



International Labour Organization

Introduction to Stata

Mabelin Villarreal-Fuentes
Data Production and Analysis Unit
ILO Department of Statistics

Measuring decent work in the context of the SDGs
November 2018 - Chiba, Japan

OBJECTIVES



This is an introduction to the statistical software Stata aiming at:

- Preparing the participants in Stata basics (interphase and commands) for the next practical sessions.
- Doing some preliminary inspection and data manipulation using the Labour Force Survey (LFS) of Myanmar 2017Q1.



WHAT IS STATA?

- It is a multi-purpose statistical package to explore, summarize and analyze information organized in datasets.
- Its first version was officially released in January 1985. The last version (Stata v.15) in 2015.
- Stata is widely used in social science research (especially economics, political science, epidemiology and medical science).
- Other statistical software: SPSS, SAS, R, etc.

STATA 13 SCREEN



The screenshot displays the Stata 13 interface with the following components:

- 3. COMMANDS:** A list of commands entered in the command window:

```
1 use "C:\Users\vilarrreal-fuentes\Desktop\MMR_LFS_2017Q1_ITC"
2 summarize b5
3 tabulate b4
```
- 1. VARIABLES IN THE DATASET:** A list of variables in the dataset, including b4 (Sex), b5 (Age), a1 (Sample FSU No.), a2 (Sample Household number), a3 (Quarter), a5 (No. of questionnaires used), a6d (Date of first interviewer visit - Day), a6m (Date of first interviewer visit - Month), a6y (Date of first interviewer visit - Year), a25 (Result of final visit), a26 (Data entry clerk number), a27d (Date of data entry - Day), a27m (Date of data entry - Month), a27y (Date of data entry - Year), a28h (Time of data entry - Hour), a28m (Time of data entry - Minute), a28s (Time of data entry - Second), and a30 (Number of household member).
- 4. RESULTS/OUTPUT WINDOW:** The output of the commands, showing the results of the summarize and tabulate commands.

```
running W:\CONTENT\BOURCES\Applications\ILO\WQ\STATISTICS\STATA13\profile.do ...
. use "C:\Users\vilarrreal-fuentes\Desktop\MMR_LFS_2017Q1_ITC"
. summarize b5
+-----+-----+-----+-----+-----+
| Variable | Obs | Mean | Std. Dev. | Min | Max |
+-----+-----+-----+-----+-----+
| b5       | 58647 | 31.61602 | 19.66695 | 0 | 99 |
+-----+-----+-----+-----+-----+
. tabulate b4
+-----+-----+-----+
| Sex | Freq. | Percent | Cum. |
+-----+-----+-----+
| Male | 27,880 | 47.54 | 47.54 |
| Female | 30,767 | 52.46 | 100.00 |
+-----+-----+-----+
| Total | 58,647 | 100.00 | |
+-----+-----+-----+

```
- 2. COMMAND WINDOW:** The command window is empty.

STATA 13 SCREEN



The screenshot shows the Stata 13.1 interface with the following components:

- Command Window:**

```

1 use "C:\Users\willarreal-fue...
2 summarize b5
3 tabulate b4
    
```
- Results Window:**

Special Edition
 Copyright 1985-2013 StataCorp LP
 Statistics/Data Analysis
 4905 Lakeway Drive
 College Station, Texas 77845 USA
 800-STATA-PC <http://www.stata.com>
 979-696-4600 stata@stata.com
 979-696-4601 (fax)

2-user Stata network perpetual license:
 Serial number: 401306217604
 Licensed to: ILO STATISTICS
 International Labour Organization

Notes:
 1. (/w# option or -set maxvar-) 5000 maximum variables

running M:\CONTENT\SOURCE3\Applications\ILO\HQ\STATISTICS\STATA13\profile.do ...

```

. use "C:\Users\willarreal-fuentes\Desktop\MMR_LFS_2017Q1_ITC"
. summarize b5
. tabulate b4
    
```

Variable	Obs	Mean	Std. Dev.	Min	Max
b5	58647	31.81602	19.86695	0	99

Sex	Freq.	Percent	Cum.
Male	27,880	47.54	47.54
Female	30,767	52.46	100.00
Total	58,647	100.00	
- Variables Window:** Lists variables b4 through x20 with their labels and properties.
- Properties Window:** Shows details for variable b4 (Name: b4, Label: Sex, Type: byte, Format: %8.0g, Value Label: B4).
- Data Window:** Shows file information (Filename: MMR_LFS_2017Q1_ITC.dta, Label, Notes, Variables: 110, Observations: 58,647, Size: 8,17M, Memory: 64M, Sorted by).

THE STATA'S TOOLBAR



The screenshot shows the Stata 13.1 interface with the toolbar highlighted. The toolbar contains 12 numbered icons:

- Open
- Save
- Print
- Log
- New Viewer
- Graph
- New Do-file Editor
- Data Editor
- Data Browser
- Variable manager
- More
- Break

- | | |
|-------------------------------|---|
| 1. Open: | Opens a new data file (use) |
| 2. Save: | Saves the current data file (save) |
| 3. Print: | Prints the content of the results window. |
| 4. Log: | Begin/close/suspend/resume a log file. |
| 5. New Viewer: | Opens a new viewer window to obtain help. |
| 6. Graph: | Bring back the graph window in front. |
| 7. New Do-file Editor: | Opens a new instance of the do-file editor (doedit). |
| 8. Data Editor: | Opens the data editor window (edit). |
| 9. Data Browser: | Opens the data browser (browse). |
| 10. Variable manager: | Manipulate variables. |
| 11. More: | Continues when paused in long output. |
| 12. Break: | Allows canceling current running calculations. |

WAYS OF WORKING WITH STATA

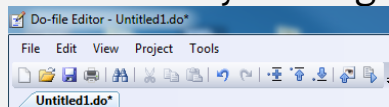
1. **Interactively:** click through the menu/toolbar or typing directly the commands in the command window.
2. **Batch mode:** type up a list of commands in a “do-file” and then execute the file.

Using the **batch mode** (do-files) is the best way to work, because it allows us:

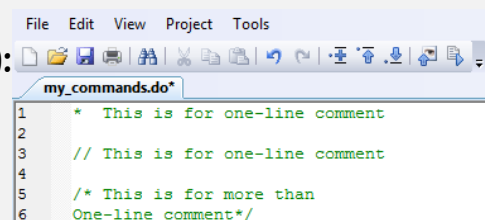
- a. To save our work and keep track of it.
- b. To repeat (copy/paste) commands at convenience.
- c. To suspend/stop our work
- d. To find and fix errors or mistakes
- e. To share our code with colleagues
- f. To better handle a long list of commands (usually the case!)

THE BATCH MODE: « DO-FILE»

1. To open a do-file, click on the «do-file» editor in the toolbar menu or type the command `doedit` on the command window.
2. Write down the commands in the do-file window.
3. Execute the entire do-file by clicking on the last menu button (Execute (do))



4. If you want to execute specific lines: select them and click the executed button.
5. To add comments (green-colored):



6. Save the do-file by clicking on the “save” icon

IMPORTING AND SAVING DATA



Importing data in Stata

Stata's dataset format: ".dta" extension

- Interactively: click on the open icon 📁
- Command: use "my_file.dta", clear

Data in other format:

- Command: insheet using filename (for formats: .csv .txt.xls)
- Interactively: click on "File" → "Import"
- Using StatTransfer software or other tools that could help us to save our data in .dta format.

Saving data in Stata

- Command: save "my_file", replace (saves as my_file; replace is necessary if a file with the same name already exists in the directory and wants to be replaced).

DATA STRUCTURE



- A **dataset** is a collection of separate sets of information usually called **variables** (commonly arranged by columns). (*The Cambridge dictionary*).

- One **variable** is a set of information containing observed, measured or reported characteristics for one or several cases/observations.

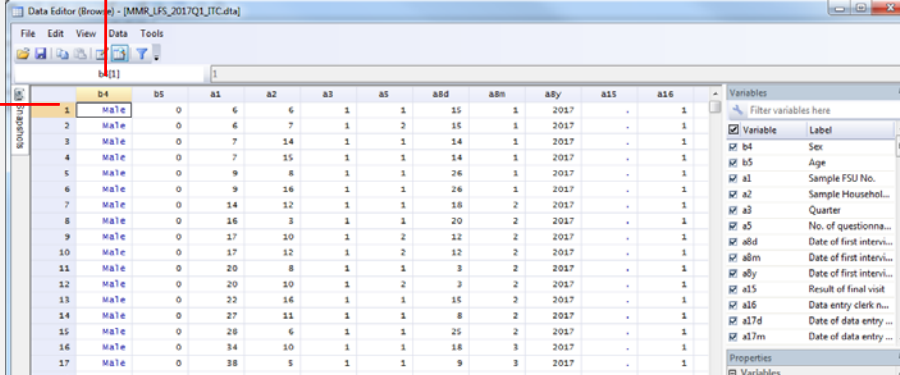
- Typical **grid** structure

- Each **row** represents the unit of observation (an individual, a firm, a region)
- Each **column** represents the values that variable takes for each observation (age, sex, educational attainment, etc.).

DATA STRUCTURE

Column: variable (e.g. sex, age, etc.)

Row: unit of observation (here, a person)



	b4	b5	a1	a2	a3	a5	a8d	a8m	a8y	a15	a16
1	Male	0	6	6	1	1	15	1	2017	-	1
2	Male	0	6	7	1	2	15	1	2017	-	1
3	Male	0	7	14	1	1	14	1	2017	-	1
4	Male	0	7	15	1	1	14	1	2017	-	1
5	Male	0	9	8	1	1	26	1	2017	-	1
6	Male	0	9	16	1	1	26	1	2017	-	1
7	Male	0	14	12	1	1	18	2	2017	-	1
8	Male	0	16	3	1	1	20	2	2017	-	1
9	Male	0	17	10	1	2	12	2	2017	-	1
10	Male	0	17	12	1	2	12	2	2017	-	1
11	Male	0	20	8	1	1	3	2	2017	-	1
12	Male	0	20	10	1	2	3	2	2017	-	1
13	Male	0	22	16	1	1	15	3	2017	-	1
14	Male	0	27	11	1	1	8	2	2017	-	1
15	Male	0	28	6	1	1	25	2	2017	-	1
16	Male	0	34	10	1	1	18	3	2017	-	1
17	Male	0	38	5	1	1	9	3	2017	-	1

VARIABLES

In Stata variables can be recorded either as:

- **Numeric:** may contain only numbers (e.g. age, wage), or
- **String:** may contain letters or numbers referring to categories (e.g. education)

Values of string variables are included in double quotes:

```
generate men=1 if sex=="male"
```

Whereas values of numeric variables not

```
generate young=1 if age<=25
```

Variables may contain missing values:

- Missing values in string variables → empty double quotes: ""
- Missing values in numeric variables → a dot: .

VARIABLES

Some information when naming variables:

- Variable names can be up to **32** written characters long
- Nonetheless – for displaying purposes – a max of **10** characters is recommended to name the variables. (otherwise it will be shown as *high_educat~n*; shorten after the 10th character)
- The name can contain lower and uppercase letters, numbers and the underscore “_” character.
- Given that Stata is case sensitive (unlike SAS for instance), it is better to use lowercase. (**age** ≠ **Age**)
- The name cannot contain blank spaces or special characters (% ! ? , ; ; .)
- It has to start with a letter or underscore (not a number)

WRITING COMMANDS

- Stata is case sensitive: all commands are lowercase.
- Standard structure for commands:
`command varlist if/in, options`
- Commands can be abbreviated up to their underlined option; thus, command can be abbreviated as:
 - ✓ `comman`
 - ✓ `comma`
 - ✓ `comm`
 - ✓ `com`
 - ✓ But not!: `co`
- If you want to check the syntax to use for each command
`help command_name`



HELP WINDOW: help rename

```
Viewer - help rename
File Edit History Help
help rename
help rename x
Dialog | Also See | h

Title
[D] rename -- Rename variable

Syntax
rename old_varname new_varname

Menu
Data > Data utilities > Rename groups of variables

Description
rename changes the name of an existing variable old_varname to new_varname; the contents of the variable are unchanged. Also see [D] rename group for renaming groups of variables.

Examples
Setup
. webuse renamexmpl
. describe

Change name of exp to experience and change name of inc to income
. rename exp experience
. rename inc income



Describe the data
. describe
```



LOGICAL AND RELATIONAL OPERATORS

==	equal to
!=	not equal to
>	greater than
>=	greater or equal to
<	less than
<=	less or equal to
&	(logical) and
	(logical) or

EXAMINING THE DATA: SOME COMMANDS

<u>command</u>	use
<u>b</u> rowse	View raw data
<u>d</u> escribe	Produces a summary of the dataset
codebook var1 var2	Examines the variables names, labels and data
<u>s</u> ummarize	Calculates and displays a variety of univariate summary statistics
<u>t</u> abulate var1	Univariate frequency table
keep var1 var2	Keeps in memory only the mentioned variables or observations
 keep and drop commands are not reversible 	
order var1 var2	Relocates var1 and var2 to the beginning of the dataset in the order in which the variables are specified
count var1 if exp	Counts the number of observation of variable if expression is true

ORGANIZING THE DATA: generate and replace

<u>command</u>	use
<u>g</u> enerate newvar=exp [if][in]	Creates a new variable
<u>r</u> eplace oldvar=exp [if][in]	Replace contents of existing variable

Example:

Generate the categorical variable “working-age population” that takes the value 1 if the person’s age is greater or equal to 15, and 0 otherwise; name it “wap”. (age is stored in variable b5)

```
generate wap = .
      replace wap=0 if b5<15
      replace wap=1 if b5>=15 & b5!=.
```

Filtering → Can
be used for data
clean up

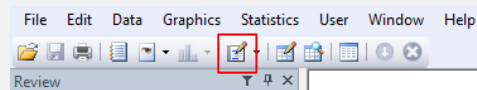
ORGANIZING THE DATA: LABELLING

- Labelling **variables** with descriptive names is useful and helps to better follow their meaning.
 - Labelling **values** of categorical variables ensures that the real-world meanings of the encodings are not forgotten.
- These points are crucial when sharing data with others, including yourself in the future.

command	use
<code>label variable var1 "my_var_name"</code>	Labels the variable
1. <code>label define set_of_val_label 0 "..." 1 "..." 2 "..."</code> 2. <code>label value var1 set_of_val_label</code>	1. Defines the set of value labels to each category. 2. Attaches the value labels previously defined to the values of var1

LET'S MOVE TO STATA V.13

1. Launch Stata version 13 by clicking the Stata icon on the desktop.
2. Open the do-file by pressing the do-file editor



3. Search in the `Introductory_Stata` folder and click on: "`Introductory_Session.do`"
4. .. Let's move to Stata

Change the variable type



- Numeric to string:
 - ToString var, force
- String to numeric:
 - Destring var, force

SOME STATA HELP:



- Stata website (www.stata.com)
- Help online
- Manuals:
 - Acock, A Gentle Introduction to Stata, 3rd Edition, Stata Press, 2010
 - Baum, An Introduction to Modern Econometrics Using Stata, 2006
 - Cameron and Trivedi, Microeconometrics Using Stata, revised edition, Stata Press 2010
- University of California resource Centre:
www.ats.ucla.edu/stat/stata



THANK YOU

www.ilo.org/ilostat