



## Regional Course on SDGs Indicators: Measuring decent work using microdata from labour force surveys

STATA basics - Getting started and Managing data

Chiba (Japan)  
30 September – 4 October 2019

## Objectives

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

- Preparing the participants in STATA basics (interphase and commands) for the next days' sessions.
- Doing some preliminary inspection and data manipulation using the Labour Force Survey (LFS) of Nepal 2017.

# 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) and the most used statistical software at the ITC-ILO campus.
- Other statistical software: SPSS, SAS, R, etc.

## Stata 13 screen

The screenshot displays the Stata 13 software interface. The main window is divided into several panes:

- Command Window (3. COMMANDS):** Located on the left, it shows the following commands:

```
1 use "C:\Users\vilarrreal-fo...  
2 summarize b5  
3 tabulate b4
```
- Results/Output Window (4. RESULTS/OUTPUT WINDOW):** The central pane displays the Stata logo and version information (13.1), copyright (1985-2013), and license details. Below this, it shows the output of the commands:

```
running W:\CONTENT\SOURCES\Applications\ILO\HQ\STATISTICS\STATA13\profile.do ...  
. use "C:\Users\vilarrreal-foentes\Desktop\MMR_LFS_2017Q1_ITC"  
( )  
. summarize b5  
+-----+-----+-----+-----+-----+  
Variable | Obs | Mean | Std. Dev. | Min | Max  
+-----+-----+-----+-----+-----+  
b5       | 58647 | 31.61602 | 19.86695 | 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 |
```
- Variables in the Dataset (1. VARIABLES IN THE DATASET):** A list on the right side of the window showing variables and their labels, such as b4 (Sex), b5 (Age), a1 (Sample FSU No.), a2 (Sample Household number), a3 (Quarter), a4 (No. of questionnaires used), a5 (Date of first interviewer visit - Day), a6 (Date of first interviewer visit - Month), a7 (Date of first interviewer visit - Year), a8 (Result of final visit), a9 (Data entry clerk number), a10 (Date of data entry - Day), a11 (Date of data entry - Month), a12 (Date of data entry - Year), a13 (Time of data entry - Hour), a14 (Time of data entry - Minute), a15 (Time of data entry - Hour), a16 (Time of end of data entry - Minute), and a17 (Number of household member).
- Properties Window (2. COMMAND WINDOW):** A small window at the bottom right showing the current variable's properties, such as Name (b4), Label (Sex), Type (byte), Format (%8.0g), and Value Label (94).

# Stata 13 screen

Stata 13.1 - C:\Users\willreal-fuentes\Desktop\MRR\_LFS\_2017Q1\_ITC.dta - [Results]

File Edit Data Graphics Statistics User Window Help

Review

# Command \_rc

```
1 use "C:\Users\willreal-fue..."
2 summarize b5
3 tabulate b4
```

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

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

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

running M:\CONTENT\SOURCE\Applications\ILO\HQ\STATISTICS\STAT13\profile.do ...

```
. use "C:\Users\willreal-fuentes\Desktop\MRR_LFS_2017Q1_ITC"
({})
```

```
. summarize b5
```

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

Command

Variables

Variable	Label
b4	Sex
b5	Age
a1	Sample FSU No.
a2	Sample Household number
a3	Quarter
a5	No. of questionnaires used
a4d	Date of first interviewer visit - Day
a4m	Date of first interviewer visit - Month
a4y	Date of first interviewer visit - Year
a15	Result of final visit
a5d	Data entry clerk number
a17d	Date of data entry - Day
a17m	Date of data entry - Month
a17y	Date of data entry - Year
a18h	Time of data entry - Hour
a18m	Time of data entry - Minute
a19h	Time of end of data entry - Hour
a19m	Time of end of data entry - Minute
x20	Number of household member

Properties

Variables

Name	b4
Label	Sex
Type	byte
Format	%8.0g
Value Label	B4

Data

Filename	MRR_LFS_2017Q1_ITC.dta
Label	
Notes	
Variables	110
Observations	58,647
Size	8,17M
Memory	64M
Sorted by	

ilostat ilo.org

5

# The Stata's toolbar

Stata/SE 13.1 - C:\Users\willreal-fuentes\Desktop\MRR\_LFS\_2017Q1\_ITC.dta - [Results]

File Edit Data Graphics Statistics User Window Help

1 2 3 4 5 6 7 8 9 10 11 12

Review

# Command \_rc

```
1 use "C:\Users\willreal-fue..."
2 summarize b5
3 tabulate b4
```

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

ilostat ilo.org

6

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):

```
1  * This is for one-line comment
2
3  // This is for one-line comment
4
5  /* This is for more than
6  One-line comment*/
```

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)

The screenshot shows the Stata Data Editor window with a dataset named 'MMR\_LFS\_2017Q1\_ITC.dta'. The main window displays a grid of data with columns labeled b4 through a16 and rows numbered 1 through 17. The first column (b4) contains the word 'Male' for all rows. The other columns contain numerical values. A red arrow points to the first row, and another red arrow points to the first column. On the right side, there is a 'Variables' list with checkboxes and labels for each variable, such as 'b4 Sex', 'b5 Age', 'a1 Sample FSU No.', etc.

## 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 10<sup>th</sup> 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)

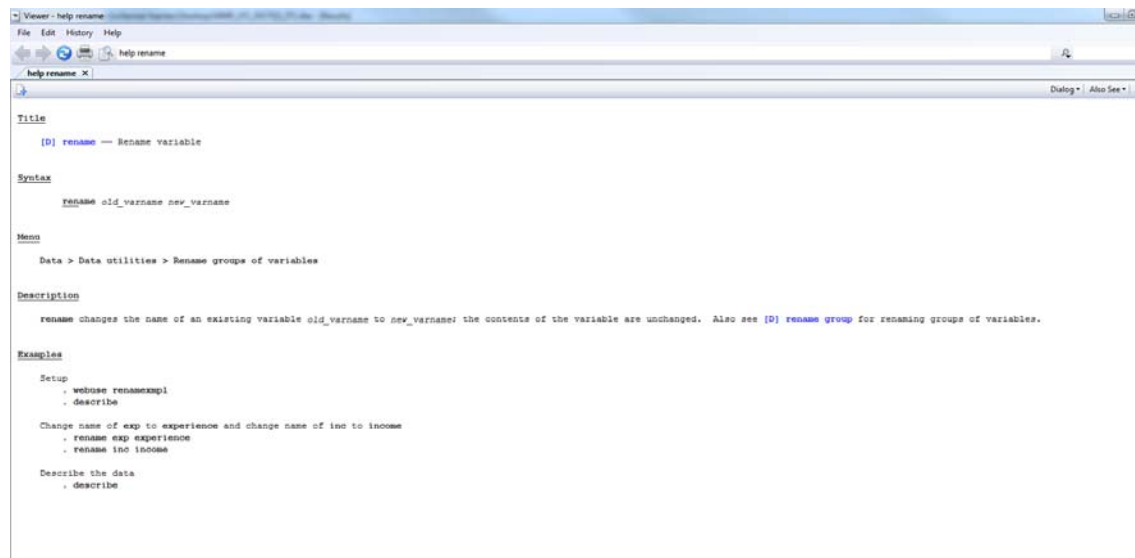
# Writing commands

- 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`



The screenshot shows a help window titled "Viewer - help rename". The window contains the following text:

```
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

<code>==</code>	equal to
<code>!=</code>	not equal to
<code>&gt;</code>	greater than
<code>&gt;=</code>	greater or equal to
<code>&lt;</code>	less than
<code>&lt;=</code>	less or equal to
<code>&amp;</code>	(logical) and
<code> </code>	(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 b02)

```
generate wap =.  
    replace wap=0 if b02<15  
    replace wap=1 if b02>=15 & b02!=.
```

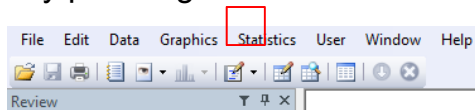
# Organizing the data: labelling

- 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.

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

## Let's move to Stata !

1. Launch your Stata version by clicking on 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

## Some Stata help can be found on:

- Stata website ([www.stata.com](http://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](http://www.ats.ucla.edu/stat/stata)