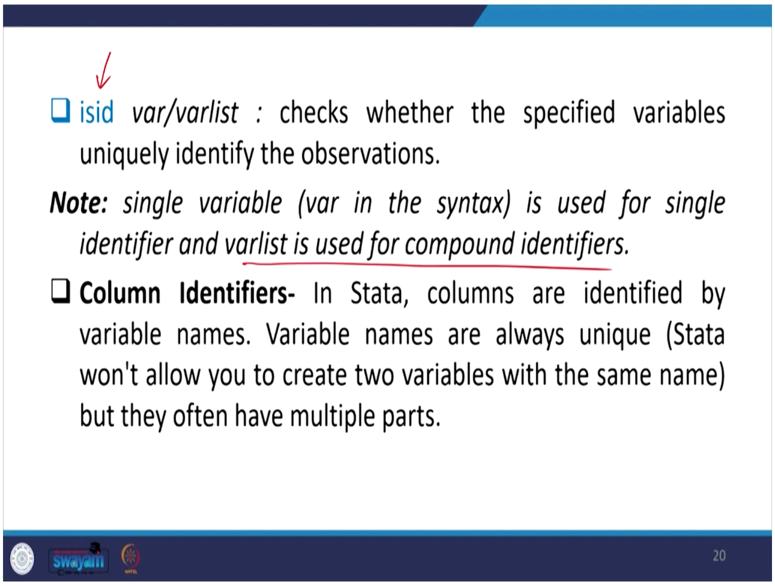
Coming to the commands for checking unique identifier, we are going to use it in detail in our extraction, I am just going to give you the basic information at this moment like suppose, we just wanted to understand the duplicate reports of variable or variable list. That is if variables have repeated values, it will produce a table with no number of copies, observation, and surplus.

If there is no repeated values it will give zero surplus value. Here like, with this duplicate report we can find out whether this variable has repetition in the database or not. From here duplicates in terms of all variables, if the surplus is coming out to be 0, if it is 0; that means, your database does not have any duplications.

But if we have observations, we can see that there surplus is 0; that means, these 23 observations have you know 0 surplus, no duplications, but in all other cases there are some duplications. So, there is a possibility of overlapping, there is a possibility of creating an identifier because of duplications of entries in each in that particular file.

So, duplication helps us to identify whether we can go for a unique identifier or not. This is the one we just derived with the help of HHID variable. So, we will experiment and let you know for sure at the time of our extraction and merging.

(Refer Slide Time: 26:19)



- ❑ isid *var/varlist* : checks whether the specified variables uniquely identify the observations.

Note: *single variable (var in the syntax) is used for single identifier and varlist is used for compound identifiers.*

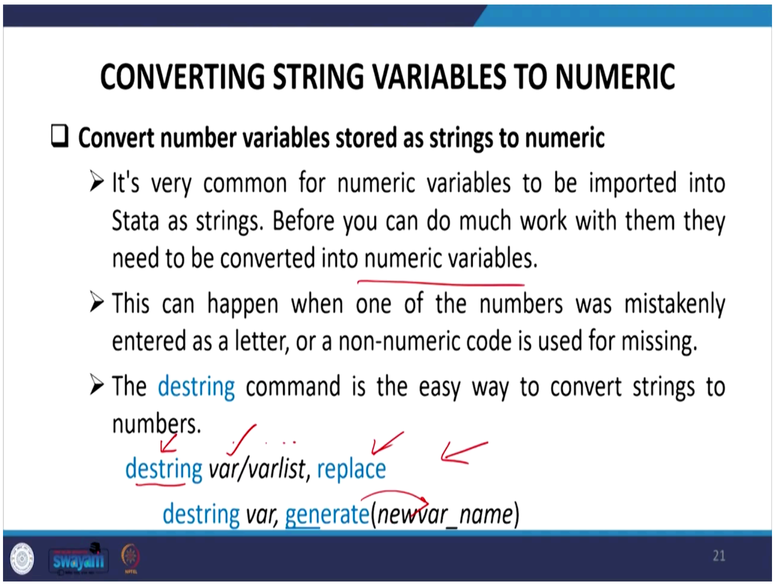
- ❑ **Column Identifiers-** In Stata, columns are identified by variable names. Variable names are always unique (Stata won't allow you to create two variables with the same name) but they often have multiple parts.

Similarly, will be using isid command as well, isid commands helps you to define whether your variable that you are thinking of making as unique or the identifier variable whether that is unique or not. Uniquely identifying the observation or not, the isid command will help you out, we are going to explain it for sure in our next lecture.

So, no we need to note here that single variable that is variable in the syntax is used for single identifier and variable list is used for compound identifiers. If there are more than one identifier then that is called varlist, otherwise it is only var as per our command. So, there might be column identifier as well. In Stata, columns are identified by variable names.

Variable names are always unique, Stata do not allow you to create two variables with the same name. But they often have multiple parts.

(Refer Slide Time: 27:29)



CONVERTING STRING VARIABLES TO NUMERIC

- ❑ **Convert number variables stored as strings to numeric**
  - ➢ It's very common for numeric variables to be imported into Stata as strings. Before you can do much work with them they need to be converted into numeric variables.
  - ➢ This can happen when one of the numbers was mistakenly entered as a letter, or a non-numeric code is used for missing.
  - ➢ The destring command is the easy way to convert strings to numbers.

```
destring var/varlist, replace
    destring var, generate(newvar_name)
```

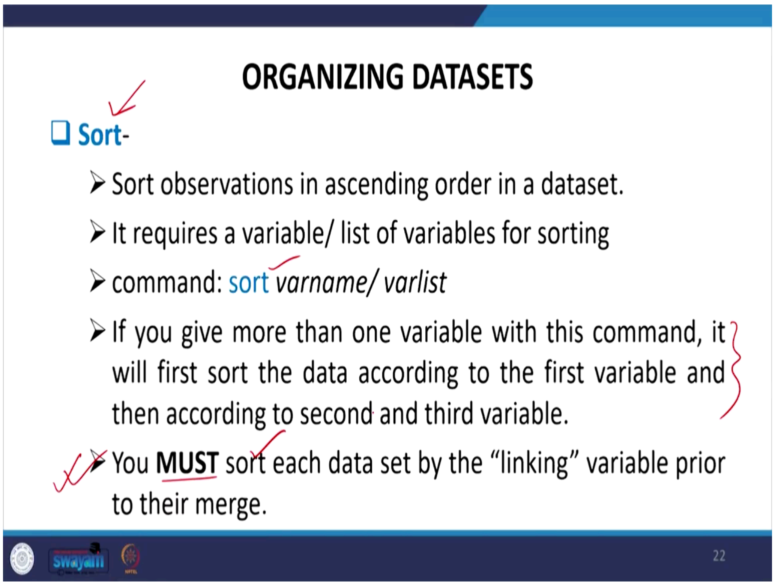
So, they have multiple parts, but they do not carry with the same name. Now converting string variables in a numeric variable, we have already said earlier we are again saying. And we will also be repeating these discussion in our next lecture as well. It is very common for numeric variables to be imported into data string.

Before we can do much work with them, they need to be converted into numeric variables. Then only most of the operations or the mathematical operations analysis could be done. This can happen when one of the numbers was mistakenly entered as a letter and or a non-numeric code is used for missing. The destring command is the easy way to convert string to numbers, and like this way you can do it, you simply write this command destring, then the variable or variable list.

If there are more than one variable, which are common identifier to connect to different block of information, in that case you have to include variable 1, variable 2, variable 3, then you need to destring it. And even like in this case we are not defining the common identifier, we are simply making them to numeric.

So, either you simply destring one variable at a go or you can also list of variables you can write it here with a space and replace it, we will destring and generate a variable name. But another approach is that it is simply replacing, another one is destring, destring command you can give it. At the end you can generate a variable, new name would be saved with its numeric values.

(Refer Slide Time: 29:23)



ORGANIZING DATASETS

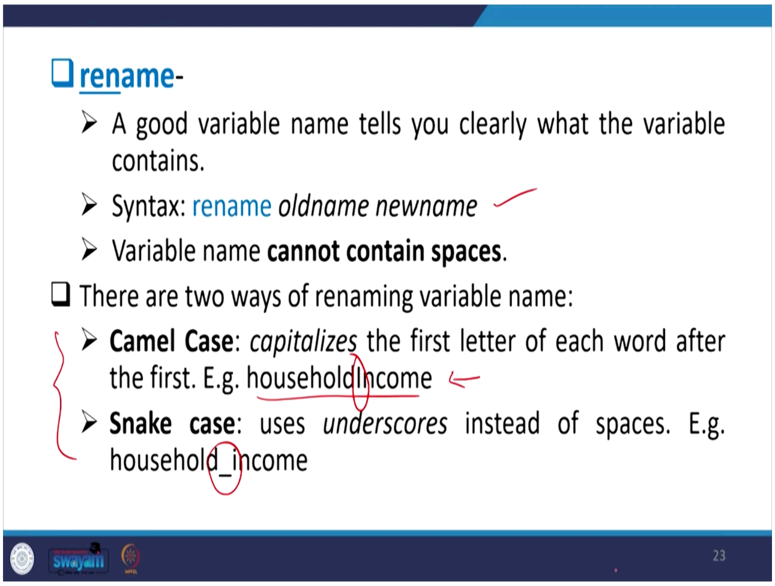
❑ Sort-

- Sort observations in ascending order in a dataset.
- It requires a variable/ list of variables for sorting
- command: sort *varname/ varlist*
- If you give more than one variable with this command, it will first sort the data according to the first variable and then according to second and third variable.
- You **MUST** sort each data set by the "linking" variable prior to their merge.

Then sort command, we will also operate it in our next lecture, at this moment I am just giving you the theoretical background about it. On the next lecture we will operate it and explain you in detail.

Sort likewise in the excel sheet we sort the data, here also we can if you sort the variable name or variable list we can able to sort the data, in ascending order or in descending order as per our requirement. If you give more than one variable with this command, it will first sort the data according to the first variable and then to the second and third. Likewise, we do it in excel page, you must sort each data set by the linking variable prior to their merge that is a must command.

(Refer Slide Time: 30:33)



❑ rename-

- A good variable name tells you clearly what the variable contains.
- Syntax: rename *oldname newname*
- Variable name **cannot contain spaces**.

❑ There are two ways of renaming variable name:

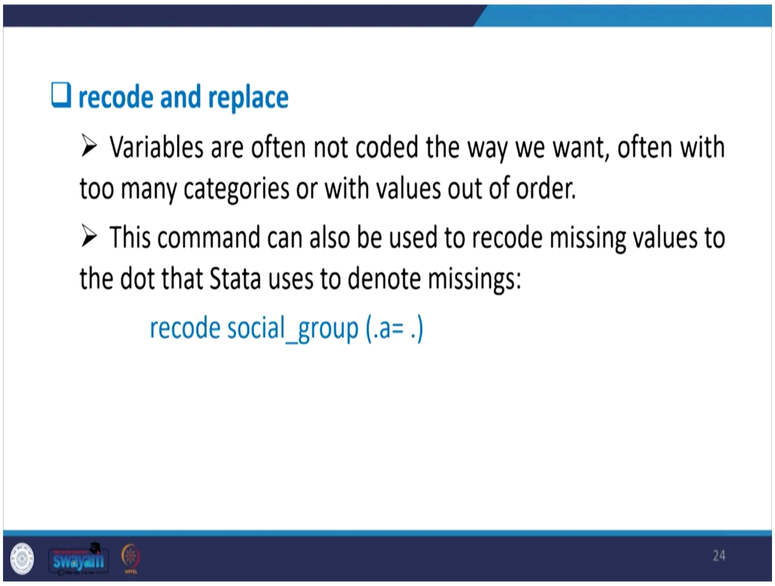
- **Camel Case**: *capitalizes* the first letter of each word after the first. E.g. householdIncome
- **Snake case**: uses *underscores* instead of spaces. E.g. household_income

Must we are saying before linking or merging the data, those identifier variable or the linking variable must have been sorted out, must be sorted sort. Must be run with the command sort before they need to be merged. So, similar rename command we have already said those basic details about Stata handlings, how to rename data, especially rename old name new name if you give it, you know it will save with the new name.

Likewise, this then variable name cannot contain spaces for sure, Stata does not read spaces. There are some names suggested how you can give the variable name, one approach called camel name, another is called snake-based name. Camel name basically likewise camel it has different shape, like different shape it has different peak point. So, similarly you can capitalize your name.

Next, if you like a household income you wanted to write it down, your first name next one you can make it as capital. So, that will be differentiating, and this kind of naming is called camel naming. Snake name is basically since it does not read spaces so you can give underscore, to differentiate the name underscore in case of snake case naming.

(Refer Slide Time: 31:48)



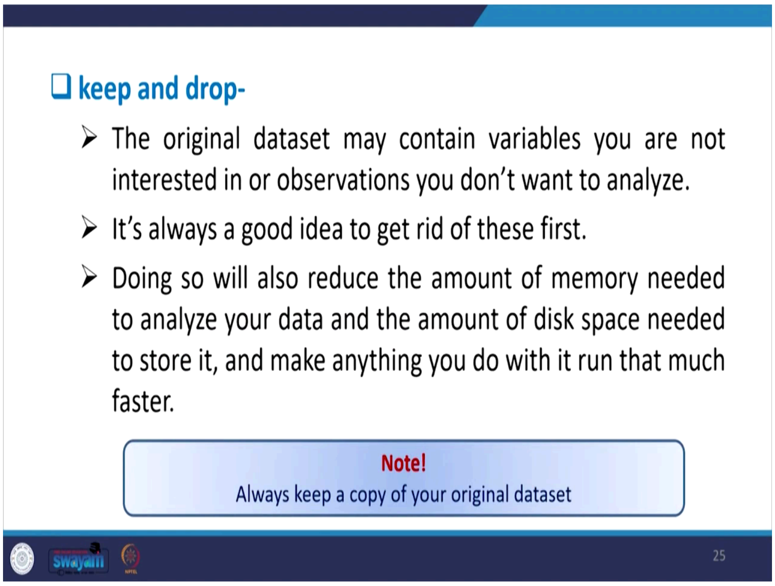
❑ recode and replace

- ➢ Variables are often not coded the way we want, often with too many categories or with values out of order.
- ➢ This command can also be used to recode missing values to the dot that Stata uses to denote missings:

```
recode social_group (.a= .)
```

Recode and replace I think I already guided you, I need not spend more time here simply recode command with the variable name. And within the bracket what you exactly do it, it we have already specified.

(Refer Slide Time: 32:02)



❑ keep and drop-

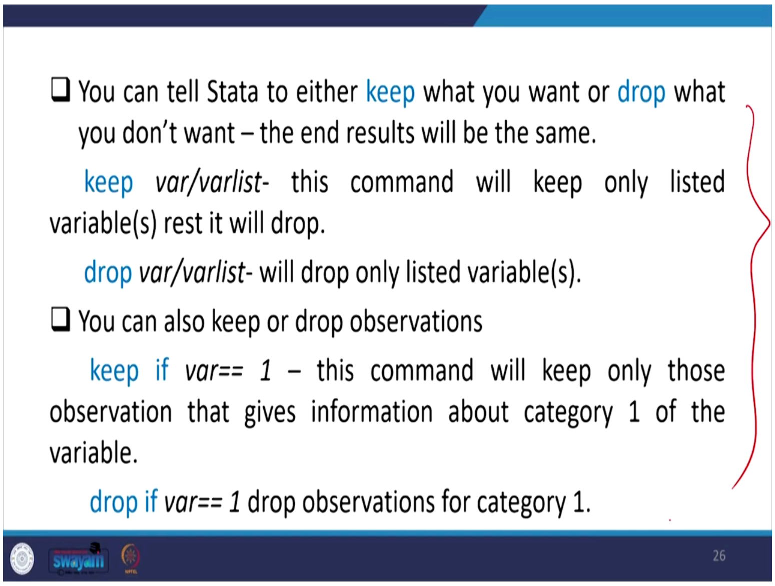
- ➢ The original dataset may contain variables you are not interested in or observations you don't want to analyze.
- ➢ It's always a good idea to get rid of these first.
- ➢ Doing so will also reduce the amount of memory needed to analyze your data and the amount of disk space needed to store it, and make anything you do with it run that much faster.

Note!
Always keep a copy of your original dataset

In this particular slide keep and drop also we have already guided you earlier, it says basically whether you are going to keep or you are going to drop some of the variables or the value also. If you wanted to drop, drop is important when very specific variables need to be dropped others to be kept so that is perfectly fine.

But in case of keep when there are so many variables you wanted to keep very less number of variables, rest you do not bother you do not mind, then you keep command is most useful. Always keep a copy of your original data set before operating this, because this is going to wipe out, with this is going to delete the others redundant variable other than your keep for drop command.

(Refer Slide Time: 32:47)



- ❑ You can tell Stata to either keep what you want or drop what you don't want – the end results will be the same.

  keep *var/varlist*- this command will keep only listed variable(s) rest it will drop.

  drop *var/varlist*- will drop only listed variable(s).

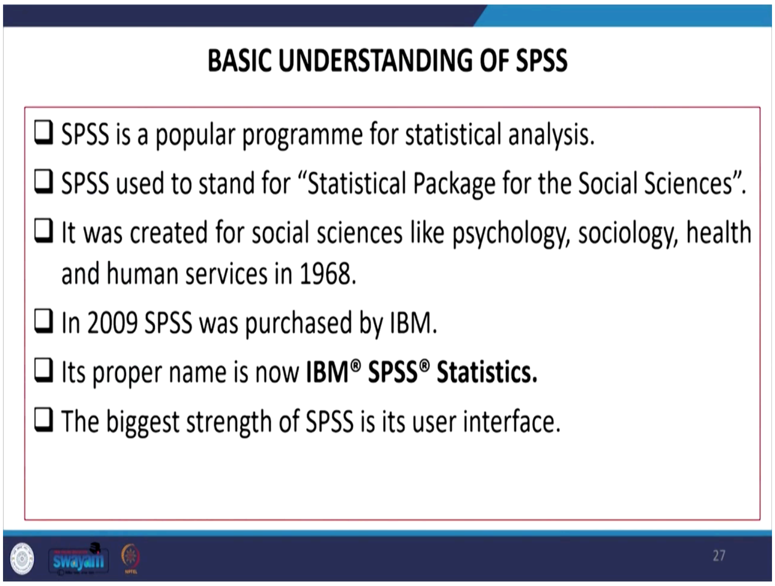
- ❑ You can also keep or drop observations

  keep if *var== 1* – this command will keep only those observation that gives information about category 1 of the variable.

  drop if *var== 1* drop observations for category 1.

So, I am not going to repeat guys you must have to read this, we have guided you earlier already I think it is time to move on.

(Refer Slide Time: 32:28)



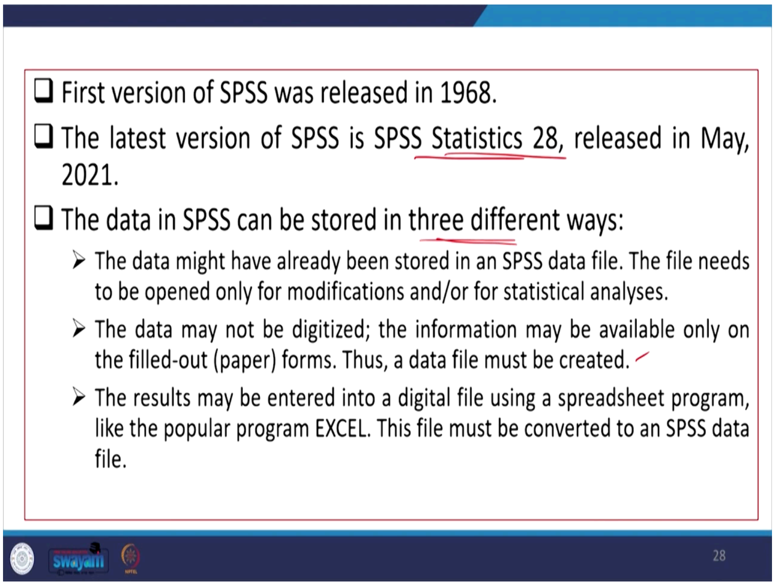
BASIC UNDERSTANDING OF SPSS

- ❑ SPSS is a popular programme for statistical analysis.
- ❑ SPSS used to stand for "Statistical Package for the Social Sciences".
- ❑ It was created for social sciences like psychology, sociology, health and human services in 1968.
- ❑ In 2009 SPSS was purchased by IBM.
- ❑ Its proper name is now **IBM® SPSS® Statistics.**
- ❑ The biggest strength of SPSS is its user interface.

Now, we are explaining about how we should go for understanding SPSS, basics of Stata, operating with Stata, some clarification on Stata which we missed in our earlier week. We did it today, on the next week also we are going to operate with its data set, with its practical data handling we will also do it. Now, I am clarifying what do you mean by SPSS and how you should go for it.

SPSS as we all know that it is very popular program and widely applied by social scientist, this is also called "Statistical Package for the Social Sciences" and it was created for the social sciences like psychology, sociology, health and health services, in 1968. In 2009 SPSS was purchased by IBM, its proper name is now IBM SPSS Statistics, the biggest strength of SPSS is its user interface.

(Refer Slide Time: 34:02)

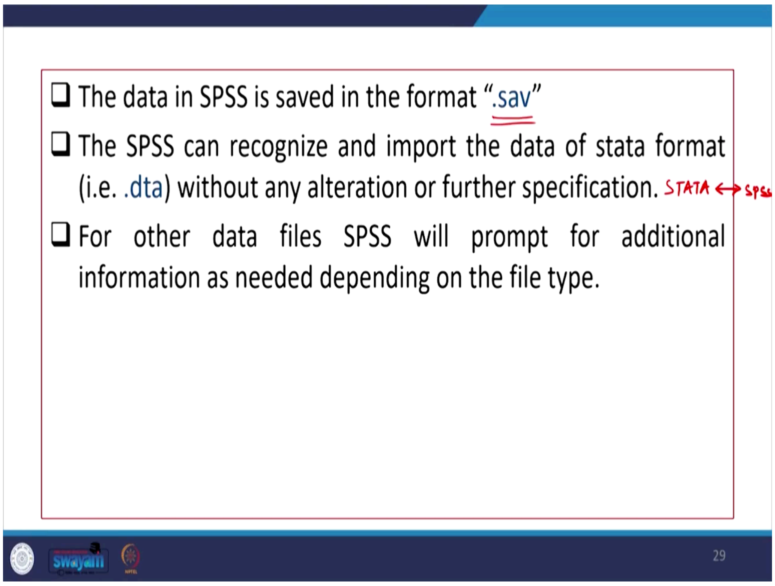


- ❑ First version of SPSS was released in 1968.
- ❑ The latest version of SPSS is SPSS Statistics 28, released in May, 2021.
- ❑ The data in SPSS can be stored in three different ways:
  - ➢ The data might have already been stored in an SPSS data file. The file needs to be opened only for modifications and/or for statistical analyses.
  - ➢ The data may not be digitized; the information may be available only on the filled-out (paper) forms. Thus, a data file must be created.
  - ➢ The results may be entered into a digital file using a spreadsheet program, like the popular program EXCEL. This file must be converted to an SPSS data file.

We have also made comparison between SPSS, Stata which one is better, Stata, SAS (Refer Time: 34:10) we have given a box earlier, you may also refer that from our very beginning of the lectures. The latest version of SPSS is available at this moment is 28th, release in May 2021. So, you may get it or the free license copy for some days you may get it with registration. The data in SPSS can be stored in three different ways, that is most important.

The data might have already been stored in SPSS data file the file needs to be opened only for modifications and for statistical analysis. The data may not be digitized, the information may be available on a filled out or paper form. Thus, a data must be created. This is another one, the third one is results may be entered into a digital file using a spread sheet program, like the popular program called excel, the file must be converted to SPSS. So, in three way we can use the SPSS data.

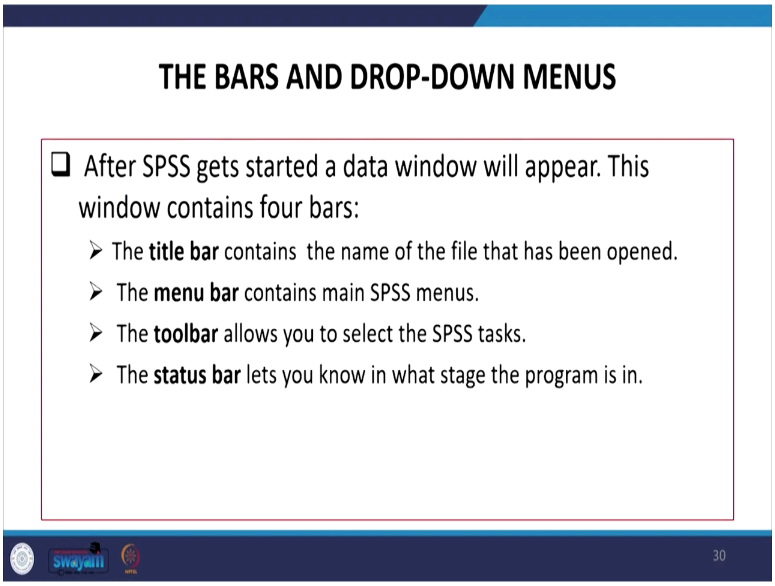
(Refer Slide Time: 35:21)



- ❑ The data in SPSS is saved in the format ".sav"
- ❑ The SPSS can recognize and import the data of stata format (i.e. .dta) without any alteration or further specification. STATA ↔ SPSS
- ❑ For other data files SPSS will prompt for additional information as needed depending on the file type.

Data in SPSS is saved in format called .sav which I already mentioned. The SPSS can recognize and import the data of Stata from the .dta file that is Stata file without any altercation or further specification. So, both ways Stata to SPSS or SPSS to Stata can be done right, without any mistake. For other I discussed just couple of minutes back in this particular lecture about this.

(Refer Slide Time: 36:06)



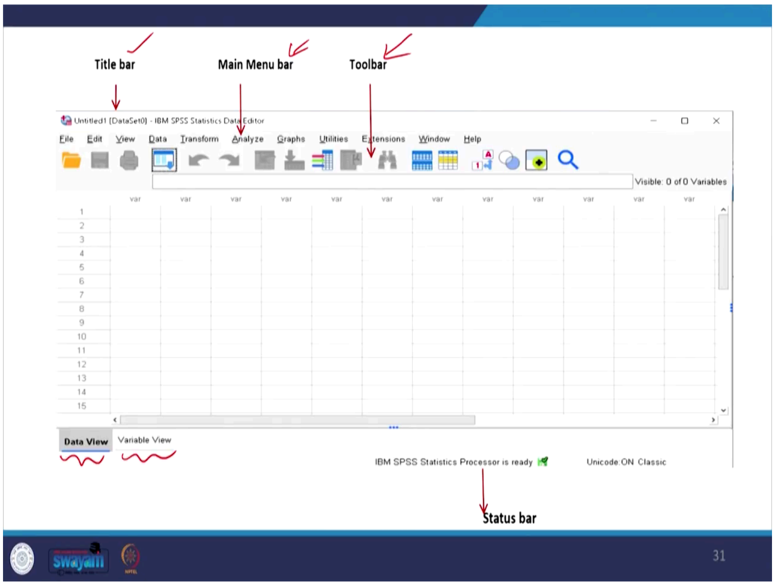
THE BARS AND DROP-DOWN MENUS

- ❑ After SPSS gets started a data window will appear. This window contains four bars:
  - ➢ The **title bar** contains the name of the file that has been opened.
  - ➢ The **menu bar** contains main SPSS menus.
  - ➢ The **toolbar** allows you to select the SPSS tasks.
  - ➢ The **status bar** lets you know in what stage the program is in.

For other data files SPSS will prompt for additional information as needed depending on the file type. The bars and drop-down menus are very important in SPSS. After SPSS gets started a data window will appear. Once that is started, data window will appear, this window contains four bars that is quite important: the title bar, menu bar, tool bar, and status bar.

So, the title bar contains specifically the name of the file that has been opened, the menu bar is containing information about SPSS menus. Then toolbar allows you to select the SPSS task, status bar is all about help you to in what stage the program is in.

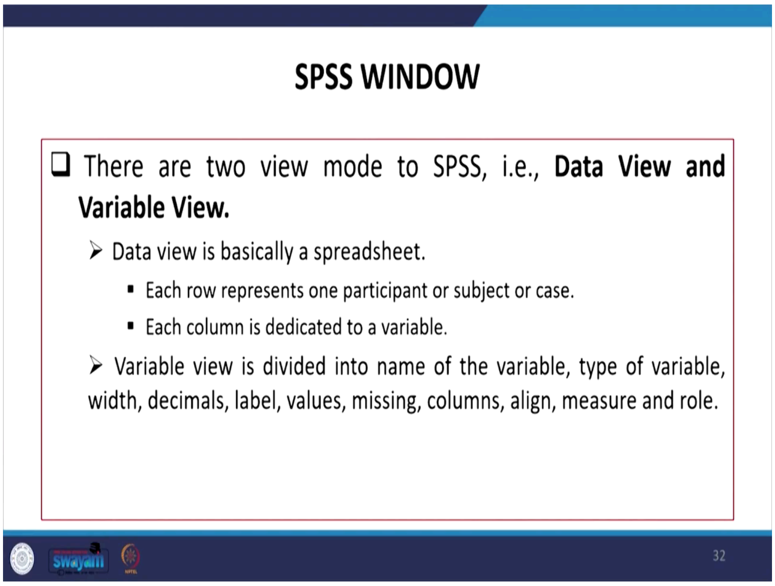
(Refer Slide Time: 36:55)


Title bar
Main Menu bar
Toolbar
Untitled1 [DataSet0] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Graphs Utilities Extensions Window Help
Visible: 0 of 0 Variables
var var var var var var var var var var var var
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Data View
Variable View
IBM SPSS Statistics Processor is ready
Unicode:ON Classic
Status bar
31

This is what is all about. We have given the snapshot of the page. Here is all about, this is the title bar, this is the menu bar here, then this is the tool bar and then the status bar you can read that SPSS statistics processor is ready to work.

Another two things we are also going to guide you is called data view and variable view. One important aspect I also suggest to everyone is that while anyone is doing survey on their own on a very small-scale basis, and then you wanted to operate in Stata or in SPSS. So, better to enter your data in SPSS format. SPSS data format is very handy and going to give you correct check, I will also guide you in between how you should do it.

(Refer Slide Time: 37:52)



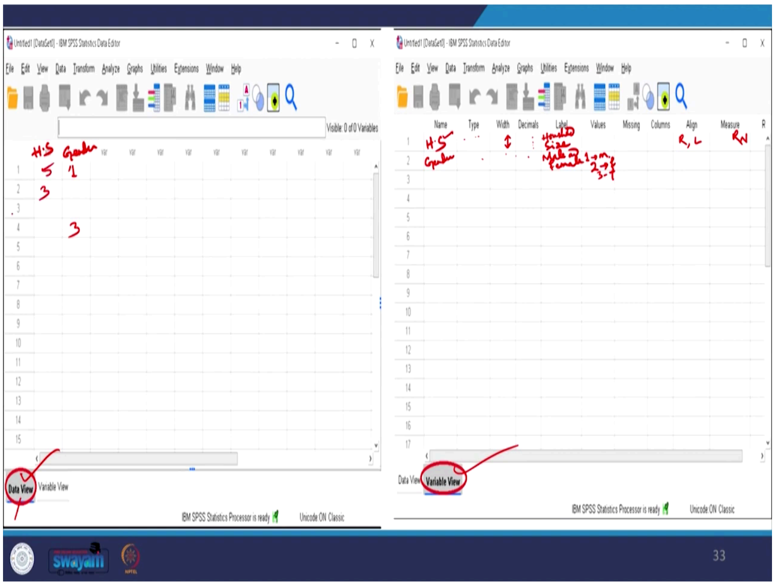
SPSS WINDOW

- ❑ There are two view mode to SPSS, i.e., **Data View and Variable View.**
  - ➢ Data view is basically a spreadsheet.
    - Each row represents one participant or subject or case.
    - Each column is dedicated to a variable.
  - ➢ Variable view is divided into name of the variable, type of variable, width, decimals, label, values, missing, columns, align, measure and role.

So, what is all about SPSS window, there are two views more to of SPSS that is called data view and variable view, which I just mentioned: data view and variable view. Data view is basically a spread sheet, which we used to show in Stata that shows each role represents one participant or subject or case, each column is dedicated to a variable.

So, variable view is divided into name of the variable, type of the variable, with decimal its width, level, values, missing columns, aligning in which, right alignment or left alignment measure which type of measurement of the data you want to specify, and other roles are important as well.

(Refer Slide Time: 38:45)


Data View
Variable View
H.S Gender
5 1
3
3
Name Type Width Decimals Label Values Missing Columns Align Measure
H.S
Gender
R, L
33

We will also show it. Here on the one side, we have shown what is called Data View and this is here this is called Variable View. On the Variable View suppose I said I write down the variable name, let it be variable name is first name is a Name like I said household size.

Suppose I said household size and what I will do I will specify its Width, as a width is till 8th number, till how much number you wanted to specify width it has to be either you increase, or you will decrease. So, then label you will specifically it write down the details, you can write down household size in detail.

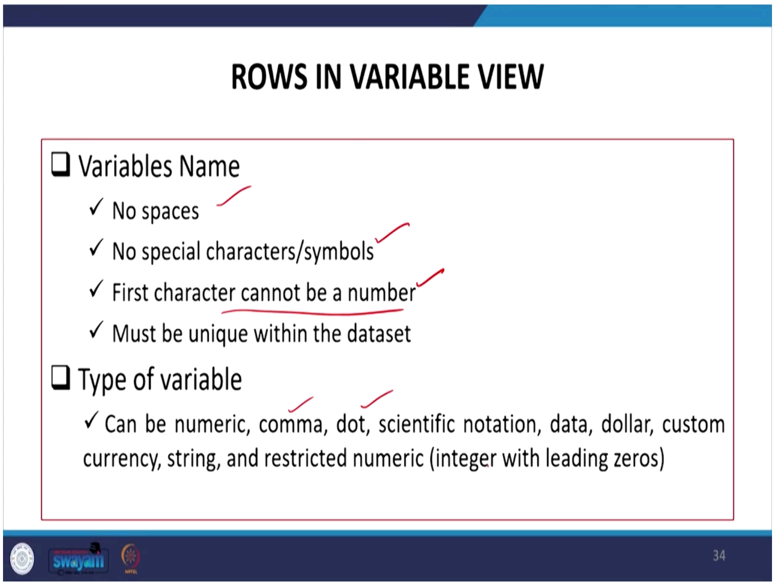
And similarly, there are other till what decimal point you can keep it, you can also specify which type of data variable is it then its alignment, right alignment or left alignment you can specify. Then most importantly its measurement, whether it is in ratio data, numeric data etc. can also be specified. Once you have defined your variable number here, once you enter it and like values, suppose this is in number.

Suppose the next variable is called gender of the person who has responded, gender likewise these you can specify, gender you can say male and female. Suppose this is what you wanted to address and gender, you can specify the values on the values point. Like you can say, on the once you click here it will redirect you to another small window, where you must give 1 stand for male, 2 stands for female, you have to specify here or 3 stands for transgender.

So, once you do it accordingly your data and on the data view all your entry will go like this. Now, this side it is your variable, on the column wise you can see your variables. Here it will automatically show you like this is household size, on the next it will automatically show you gender, since you have already entered there. So, the number it will come like first observation, if the first observation is male then 1 will be entered, if you have entered 1 there.

If fourth is the third gender, then 3 will come. So, if these are quite number maybe first individual saying that in a household, we have five members. Another might be saying three members likewise these others will respond.

(Refer Slide Time: 41:49)



ROWS IN VARIABLE VIEW

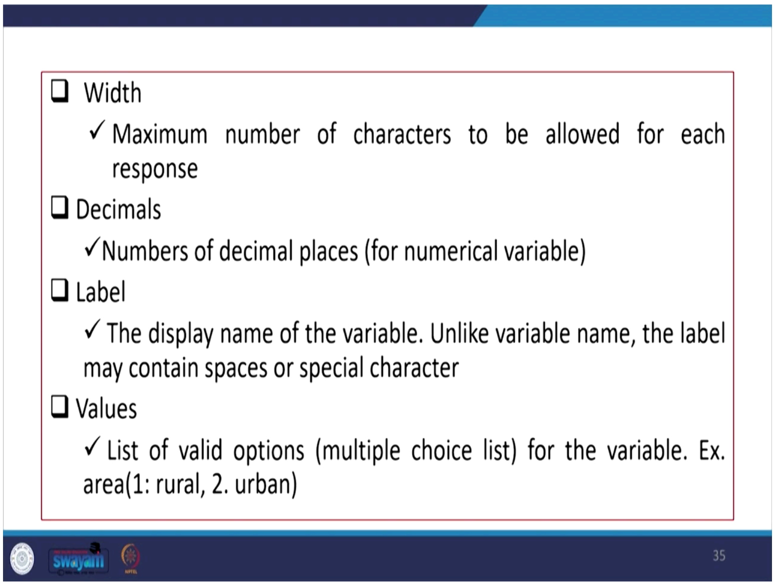
❑ Variables Name
- ✓ No spaces
- ✓ No special characters/symbols
- ✓ First character cannot be a number
- ✓ Must be unique within the dataset

❑ Type of variable
- ✓ Can be numeric, comma, dot, scientific notation, data, dollar, custom currency, string, and restricted numeric (integer with leading zeros)

So, what are the rows in variable view and like rows I have already said that they are basically the cases, a row in variable view basically you need to specify the variable names. And rows in data view that is basically cases.

So, variable name you need to give with no spaces, that is important no space no special character or symbols. First character cannot be a number that is another important aspect and that must be unique within the data set, should not be repeating with the same name. Like type of variable can be numeric, comma, dot, you must specify, dollar etc. a rest of the detail you will find out by your own operation.

(Refer Slide Time: 42:42)



□ Width

✓ Maximum number of characters to be allowed for each response

□ Decimals

✓Numbers of decimal places (for numerical variable)

□ Label

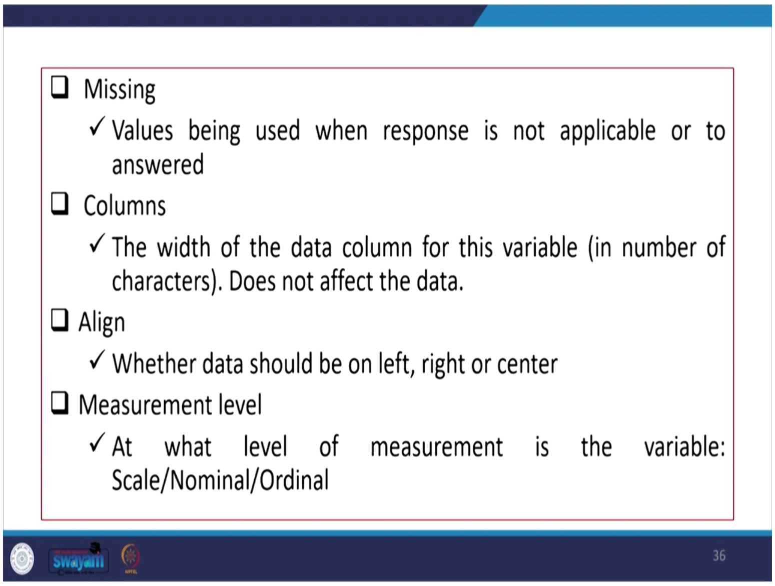
✓ The display name of the variable. Unlike variable name, the label may contain spaces or special character

□ Values

✓ List of valid options (multiple choice list) for the variable. Ex. area(1: rural, 2. urban)

Width, similarly, decimals point, label, values, etc. I have already defined to you.

(Refer Slide Time: 42:48)



□ Missing

✓ Values being used when response is not applicable or to answered

□ Columns

✓ The width of the data column for this variable (in number of characters). Does not affect the data.

□ Align

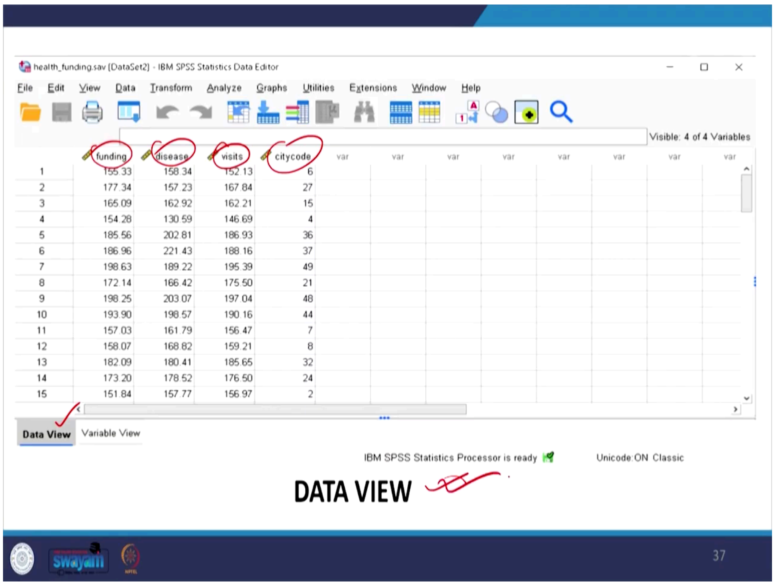
✓ Whether data should be on left, right or center

□ Measurement level

✓ At what level of measurement is the variable: Scale/Nominal/Ordinal

This one is having missing values, like values being used when response is not applicable or not to be answered, then you can add as missing. Similarly, cost columns the width of the data column for this variable should also be mentioned, left, right alignment should be given. Measurements in scale, nominal ordinal etc. should be defined.

(Refer Slide Time: 43:16)



health_funding.sav [DataSet2] - IBM SPSS Statistics Data Editor

Visible: 4 of 4 Variables

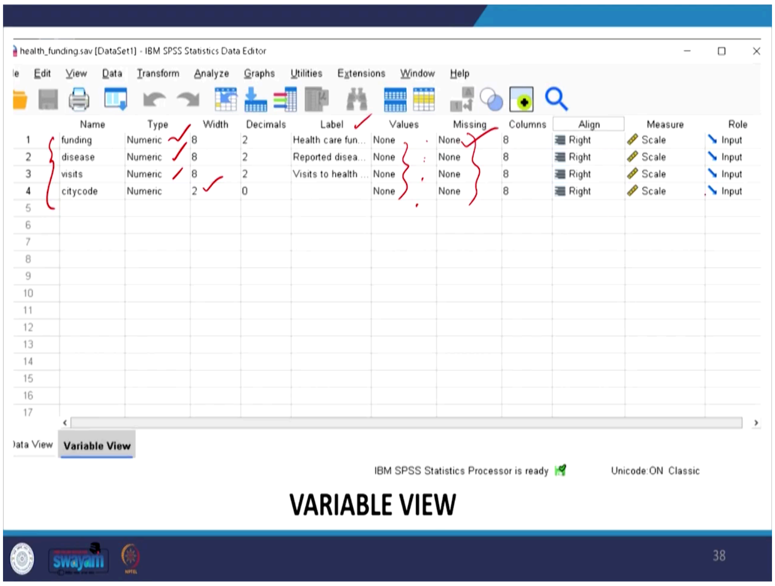
	funding	disease	visits	citycode	var	var	var	var	var	var	var	var
1	155.33	158.34	152.13	6								
2	177.34	157.23	167.84	27								
3	165.09	162.92	162.21	15								
4	154.28	130.59	146.69	4								
5	185.56	202.81	186.93	36								
6	186.96	221.43	188.16	37								
7	198.63	189.22	195.39	49								
8	172.14	166.42	175.50	21								
9	198.25	203.07	197.04	48								
10	193.90	198.57	190.16	44								
11	157.03	161.79	156.47	7								
12	158.07	168.82	159.21	8								
13	182.09	180.41	185.65	32								
14	173.20	178.52	176.50	24								
15	151.84	157.77	156.97	2								

IBM SPSS Statistics Processor is ready   Unicode:ON Classic

DATA VIEW

Here we have given the decimal point, the data with decimal point city code without decimal point. So far as data view is concerned, in variable view basically these are entered in the variable view, and this is in data view we have presented.

(Refer Slide Time: 43:39)



health_funding.sav [DataSet1] - IBM SPSS Statistics Data Editor

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	funding	Numeric	8	2	Health care fun...	None	None	8	Right	Scale	Input
2	disease	Numeric	8	2	Reported disea...	None	None	8	Right	Scale	Input
3	visits	Numeric	8	2	Visits to health ...	None	None	8	Right	Scale	Input
4	citycode	Numeric	2	0		None	None	8	Right	Scale	Input
5											
6											
7											
8											
9											
10											
11											
12											
13											
14											
15											
16											
17											

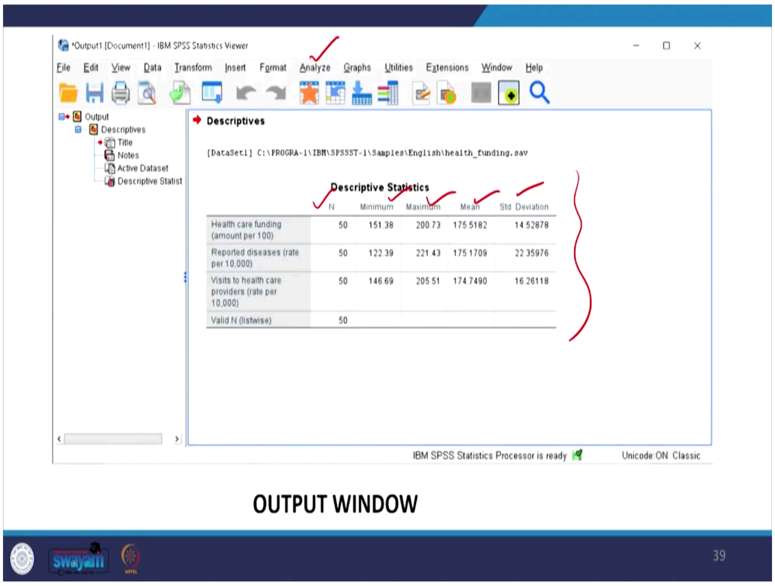
IBM SPSS Statistics Processor is ready   Unicode:ON Classic

VARIABLE VIEW

And this is in variable view, this is what the variable we have shown to you just now. Numeric we have specified as numeric here. Then its width at maximum it is going to read 8th width spaces at since 8th you have defined. City code we know that it is not going to be more than 3 entries. So, 2 space at maximum 2 width; so, 2 we have given.

Similar Decimal point wherever it is not required we can give it 0. Label is defined, it is detailing or detailed name of that particular variable, Values if it is categorical and then you could have defined the values accordingly. Missing, since no missing is there, we have mentioned as none, usually it goes with none, then alignment scale etc. can also be defined clearly in the variable view.

(Refer Slide Time: 44:32)


*Output1 [Document1] - IBM SPSS Statistics Viewer
File Edit View Data Transform Insert Format Analyze Graphs Utilities Extensions Window Help
Output
Descriptives
Title
Notes
Active Dataset
Descriptive Statist
IBM SPSS Statistics Processor is ready
Unicode:ON Classic
39

Descriptives

```
[DataSet1] C:\PROGRA~1\IBM\SPSSST~1\Samples\English\health_funding.sav
```

Descriptive Statistics

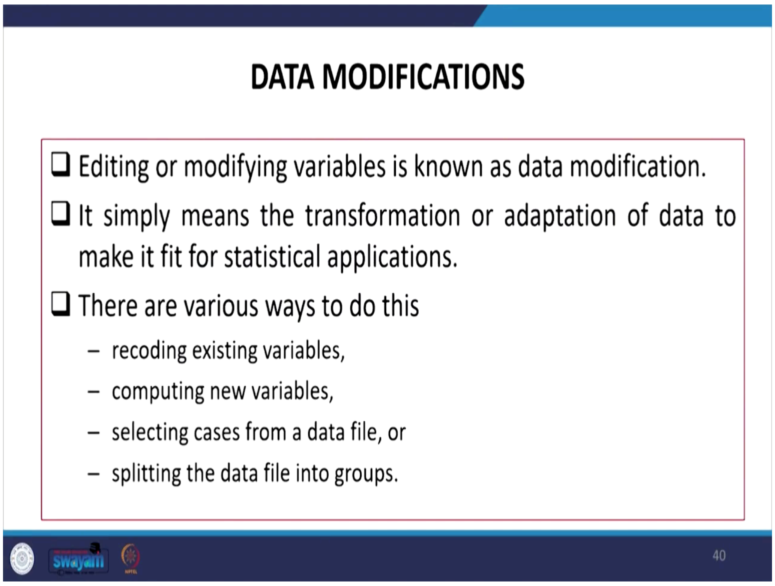
	N	Minimum	Maximum	Mean	Std. Deviation
Health care funding (amount per 100)	50	151.38	200.73	175.5182	14.52878
Reported diseases (rate per 10,000)	50	122.39	221.43	175.1709	22.35976
Visits to health care providers (rate per 10,000)	50	146.69	205.51	174.7490	16.26118
Valid N (listwise)	50				

OUTPUT WINDOW

Now, after making sure that we have done those basic details, we should go for some descriptive analysis. How to do the descriptive analysis? Basically, that will be displayed in our output window, this is what is our output window, we can do it through the Analyze.

Within the Analyze once you click on it will ask for descriptive. It is there in our slide, descriptive likewise we did in our Stata, we will have the same thing as number of observation minimum, maximum, mean value and standard deviation. All those things are also available here.

(Refer Slide Time: 45:16)



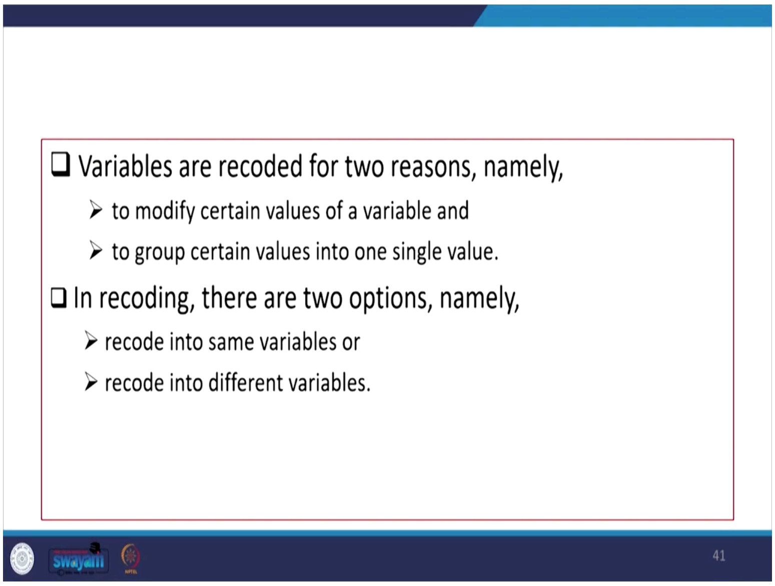
DATA MODIFICATIONS

- ❑ Editing or modifying variables is known as data modification.
- ❑ It simply means the transformation or adaptation of data to make it fit for statistical applications.
- ❑ There are various ways to do this
  - recoding existing variables,
  - computing new variables,
  - selecting cases from a data file, or
  - splitting the data file into groups.

Some data modifications can be done editing or modifying variables is known as data modification. It simply means the transformation or adoption of data to make it fit for statistical application. There are various ways to do this, recoding of the existing variables, computing new variables, selecting cases from a data file, or splitting the data into groups.

Likewise in Stata we go for do file here in SPSS we have syntax. In syntax we can also save our command and we can also operate in SPSS as well. But you know most of the people prefer to walk through the drop-down menus, because it is very friendly and in very fast.

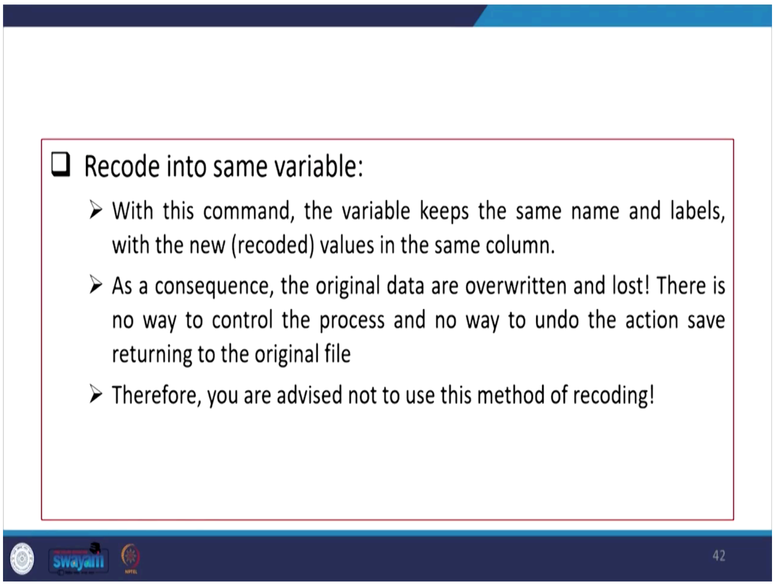
(Refer Slide Time: 46:07)



- ❑ Variables are recoded for two reasons, namely,
  - ➢ to modify certain values of a variable and
  - ➢ to group certain values into one single value.
- ❑ In recoding, there are two options, namely,
  - ➢ recode into same variables or
  - ➢ recode into different variables.

Variables are recoded for two reasons, because of modifying certain values of a variable or to group certain values into one single value. In recoding there are two options that is called recoding into same variable or recode into different variable. So, both are possible.

(Refer Slide Time: 46:30)



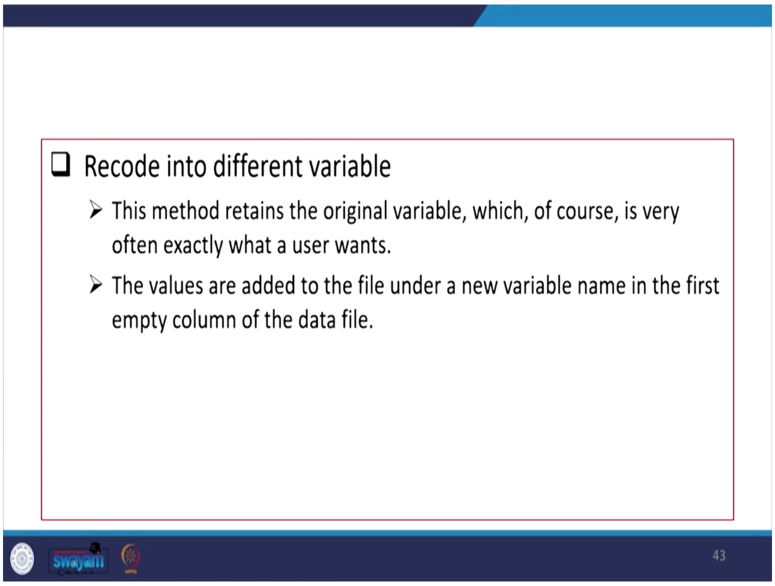
❑ Recode into same variable:

- ➢ With this command, the variable keeps the same name and labels, with the new (recoded) values in the same column.
- ➢ As a consequence, the original data are overwritten and lost! There is no way to control the process and no way to undo the action save returning to the original file
- ➢ Therefore, you are advised not to use this method of recoding!

If it is recoding into same variable, then within this command the variable keeps the same name. Labels with the new values in the same column. New values in the same column, the same column or but the new values will be entered. As a consequences of this command, the original data are overwritten now and lost with the previous data that is stored.

There is no way to control the process and no way to undo the action save action save returning to the original file. Therefore, you are advised to use this method of recoding, you are suggested not to use this method for recoding rather the second method should be adopted, that is to go for a new recode to a different variable, that is always suggested, this one is most preferred.

(Refer Slide Time: 47:33)

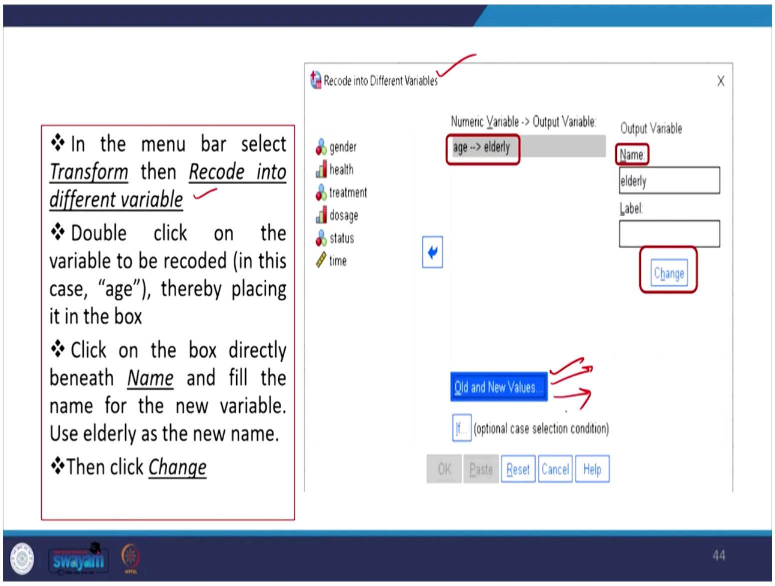


❑ Recode into different variable

- ➢ This method retains the original variable, which, of course, is very often exactly what a user wants.
- ➢ The values are added to the file under a new variable name in the first empty column of the data file.

Since this is replacing the original data, so you might miss those things for further work. If you are adapting decode into different variable then basically a new original variable you can create and based on your requirement, those new values are added to the file under a new variable name in the first empty column of the data set. We will show it.

(Refer Slide Time: 47:57)



❖ In the menu bar select *Transform* then *Recode into different variable*

❖ Double click on the variable to be recoded (in this case, "age"), thereby placing it in the box

❖ Click on the box directly beneath *Name* and fill the name for the new variable. Use elderly as the new name.

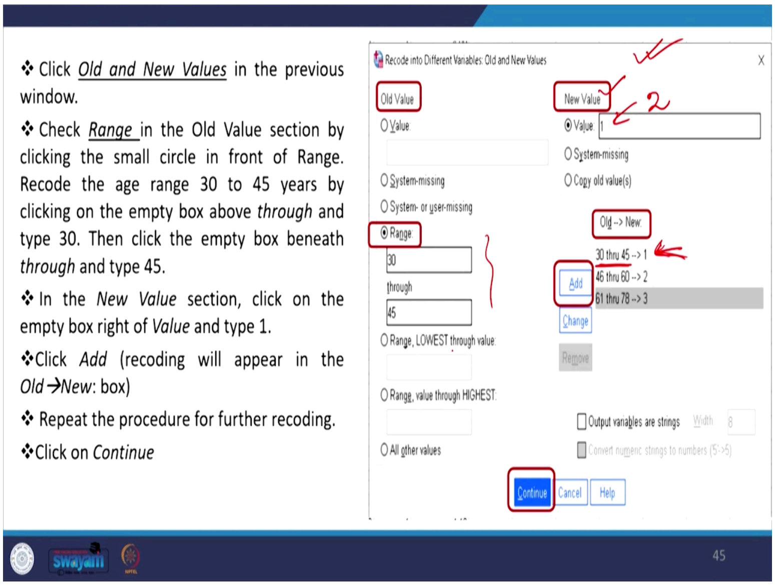
❖Then click *Change*

Here on the command recode into different variable if you go to edit data, a menu like on the menu bar select transform, transform it is there on the window. Here is the transform, transform comma you want to transform the data. You go there, you will see this particular command. Recode into different variable, it will give you the option like recode into same variable and recode into different variable. So, better to click on this.

Then double click on the variable to be recoded, that is the variable you wanted to recode you can just simply double click on it, it will come over here on the screen. Then like double click in this case we have used age, age, thereby replacing it in the box we have kept it in the box.

Now name we have given as elderly, elderly click on the box directly beneath name and fill the name for the new variable, use elderly as a new name and it will automatically change after once you save it. But if you have different values, within the variable if you want to change certain values, some new values you wanted to give it you can also take this option.

(Refer Slide Time: 49:21)



- ❖ Click *Old and New Values* in the previous window.
- ❖ Check *Range* in the Old Value section by clicking the small circle in front of Range. Recode the age range 30 to 45 years by clicking on the empty box above *through* and type 30. Then click the empty box beneath *through* and type 45.
- ❖ In the *New Value* section, click on the empty box right of *Value* and type 1.
- ❖Click *Add* (recoding will appear in the *Old→New*: box)
- ❖ Repeat the procedure for further recoding.
- ❖Click on *Continue*

Recode into Different Variables: Old and New Values

Old Value

○ Value:

○ System-missing

○ System- or user-missing

● Range:

30

through

45

○ Range, LOWEST through value:

○ Range, value through HIGHEST:

○ All other values

New Value

● Value: 1

○ System-missing

○ Copy old value(s)

Old --> New:

30 thru 45 --> 1

46 thru 60 --> 2

61 thru 78 --> 3

□ Output variables are strings Width: 8

□ Convert numeric strings to numbers ('5'->5)

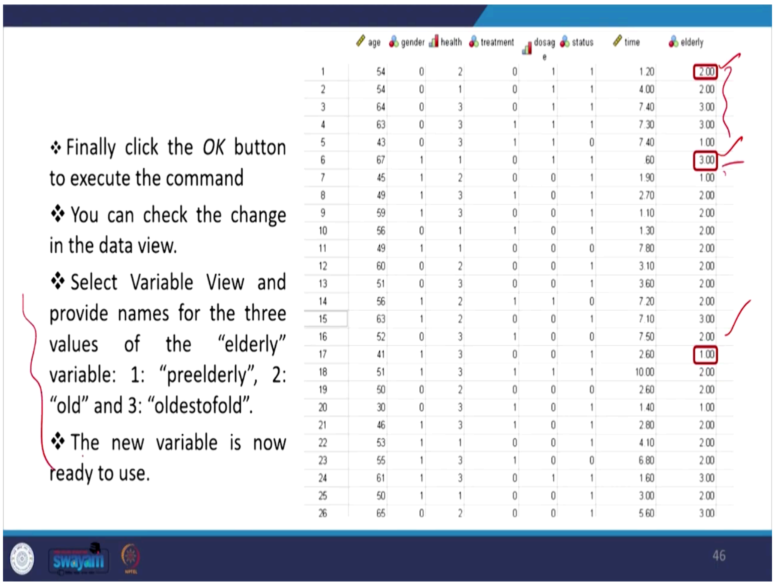
And on once you click it you will find out all those things, yes. We have given it here, I recode into different variables with Old and New Values, that values window we have kept it in this particular page. So, Old and New Values in the previous window we are now carrying forward.

Simply check Range, then range you can define 30 through 45, if you give it will automate then 30 to 45 which is code name. Suppose we wanted to give 30 to 45 as with the code as 1, like in the standard of living if you wanted to give their standard till 20000 is their expenditure that will be carrying one code. Then 20000 one expenditure till 50000 expenditure you can give code 2.

You are basically transferring the values. You may make it a categorical variable. So, in that case so what we are defining we are taking the option called range, we have defined this and given a value 1. Once in the new value it if it is one, then once you add it will be added here on this box then again if I change, then 46 through 60 then we will change it to 2 here, we will change it to 2 write it down 2 here.

Then add it will actually change it and similarly for 3 as well. So, similarly we can also define lowest and its highest value and can find out which one is the best and rest of the details we have guided. And I am sure you will enjoy and dealing with these things.

(Refer Slide Time: 51:00)



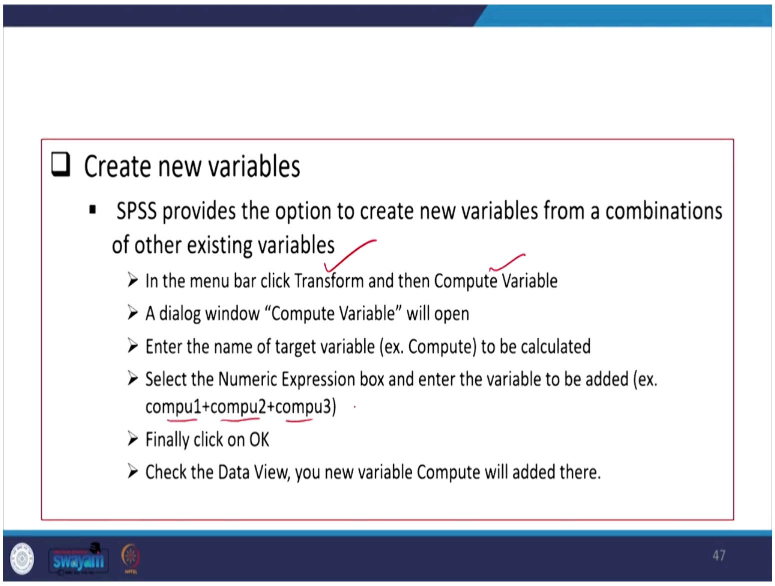
- Finally click the *OK* button to execute the command
- You can check the change in the data view.
- Select Variable View and provide names for the three values of the "elderly" variable: 1: "preelderly", 2: "old" and 3: "oldestofold".
- The new variable is now ready to use.

	age	gender	health	treatment	dosage	status	time	elderly
1	54	0	2	0	1	1	1.20	2.00
2	54	0	1	0	1	1	4.00	2.00
3	64	0	3	0	1	1	7.40	3.00
4	63	0	3	1	1	1	7.30	3.00
5	43	0	3	1	1	0	7.40	1.00
6	67	1	1	0	1	1	.60	3.00
7	45	1	2	0	0	1	1.90	1.00
8	49	1	3	1	0	1	2.70	2.00
9	59	1	3	0	0	1	1.10	2.00
10	56	0	1	1	0	1	1.30	2.00
11	49	1	1	0	0	0	7.80	2.00
12	60	0	2	0	0	1	3.10	2.00
13	51	0	3	0	0	1	3.60	2.00
14	56	1	2	1	1	0	7.20	2.00
15	63	1	2	0	0	1	7.10	3.00
16	52	0	3	1	0	0	7.50	2.00
17	41	1	3	0	0	1	2.60	1.00
18	51	1	3	1	1	1	10.00	2.00
19	50	0	2	0	0	0	2.60	2.00
20	30	0	3	1	0	1	1.40	1.00
21	46	1	3	1	0	1	2.80	2.00
22	53	1	1	0	0	1	4.10	2.00
23	55	1	3	1	0	0	6.80	2.00
24	61	1	3	0	1	1	1.60	3.00
25	50	1	1	0	0	1	3.00	2.00
26	65	0	2	0	0	1	5.60	3.00

46

So, now, the elderly the data whatever is changed, it is highlighted, the data that can be recoded the age can be recoded to different groups. So, 2 here, we have defined 3 and accordingly I think some guidance I have given it here for your reference I am sure you will enjoy operating on your own.

(Refer Slide Time: 51:28)



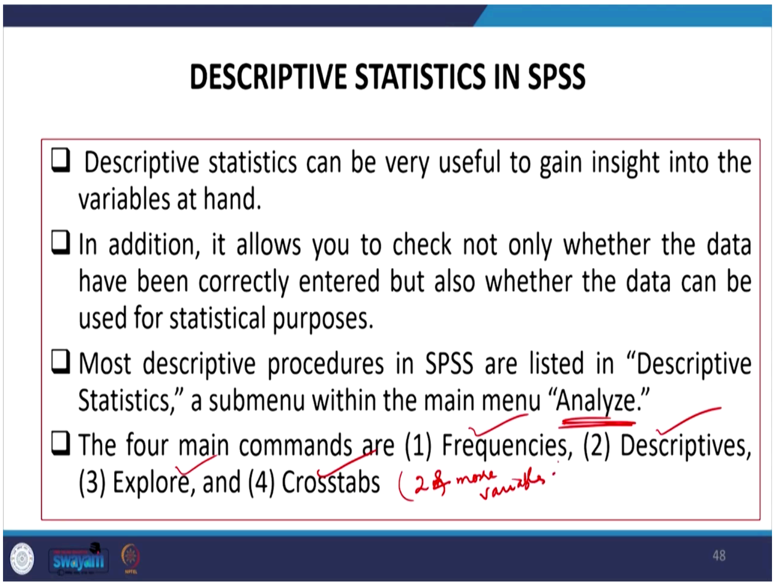
❑ Create new variables

- SPSS provides the option to create new variables from a combinations of other existing variables
  - ➢ In the menu bar click Transform and then Compute Variable
  - ➢ A dialog window "Compute Variable" will open
  - ➢ Enter the name of target variable (ex. Compute) to be calculated
  - ➢ Select the Numeric Expression box and enter the variable to be added (ex. compu1+compu2+compu3)
  - ➢ Finally click on OK
  - ➢ Check the Data View, you new variable Compute will added there.

Similarly, creating a new variable it is possible, SPSS provide the options to create new variables from a combination of other existing variables, in again from the transform option on the menu bar, you click there then you compute variable this is what is required.

A dialogue window will be coming as Compute Variable, then you can execute with options, some options we can give it like column compute 1, 2, 3 etc. will also we can able to get all those details.

(Refer Slide Time: 52:08)



DESCRIPTIVE STATISTICS IN SPSS

❑ Descriptive statistics can be very useful to gain insight into the variables at hand.

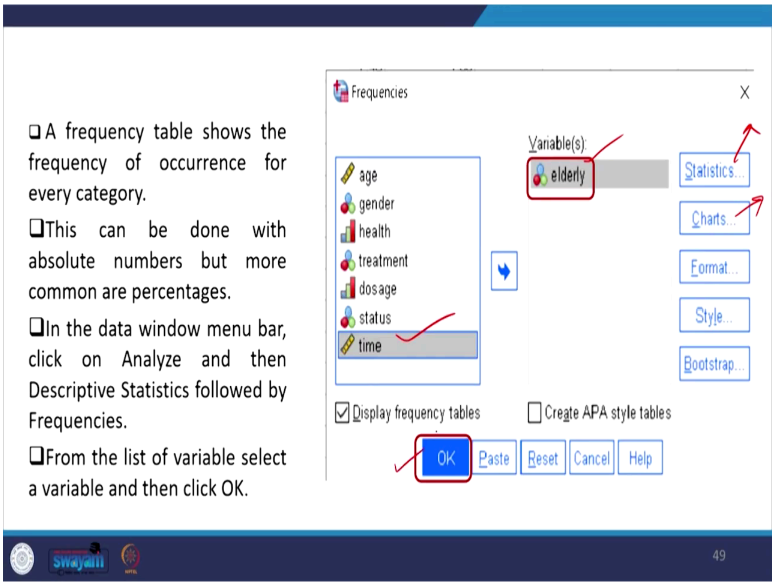
❑ In addition, it allows you to check not only whether the data have been correctly entered but also whether the data can be used for statistical purposes.

❑ Most descriptive procedures in SPSS are listed in "Descriptive Statistics," a submenu within the main menu "Analyze."

❑ The four main commands are (1) Frequencies, (2) Descriptives, (3) Explore, and (4) Crosstabs (2 & more variables.

Descriptive statistics is possible, likewise we do in Stata. In SPSS also we go to "Analyze"; Analyze and then we will get four options, Frequencies, Descriptive, Explore or Cross tabulation. Cross tabulation where we require more than one variable, may be two and more variable.

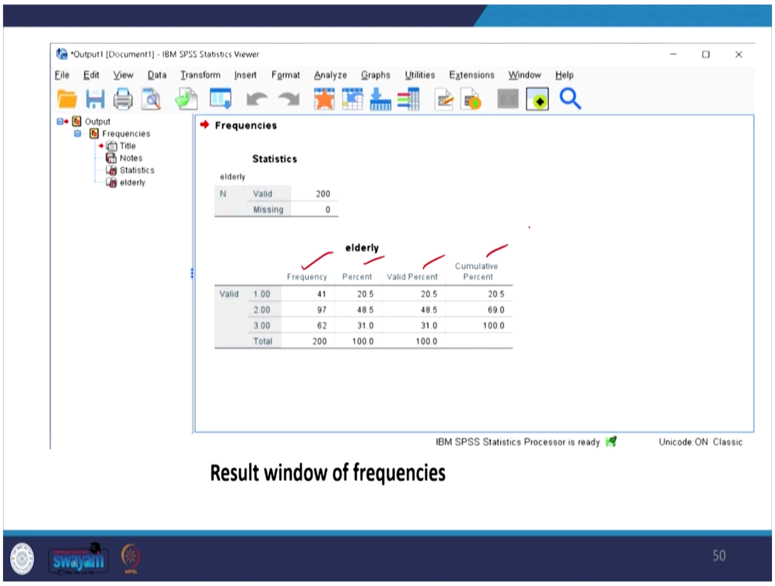
(Refer Slide Time: 52:42)



- ❑A frequency table shows the frequency of occurrence for every category.
- ❑This can be done with absolute numbers but more common are percentages.
- ❑In the data window menu bar, click on Analyze and then Descriptive Statistics followed by Frequencies.
- ❑From the list of variable select a variable and then click OK.

I am not explaining much, rest you will be guided on your own. If you click on frequencies, you want these variables or elderly variable and its frequency. You double click on the elderly and then click it here, you can take options for your statistics what sort of statistics I want, in my result all so. Whether I want in chart also, which style I want, which format I want and once you click, it will give you the table like this.

(Refer Slide Time: 53:09)


*Output1 [Document1] - IBM SPSS Statistics Viewer
File Edit View Data Transform Insert Format Analyze Graphs Utilities Extensions Window Help
Output
Frequencies
Title
Notes
Statistics
elderly
IBM SPSS Statistics Processor is ready
Unicode:ON Classic
50

Frequencies

Statistics

elderly

N	Valid	200
	Missing	0

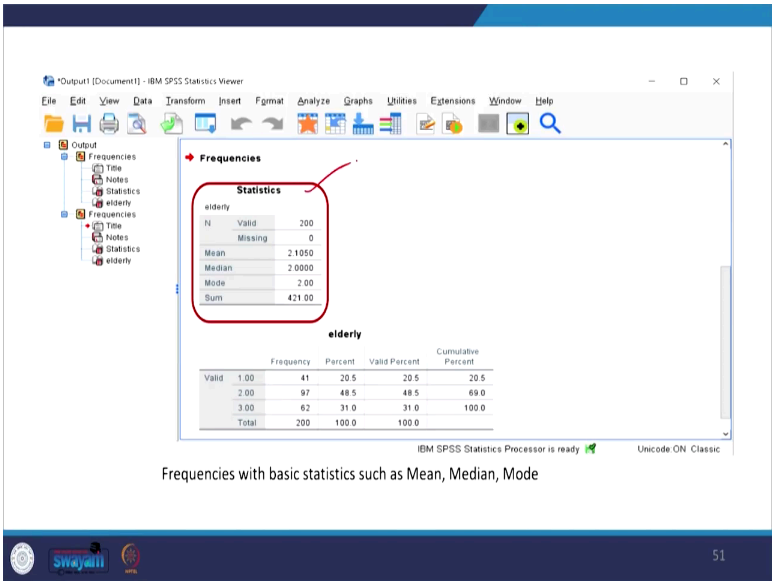
elderly

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1.00	41	20.5	20.5	20.5
	2.00	97	48.5	48.5	69.0
	3.00	62	31.0	31.0	100.0
	Total	200	100.0	100.0	

Result window of frequencies

It gives you Frequency, Percentage, Valid Percentage and Cumulative Percentage as well.

(Refer Slide Time: 53:16)


*Output1 [Document1] - IBM SPSS Statistics Viewer
File Edit View Data Transform Insert Format Analyze Graphs Utilities Extensions Window Help
Output
Frequencies
Title
Notes
Statistics
elderly
Frequencies
Title
Notes
Statistics
elderly
IBM SPSS Statistics Processor is ready
Unicode:ON Classic
51

Frequencies

Statistics

elderly

N	Valid	200
	Missing	0
Mean		2.1050
Median		2.0000
Mode		2.00
Sum		421.00

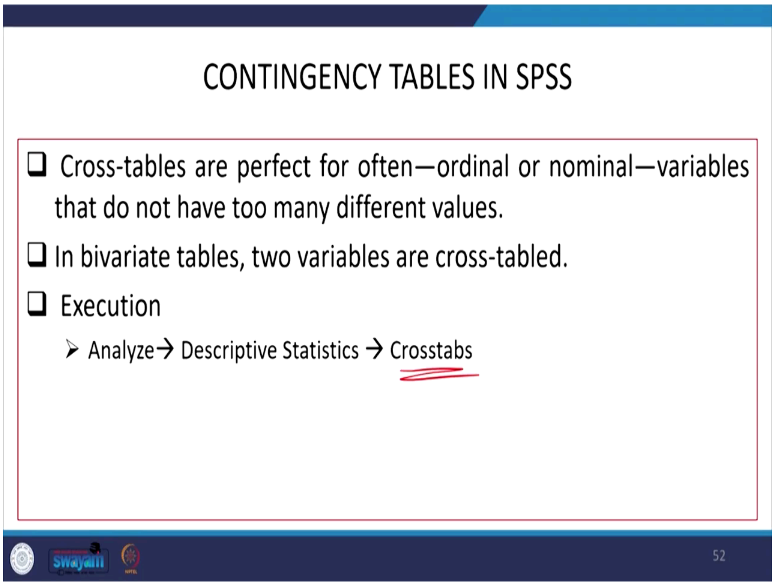
elderly

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1.00	41	20.5	20.5	20.5
	2.00	97	48.5	48.5	69.0
	3.00	62	31.0	31.0	100.0
	Total	200	100.0	100.0	

Frequencies with basic statistics such as Mean, Median, Mode

Similarly, if you take some statistics options as well, it will give you mean value, median, mode, etc.

(Refer Slide Time: 53:25)

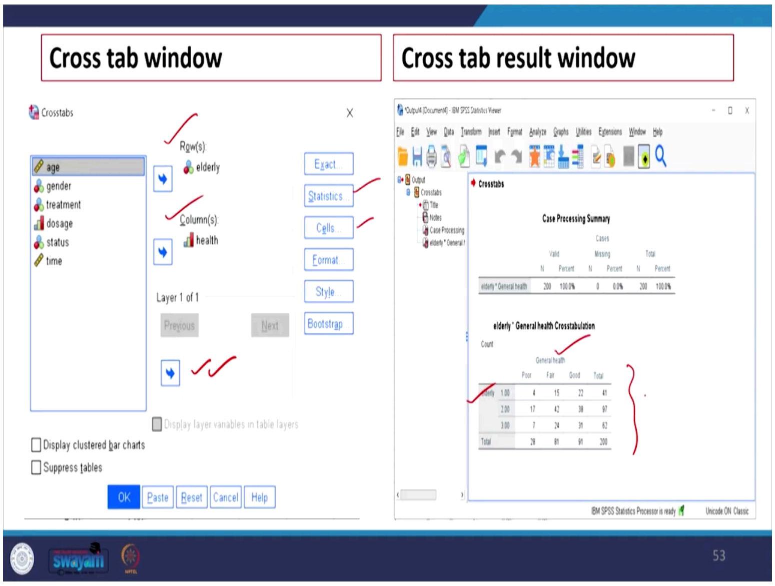


CONTINGENCY TABLES IN SPSS

- ❑ Cross-tables are perfect for often—ordinal or nominal—variables that do not have too many different values.
- ❑ In bivariate tables, two variables are cross-tabled.
- ❑ Execution
  - ➤ Analyze→ Descriptive Statistics → Crosstabs

Contingency table in Stata through cross tabulation, contingency table basically if the data are original or nominal. These kinds of tables are preferred, and if this is basically called bivariate tables, we can do it.

(Refer Slide Time: 53:42)



Cross tab window

Crosstabs

age
gender
treatment
dosage
status
time

Row(s): elderly

Column(s): health

Layer 1 of 1

□ Display layer variables in table layers

□ Display clustered bar charts

□ Suppress tables

Cross tab result window

*Output4 [Document4] - IBM SPSS Statistics Viewer

Crosstabs

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
elderly * General health	200	100.0%	0	0.0%	200	100.0%

elderly * General health Crosstabulation

Count

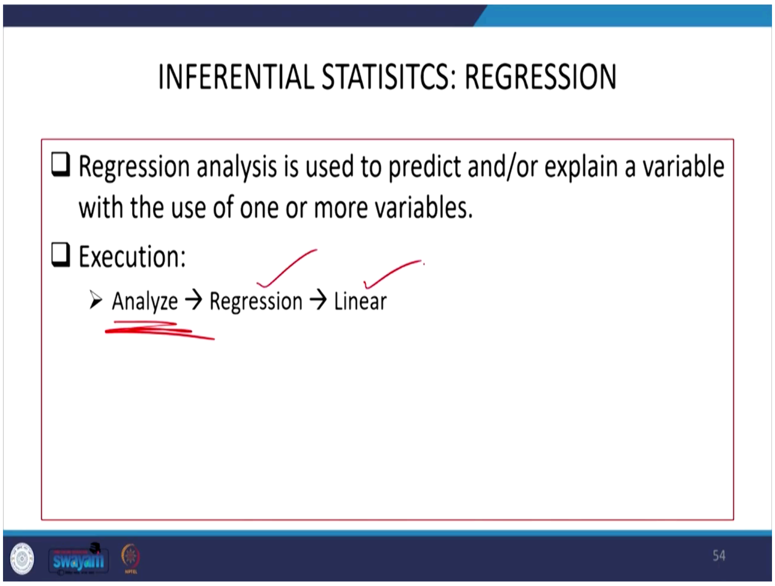
		General health			
		Poor	Fair	Good	Total
elderly	1.00	4	15	22	41
	2.00	17	42	38	97
	3.00	7	24	31	62
Total		28	81	91	200

IBM SPSS Statistics Processor is ready   Unicode:ON Classic

We have to select row variable and column variable. One variable in Row and another variable in Column. Then accordingly we can take it Statistics, sale percentage or column percentage those things we will have to be defined. Then once you click on ok or if you want to have a third variable, likewise you did it by sort in Stata you can keep it the third variable over here, layer 1 of 1. So, here gender and age or elderly; that is health and elderly we have taken a cross tabulation at contingency table by for 2 and 2 cross table.

We have taken some basic statistics as well and the result is displayed here and once you operate on the same way, I am sure you will get the result very correctly.

(Refer Slide Time: 54:32)

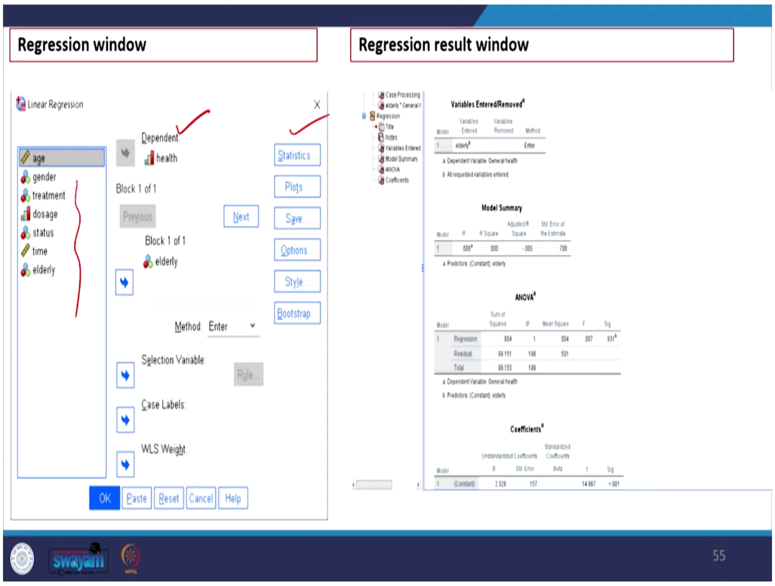


INFERENTIAL STATISITCS: REGRESSION

- ❑ Regression analysis is used to predict and/or explain a variable with the use of one or more variables.
- ❑ Execution:
  - ➢ Analyze → Regression → Linear

Similarly, from the analyze window analyze point or on the main bar menu bar, you will get regression, you click on linear regression, and you specify the variables.

(Refer Slide Time: 54:46)



Regression window

Regression result window

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	elderly[b]		Enter

a Dependent Variable: General health

b All requested variables entered

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.006[a]	.000	-.005	.708

a Predictors: (Constant), elderly

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	.004	1	.004	.007	.931[b]
	Residual	99.151	198	.501		
	Total	99.155	199			

a Dependent Variable: General health

b Predictors: (Constant), elderly

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	2.328	.157		14.867	<.001

Specify variable means integration you are supposed to discuss your dependent variable, your independent variable correctly. Then if you click on the result, with some statistics then that will surely give you give you the result correctly. Or other details I am not guiding because of time paucity, we may continue if your requirements are there, we will also be happy to guide you further.

But here all selections of variables you can specify, all other variables the other than dependent can also be mentioned. And once you mention it will help you to derive the result. So, these are all for SPSS, to start with I know that many of the discussions still required once you start operating on your own, I am sure you will understand many things of SPSS for sure.

With these I think I am looking forward for your queries and from the next class we will be happy to address you on data extraction and merging using an access data.

Thank you.