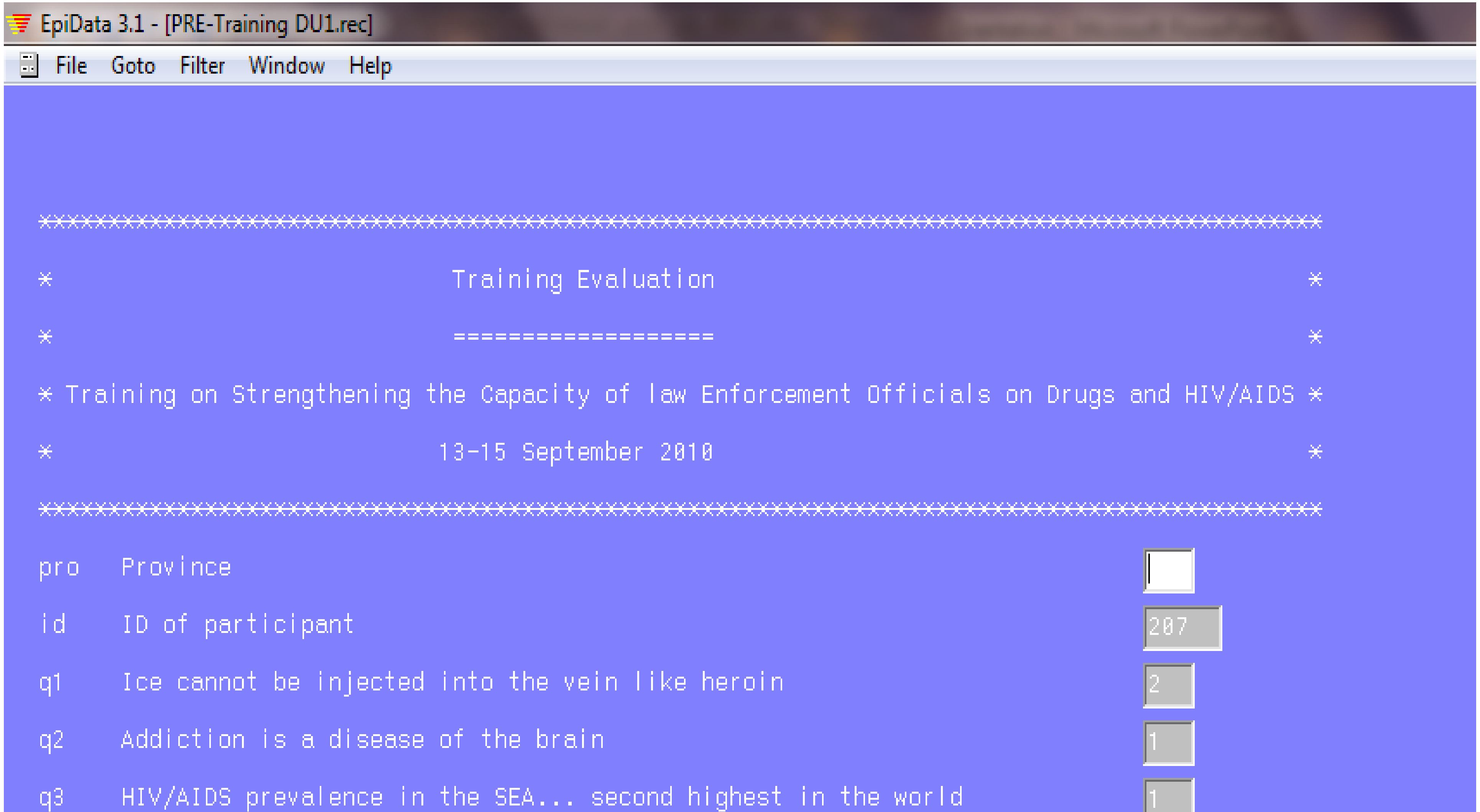
Introduction of STATA program

Prepared by Associate prof. Prum Virak

Research Skills

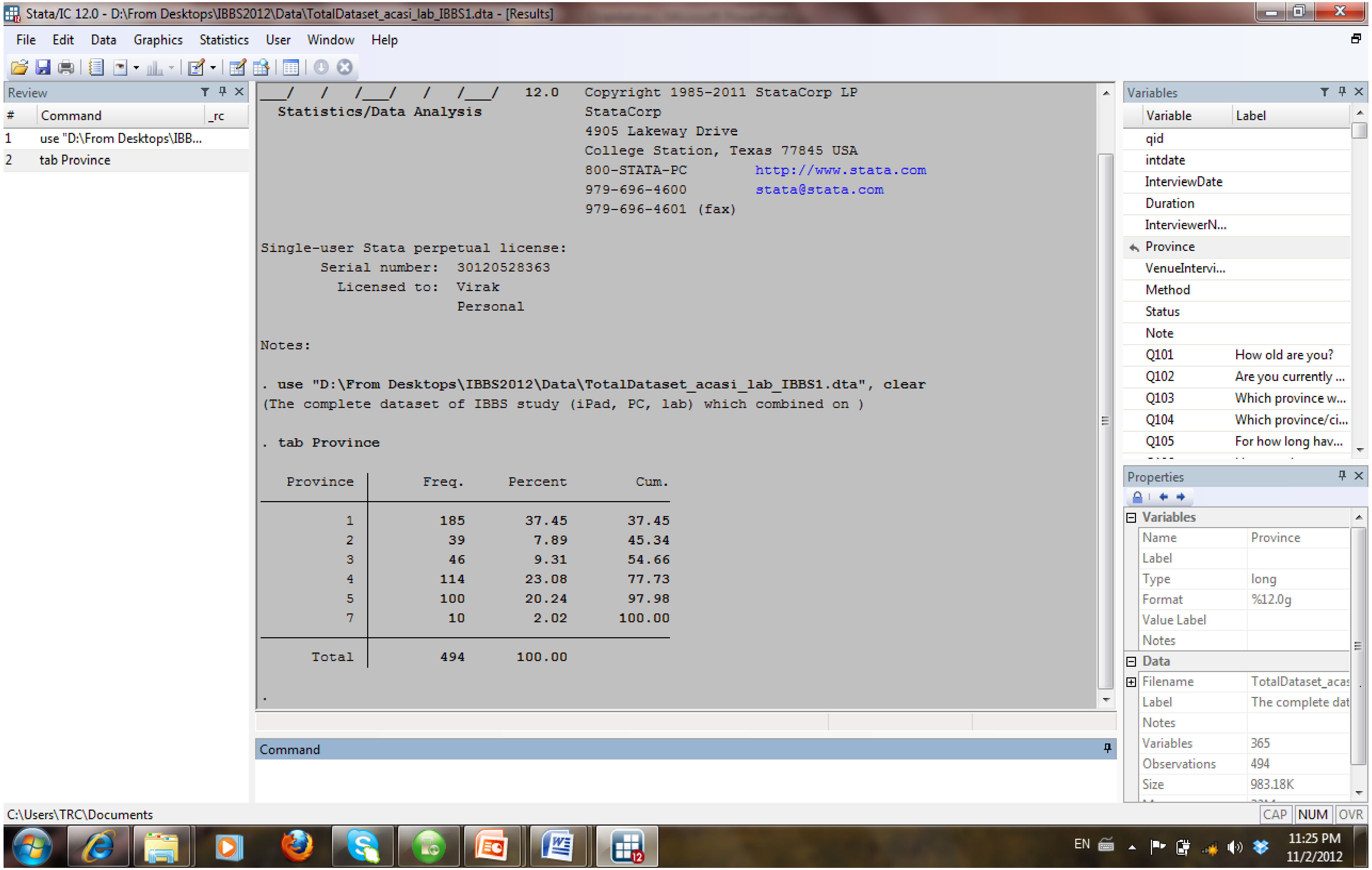
Proposal writing	• Interest/Idea/theory • Conceptualization/Research Method/Population & sampling
Data collection	• Interviewer, field team leader, coordinator
Data processing	• Data frame design and Entering collected question by using Epi-Data
Data management & analysis	• labeling, coding, generating, analyzing.... by using STATA, SPSS
Report writing	• More international standard report writing
Presentation	• More effective and convenience presentation

Epi-Data program is used to enter the collected information (in Questionnaire) and transferring data.



```
*********************************************************************************************
*                             Training Evaluation                                           *
*                             ===================                                           *
* Training on Strengthening the Capacity of law Enforcement Officials on Drugs and HIV/AIDS *
*                            13-15 September 2010                                           *
*********************************************************************************************
pro   Province                                                                  |
id    ID of participant                                                         207
q1    Ice cannot be injected into the vein like heroin                          2
q2    Addiction is a disease of the brain                                       1
q3    HIV/AIDS prevalence in the SEA... second highest in the world             1
```

STATA program is used to manage and to analyze data



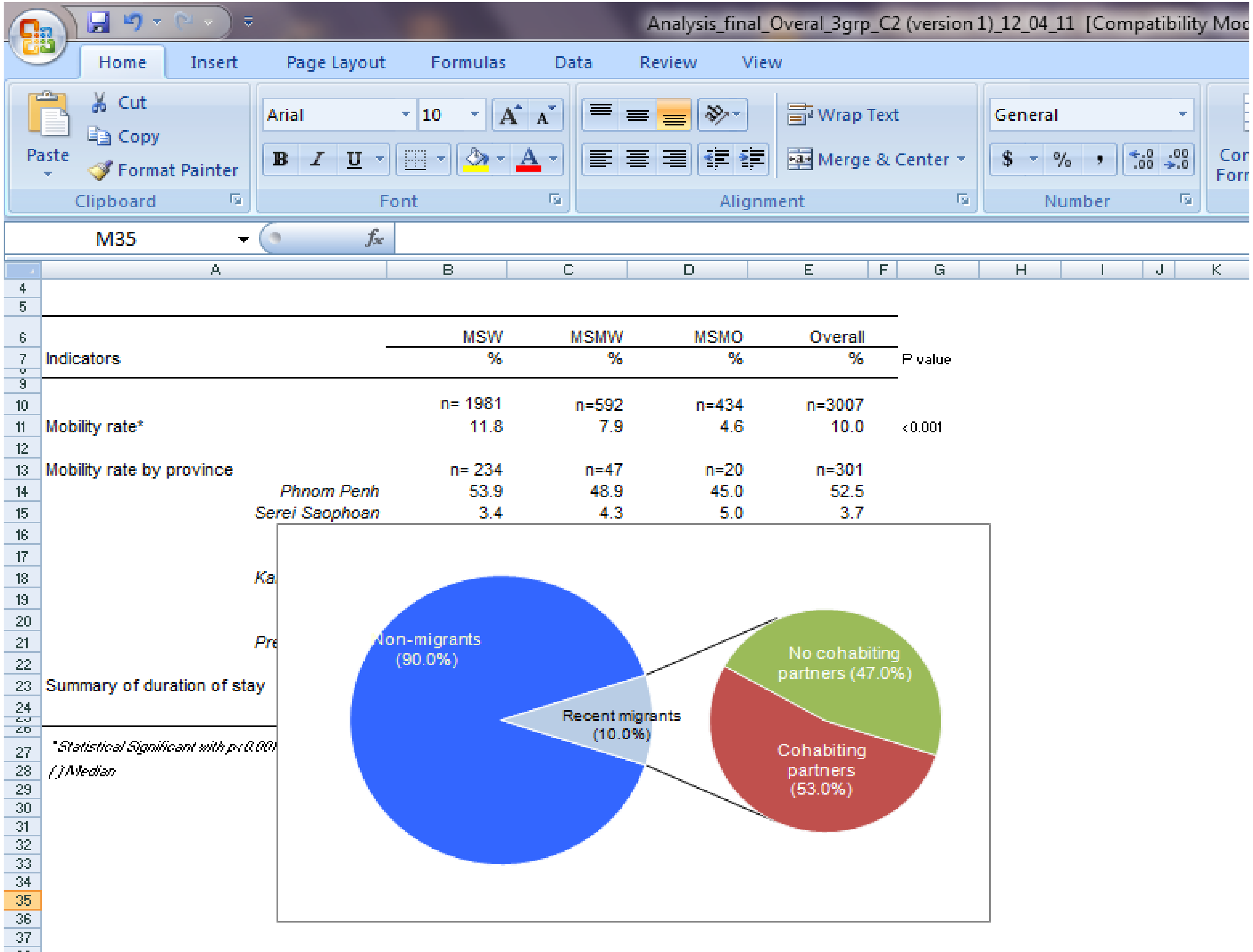
#	Command	_rc
1	use "D:\From Desktops\IBB...	
2	tab Province	

```
___/   /   /___/   /   /___/   12.0   Copyright 1985-2011 StataCorp LP
  Statistics/Data Analysis            StataCorp
                                      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)

Single-user Stata perpetual license:
       Serial number:  30120528363
         Licensed to:  Virak
                       Personal

Notes:

. use "D:\From Desktops\IBBS2012\Data\TotalDataset_acasi_lab_IBBS1.dta", clear
(The complete dataset of IBBS study (iPad, PC, lab) which combined on )

. tab Province

   Province |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |        185       37.45       37.45
          2 |         39        7.89       45.34
          3 |         46        9.31       54.66
          4 |        114       23.08       77.73
          5 |        100       20.24       97.98
          7 |         10        2.02      100.00
------------+-----------------------------------
      Total |        494      100.00

.
```

Variable	Label
qid	
intdate	
InterviewDate	
Duration	
InterviewerN...	
Province	
VenueIntervi...	
Method	
Status	
Note	
Q101	How old are you?
Q102	Are you currently ...
Q103	Which province w...
Q104	Which province/ci...
Q105	For how long hav...

Variables	
Name	Province
Label	
Type	long
Format	%12.0g
Value Label	
Notes	
Data	
Filename	TotalDataset_acas
Label	The complete dat
Notes	
Variables	365
Observations	494
Size	983.18K

Excel program is used to present the data analysis result



	MSW	MSMW	MSMO	Overall	
Indicators	%	%	%	%	P value
	n= 1981	n=592	n=434	n=3007	
Mobility rate*	11.8	7.9	4.6	10.0	<0.001
Mobility rate by province	n= 234	n=47	n=20	n=301	
Phnom Penh	53.9	48.9	45.0	52.5	
Serei Saophoan	3.4	4.3	5.0	3.7	
Ka					
Pre					
Summary of duration of stay					

**Statistical Significant with p<0.001*

( ) Median

Expectation

- **Know** how to process, manage and to analyze data by using STATA program.
- **Understand** the importance of data cleaning before starting data analyses.
- **Be able not only to create** convenient graphs or tables in Excel program but also earn the basic knowledge of how to translate the produced result to program activity or real life.

INTROCTION

TO

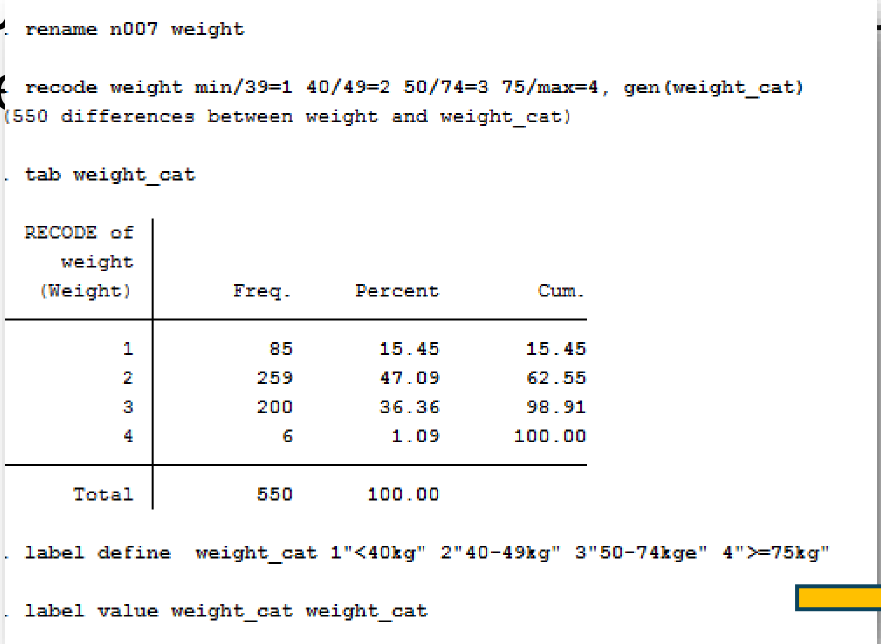
STATA PROGRAM

Why use STATA?

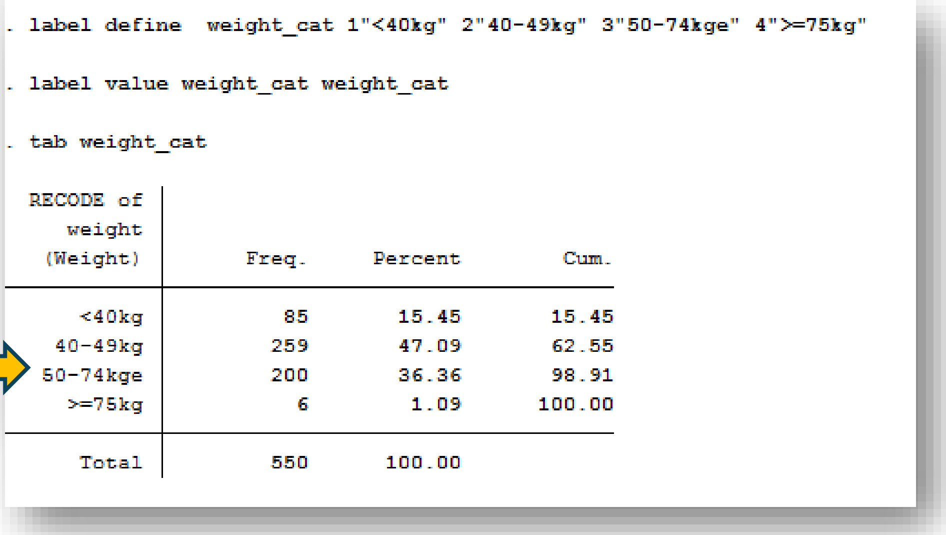
- A command-driven package especially for statistic analyses and data management.
- Easy to use for both simple and complicated tasks.

Data management

Data management: STATA program can **assist you more effectively in preparing research data before doing data analysis** such as *labeling variables, labeling codes of variables, recoding variables, generating a new variable, finding and correcting data error e*



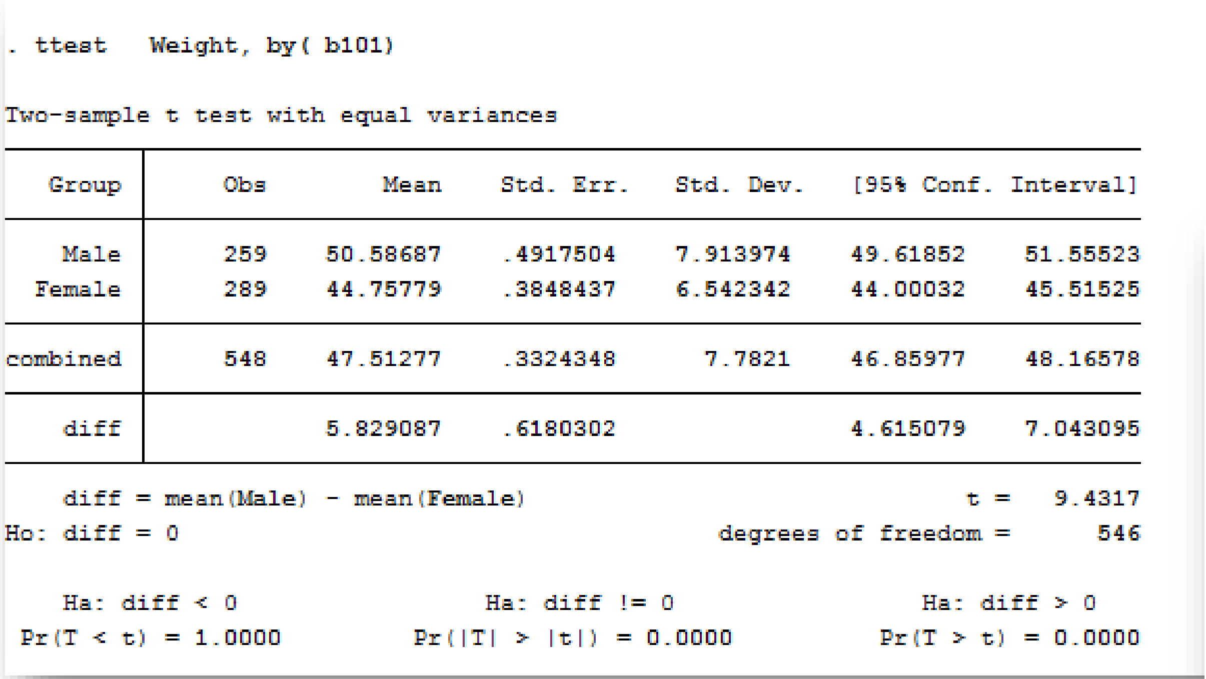
```
. rename n007 weight

. recode weight min/39=1 40/49=2 50/74=3 75/max=4, gen(weight_cat)
(550 differences between weight and weight_cat)

. tab weight_cat

  RECODE of |
     weight |
   (Weight) |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |         85       15.45       15.45
          2 |        259       47.09       62.55
          3 |        200       36.36       98.91
          4 |          6        1.09      100.00
------------+-----------------------------------
      Total |        550      100.00

. label define  weight_cat 1"<40kg" 2"40-49kg" 3"50-74kge" 4">=75kg"

. label value weight_cat weight_cat
```



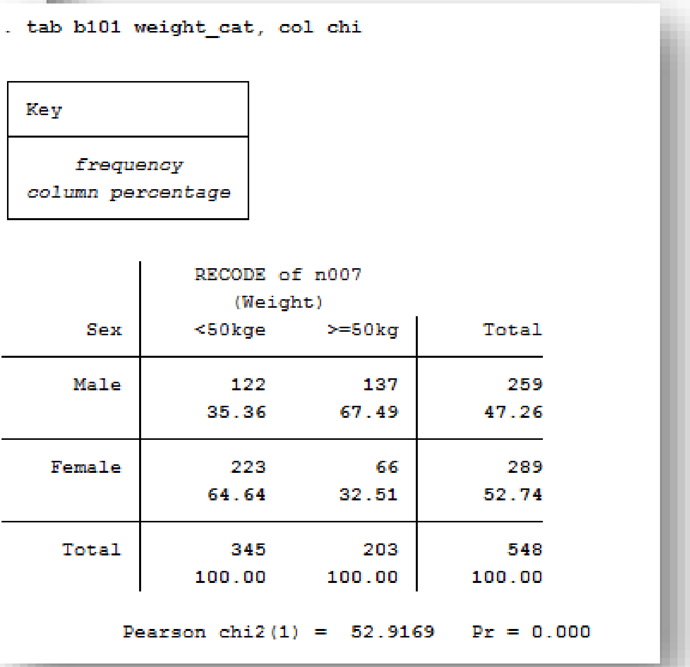
```
. label define  weight_cat 1"<40kg" 2"40-49kg" 3"50-74kge" 4">=75kg"

. label value weight_cat weight_cat

. tab weight_cat

  RECODE of |
     weight |
   (Weight) |      Freq.     Percent        Cum.
------------+-----------------------------------
      <40kg |         85       15.45       15.45
    40-49kg |        259       47.09       62.55
   50-74kge |        200       36.36       98.91
     >=75kg |          6        1.09      100.00
------------+-----------------------------------
      Total |        550      100.00
```

Statistical Analysis

- STATA program can **assist you more effectively in analyzing your res**



```
. ttest   Weight, by( b101)

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
    Male |     259    50.58687    .4917504    7.913974    49.61852    51.55523
  Female |     289    44.75779    .3848437    6.542342    44.00032    45.51525
---------+--------------------------------------------------------------------
combined |     548    47.51277    .3324348      7.7821    46.85977    48.16578
---------+--------------------------------------------------------------------
    diff |            5.829087    .6180302                4.615079    7.043095
------------------------------------------------------------------------------
    diff = mean(Male) - mean(Female)                              t =   9.4317
Ho: diff = 0                                     degrees of freedom =      546

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000
```



```
. tab b101 weight_cat, col chi

+-------------------+
| Key               |
|-------------------|
|     frequency     |
| column percentage |
+-------------------+

           |    RECODE of n007
           |       (Weight)
       Sex |    <50kge     >=50kg |     Total
-----------+----------------------+----------
      Male |       122        137 |       259 
           |     35.36      67.49 |     47.26 
-----------+----------------------+----------
    Female |       223         66 |       289 
           |     64.64      32.51 |     52.74 
-----------+----------------------+----------
     Total |       345        203 |       548 
           |    100.00     100.00 |    100.00 

          Pearson chi2(1) =  52.9169   Pr = 0.000
```

STATA types

There are many types of STATA packages have designed for students or professional researchers:

- **Stata/MP: It is the fastest and largest version of Stata.** Most computers purchased since mid 2006 can take advantage of the advanced multiprocessing of Stata/MP. This includes the Intel Core^{TM} 2 Duo, i3, i5, i7, and the AMD X2 dual-core chips.

STATA types

- **Small Stata**: *A version of Stata that handles small datasets (for educational purchases only)*. It is limited to analyzing data sets with a maximum of 99 variables on approximately 1,200 observations.
- **Stata/IC:** *Stata for moderate-sized datasets*. It allows data sets with as many as 2,047 variables. The number of observations is limited only by the amount of RAM in your computer.
- **Stata/SE:** *Stata for large datasets*. It allows datasets with up to 32,767 variables. The number of observations is limited only by the amount of RAM in your computer.

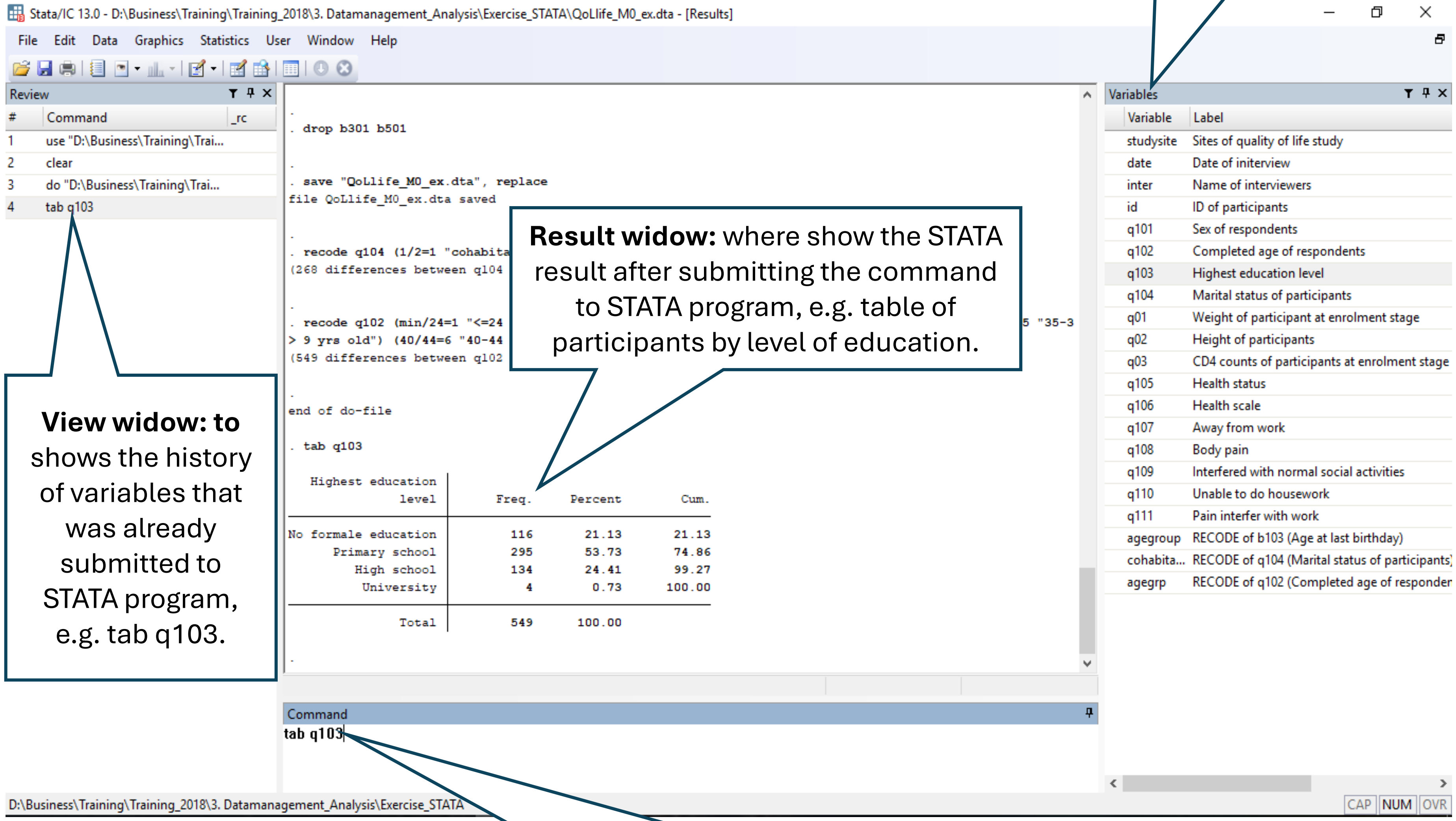
Getting start

STATA program

- A command-driven package, easy to use for both simple and complicated task for statistic analyses
- **Data management:** *Assist more effectively in preparing research data such as labeling variables, labeling codes of variables, recoding variables, generating a new variable,* ***finding and correcting data error*** *etc.*
- **Data analysis**: *Assist more effectively in analyzing research dataset.*
- **Graphic**: *Assist you to create graphs for presenting your data.*

4 Main Windows of STATA Program

Variable widow: shows the list of variables and the labels of variables



Stata/IC 13.0 - D:\Business\Training\Training_2018\3. Datamanagement_Analysis\Exercise_STATA\QoLlife_M0_ex.dta - [Results]
File Edit Data Graphics Statistics User Window Help
Review
Variables
Command
D:\Business\Training\Training_2018\3. Datamanagement_Analysis\Exercise_STATA
CAP NUM OVR

#	Command	_rc
1	use "D:\Business\Training\Trai...	
2	clear	
3	do "D:\Business\Training\Trai...	
4	tab q103	

```
.
. drop b301 b501

.
. save "QoLlife_M0_ex.dta", replace
file QoLlife_M0_ex.dta saved

.
. recode q104 (1/2=1 "cohabita
(268 differences between q104

.
. recode q102 (min/24=1 "<=24                                                5 "35-3
> 9 yrs old") (40/44=6 "40-44
(549 differences between q102

.
end of do-file

. tab q103

  Highest education |
              level |      Freq.     Percent        Cum.
--------------------+-----------------------------------
No formale education |        116       21.13       21.13
     Primary school |        295       53.73       74.86
        High school |        134       24.41       99.27
         University |          4        0.73      100.00
--------------------+-----------------------------------
              Total |        549      100.00

.
```

Variable	Label
studysite	Sites of quality of life study
date	Date of interview
inter	Name of interviewers
id	ID of participants
q101	Sex of respondents
q102	Completed age of respondents
q103	Highest education level
q104	Marital status of participants
q01	Weight of participant at enrolment stage
q02	Height of participants
q03	CD4 counts of participants at enrolment stage
q105	Health status
q106	Health scale
q107	Away from work
q108	Body pain
q109	Interfered with normal social activities
q110	Unable to do housework
q111	Pain interfer with work
agegroup	RECODE of b103 (Age at last birthday)
cohabita...	RECODE of q104 (Marital status of participants)
agegrp	RECODE of q102 (Completed age of responder

Command:

```
tab q103
```

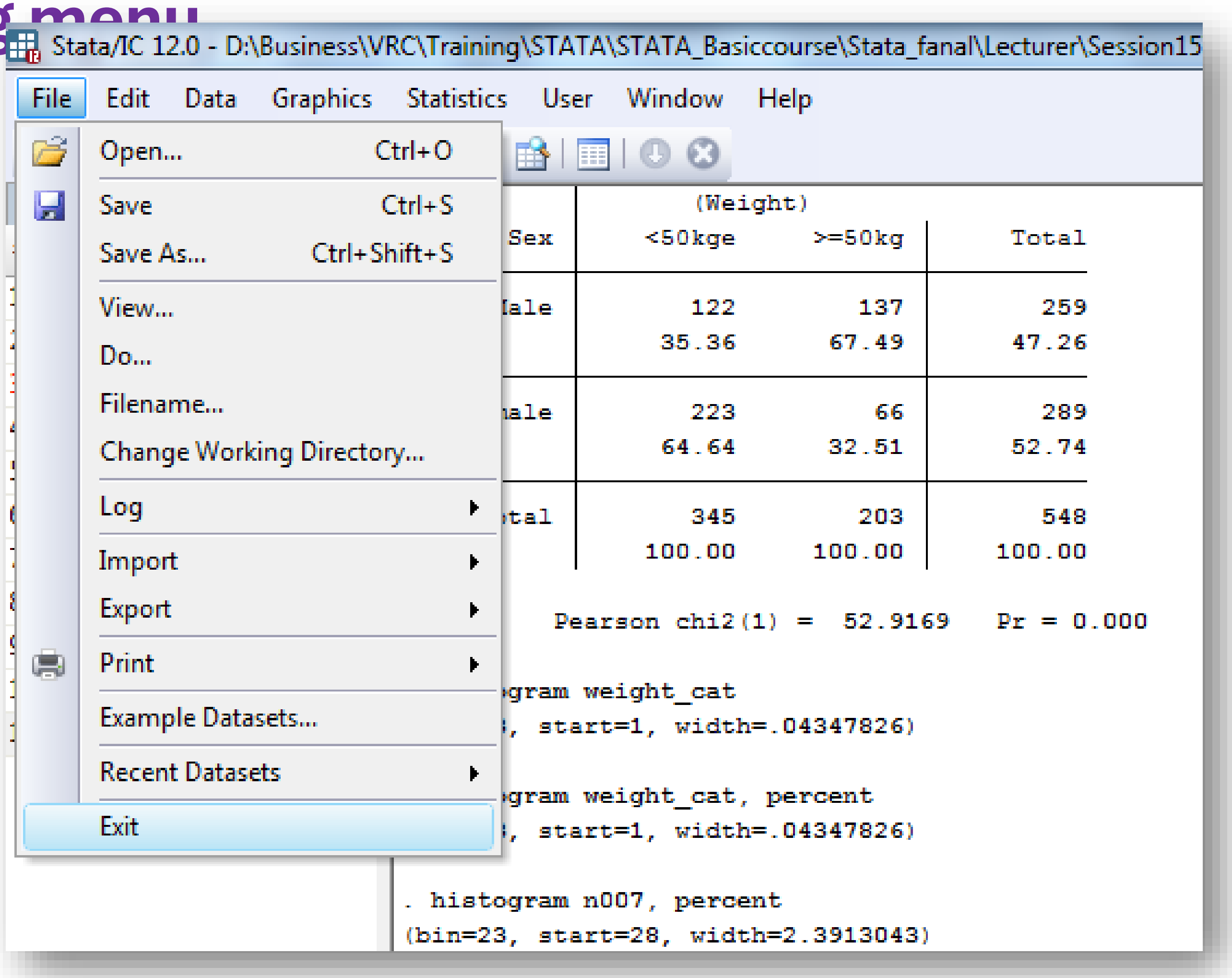
View widow: to shows the history of variables that was already submitted to STATA program, e.g. tab q103.

Result widow: where show the STATA result after submitting the command to STATA program, e.g. table of participants by level of education.

Command widow: where a command is submitted to STATA program, e.g. tab q103

Exiting STATA by using menu

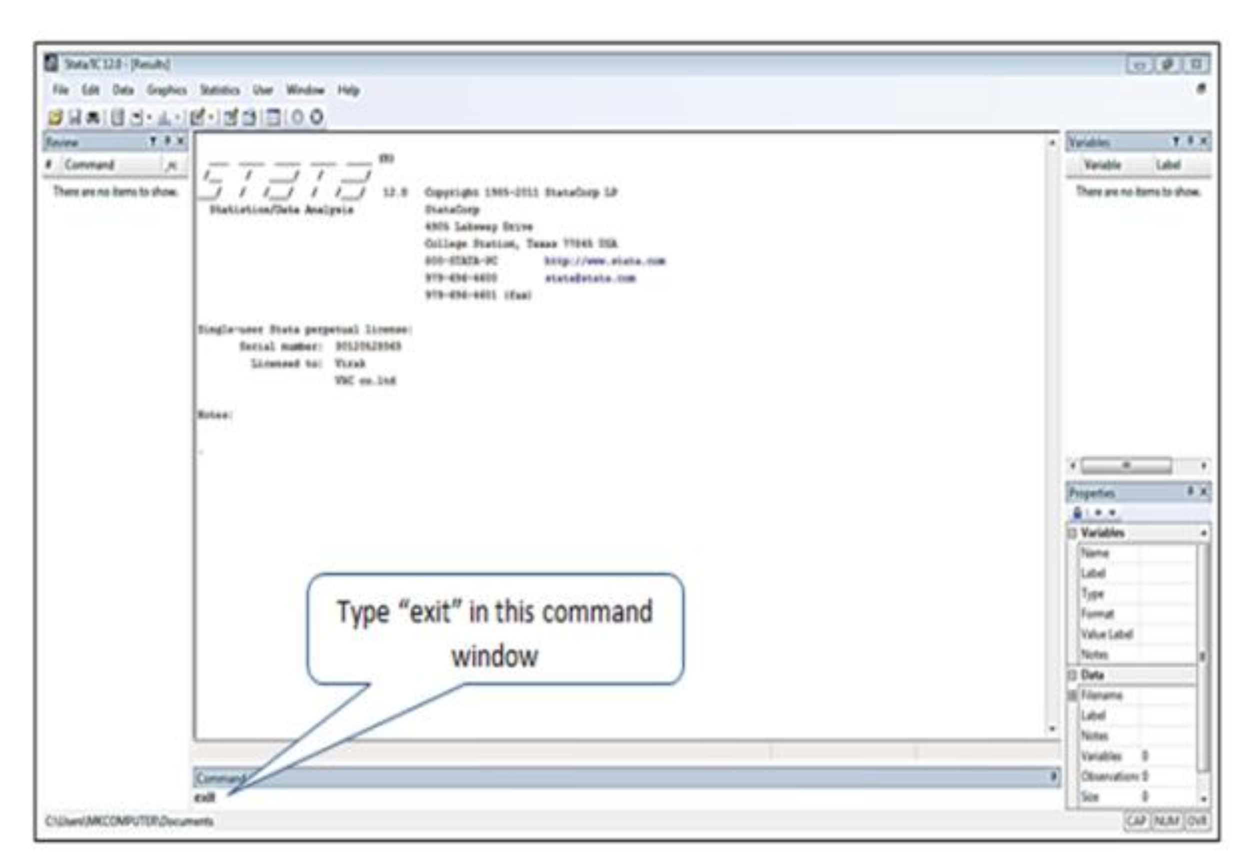
Click on **File** ⇨ then



```
           |      (Weight)
       Sex |  <50kge    >=50kg |     Total
-----------+-------------------+----------
      Male |     122       137 |       259
           |   35.36     67.49 |     47.26
-----------+-------------------+----------
      Female |     223        66 |       289
           |   64.64     32.51 |     52.74
-----------+-------------------+----------
     Total |     345       203 |       548
           |  100.00    100.00 |    100.00

          Pearson chi2(1) =  52.9169   Pr = 0.000

. histogram weight_cat
(bin=23, start=1, width=.04347826)

. histogram weight_cat, percent
(bin=23, start=1, width=.04347826)

. histogram n007, percent
(bin=23, start=28, width=2.3913043)
```

Exiting STATA by using command

… type **exit** in the **command Window** and then **Press Enter**



```
  ___  ____  ____  ____  ____ (R)
 /__    /   ____/   /   ____/
___/   /   /___/   /   /___/   12.0   Copyright 1985-2011 StataCorp LP
  Statistics/Data Analysis            StataCorp
                                      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)

Single-user Stata perpetual license:
       Serial number:  [illegible]
         Licensed to:  Virak
                       VNC co.ltd

Notes:

.
```

```
exit
```

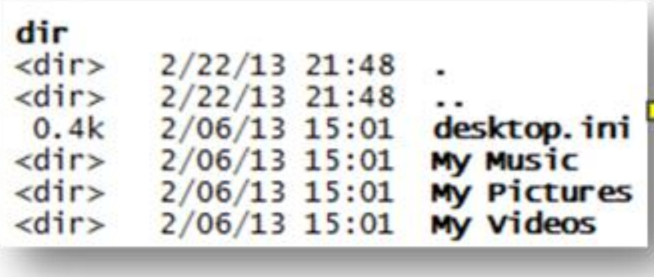
Type "exit" in this command window

How to create your directory by Stata Window commands?

Ex: I want to create a folder, which names **“practice”** in drive D by STATA command.

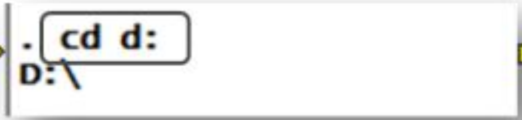
Make directory

Content check



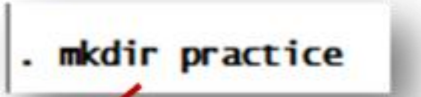
```
dir
<dir>     2/22/13 21:48   .
<dir>     2/22/13 21:48   ..
 0.4k     2/06/13 15:01   desktop.ini
<dir>     2/06/13 15:01   My Music
<dir>     2/06/13 15:01   My Pictures
<dir>     2/06/13 15:01   My Videos
```

Chang from drive C:\
to drive D:\

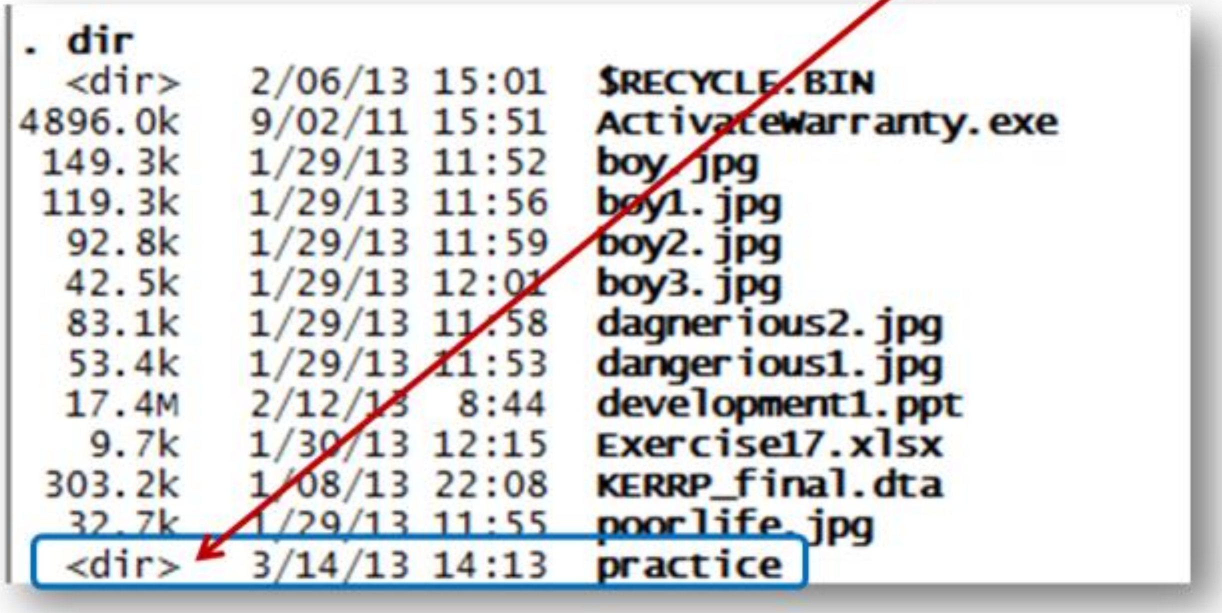


```
. cd d:
D:\
```

Creat directory
“practice”



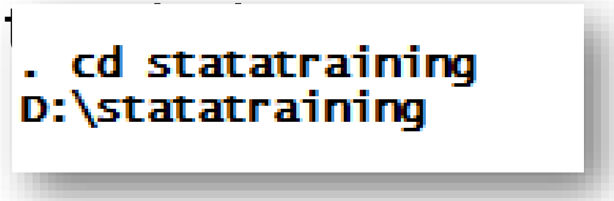
```
. mkdir practice
```



```
. dir
  <dir>     2/06/13 15:01   $RECYCLE.BIN
4896.0k     9/02/11 15:51   ActivateWarranty.exe
 149.3k     1/29/13 11:52   boy.jpg
 119.3k     1/29/13 11:56   boy1.jpg
  92.8k     1/29/13 11:59   boy2.jpg
  42.5k     1/29/13 12:01   boy3.jpg
  83.1k     1/29/13 11:58   dagnerious2.jpg
  53.4k     1/29/13 11:53   dangerious1.jpg
  17.4M     2/12/13  8:44   development1.ppt
   9.7k     1/30/13 12:15   Exercise17.xlsx
 303.2k     1/08/13 22:08   KERRP_final.dta
  32.7k     1/29/13 11:55   poorlife.jpg
  <dir>     3/14/13 14:13   practice
```

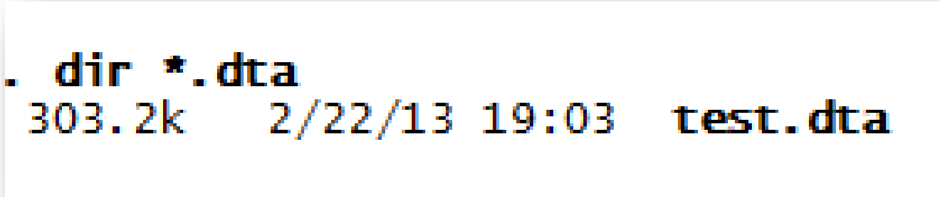
Directory content checking

I get into **statatraining** directory by typing **CD statatraining** ⇨ press **see** pict



```
. cd statatraining
D:\statatraining
```

Then, I type **dir *.dta** ⇨ press **enter** ...to see **<u>all Stata files</u>** in this director. See picture below

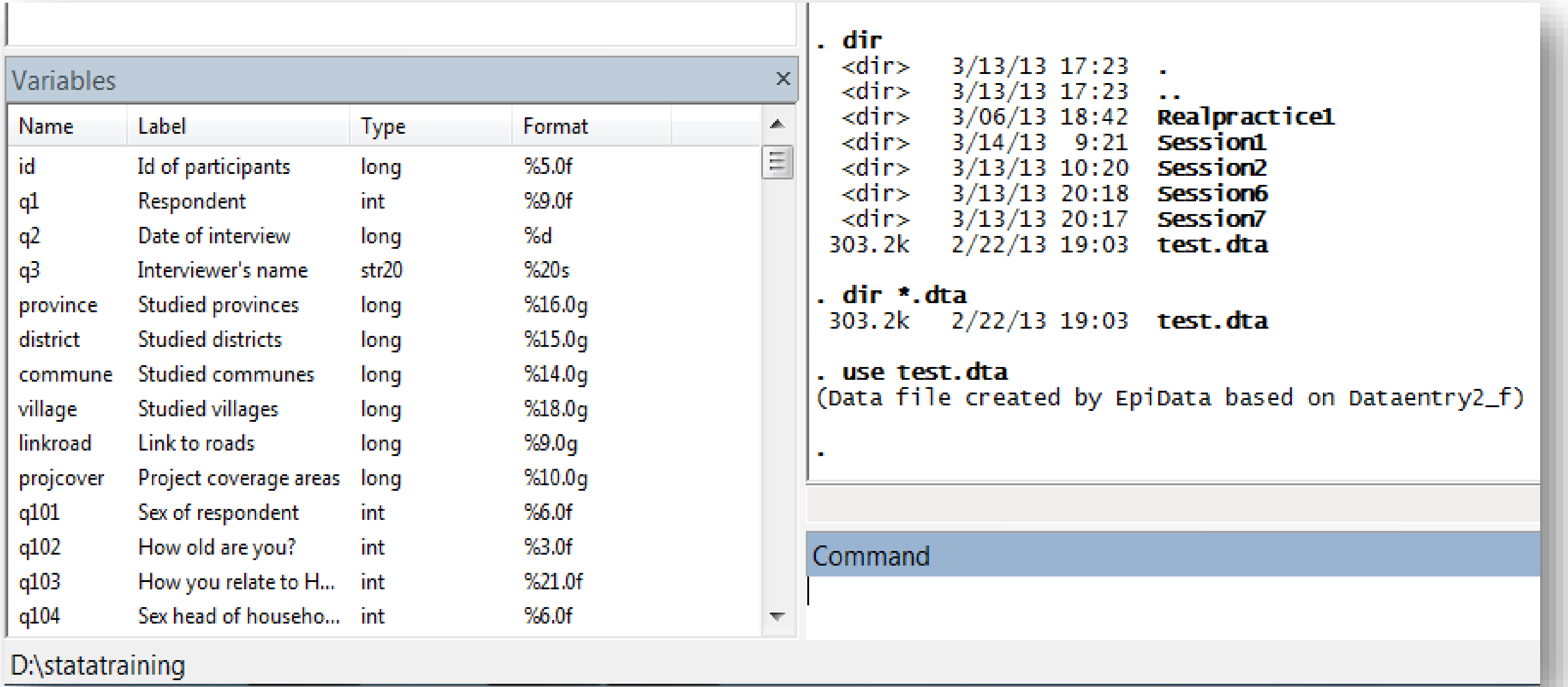


```
. dir *.dta
 303.2k   2/22/13 19:03   test.dta
```

“use” command

It is used to open a dataset e.g test.dta.

…type **use test.dta** ⇨ press **enter.** See pictures below



Name	Label	Type	Format
id	Id of participants	long	%5.0f
q1	Respondent	int	%9.0f
q2	Date of interview	long	%d
q3	Interviewer's name	str20	%20s
province	Studied provinces	long	%16.0g
district	Studied districts	long	%15.0g
commune	Studied communes	long	%14.0g
village	Studied villages	long	%18.0g
linkroad	Link to roads	long	%9.0g
projcover	Project coverage areas	long	%10.0g
q101	Sex of respondent	int	%6.0f
q102	How old are you?	int	%3.0f
q103	How you relate to H...	int	%21.0f
q104	Sex head of househo...	int	%6.0f

```
. dir
  <dir>   3/13/13 17:23  .
  <dir>   3/13/13 17:23  ..
  <dir>   3/06/13 18:42  Realpractice1
  <dir>   3/14/13  9:21  Session1
  <dir>   3/13/13 10:20  Session2
  <dir>   3/13/13 20:18  Session6
  <dir>   3/13/13 20:17  Session7
 303.2k   2/22/13 19:03  test.dta

. dir *.dta
 303.2k   2/22/13 19:03  test.dta

. use test.dta
(Data file created by EpiData based on Dataentry2_f)

.
```

In case you know the path and name of your file,

- e.g. **use "D:\statatraining\test.dta”** ⇨ **press enter**

Setting your file path

How to Set your file path

Should always have a **unique directory** for each specific project or piece of work you do.

- E.g. **D:\statatraining\test.dta**

How to Set your file path

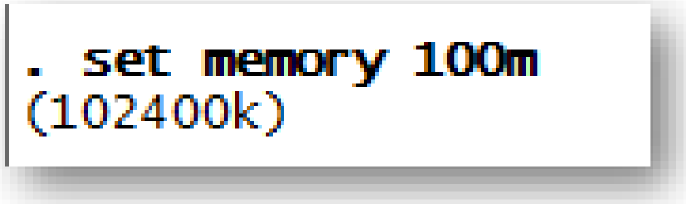
Sometimes you have problems with setting your path: directory or file name have spaces in them. So you have to put the entire path name in quotation marks “..”

E.g. I have one file names “**test.dta**” in my directory statatraining.

use "D:\statatraining\test.dta” ⇨ press **enter**

How to increase memory allocation

- When you work with large datasets the amount of memory allocated to STATA may need to be changed.
- Have to increase the memory before reading in any data. If you need to increase memory 10 megabytes to read your dataset… by typing , e.g., **set memory 100m** ⇨ press **enter**, see picture below.



```
. set memory 100m
(102400k)
```

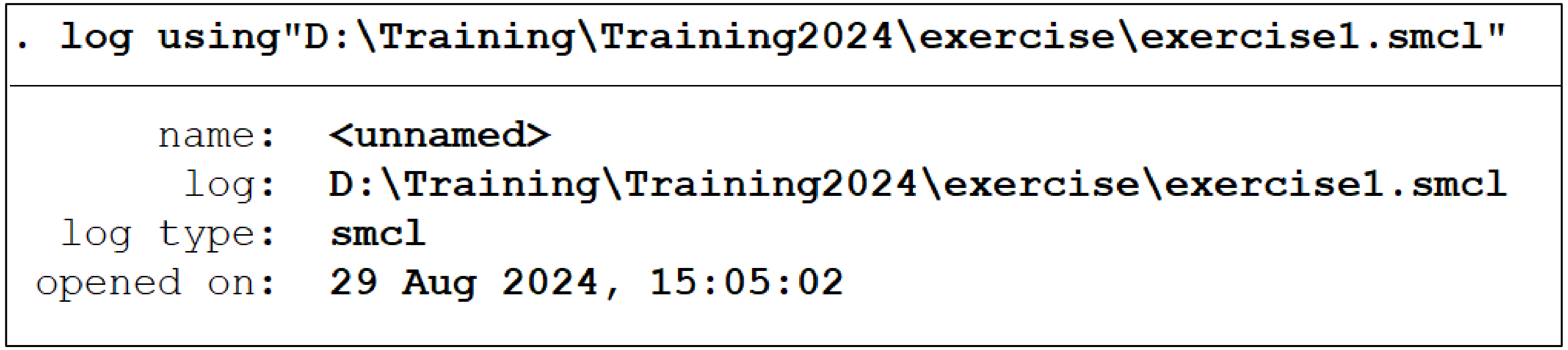
How to create a new log file (Save your outputs)

You should always open a log file to save the work you do and the output you produce...you can create a log file by using commands and by menu.

Command: log using *“path and file name”* ⇨ press enter.

E.g. log using"D:\Training\Training2024\exercise\exercise1.smcl”

⇨ press enter



```
. log using"D:\Training\Training2024\exercise\exercise1.smcl"
--------------------------------------------------------------

      name:  <unnamed>
       log:  D:\Training\Training2024\exercise\exercise1.smcl
  log type:  smcl
 opened on:  29 Aug 2024, 15:05:02
```

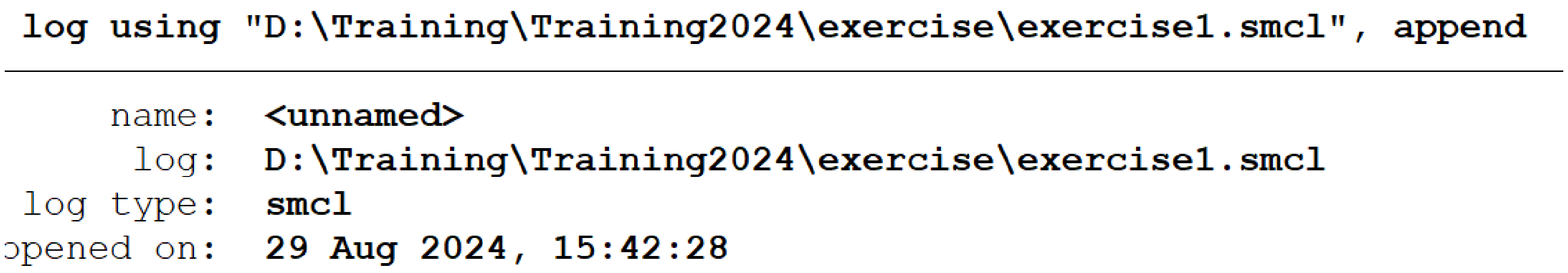
How to open your outputs

Command: log using *“path and file name”* , append ⇨ press enter.

E.g. typing

log using "D:\Training\Training2024\exercise\exercise1.smcl", append

⇨ press enter



```
  log using "D:\Training\Training2024\exercise\exercise1.smcl", append
--------------------------------------------------------------------------
      name:  <unnamed>
       log:  D:\Training\Training2024\exercise\exercise1.smcl
  log type:  smcl
 opened on:  29 Aug 2024, 15:42:28
```

- close command log, temporarily suspend logging, or resume logging
  - log close
  - log of,
  - log on

WORKING WITH DO FILE

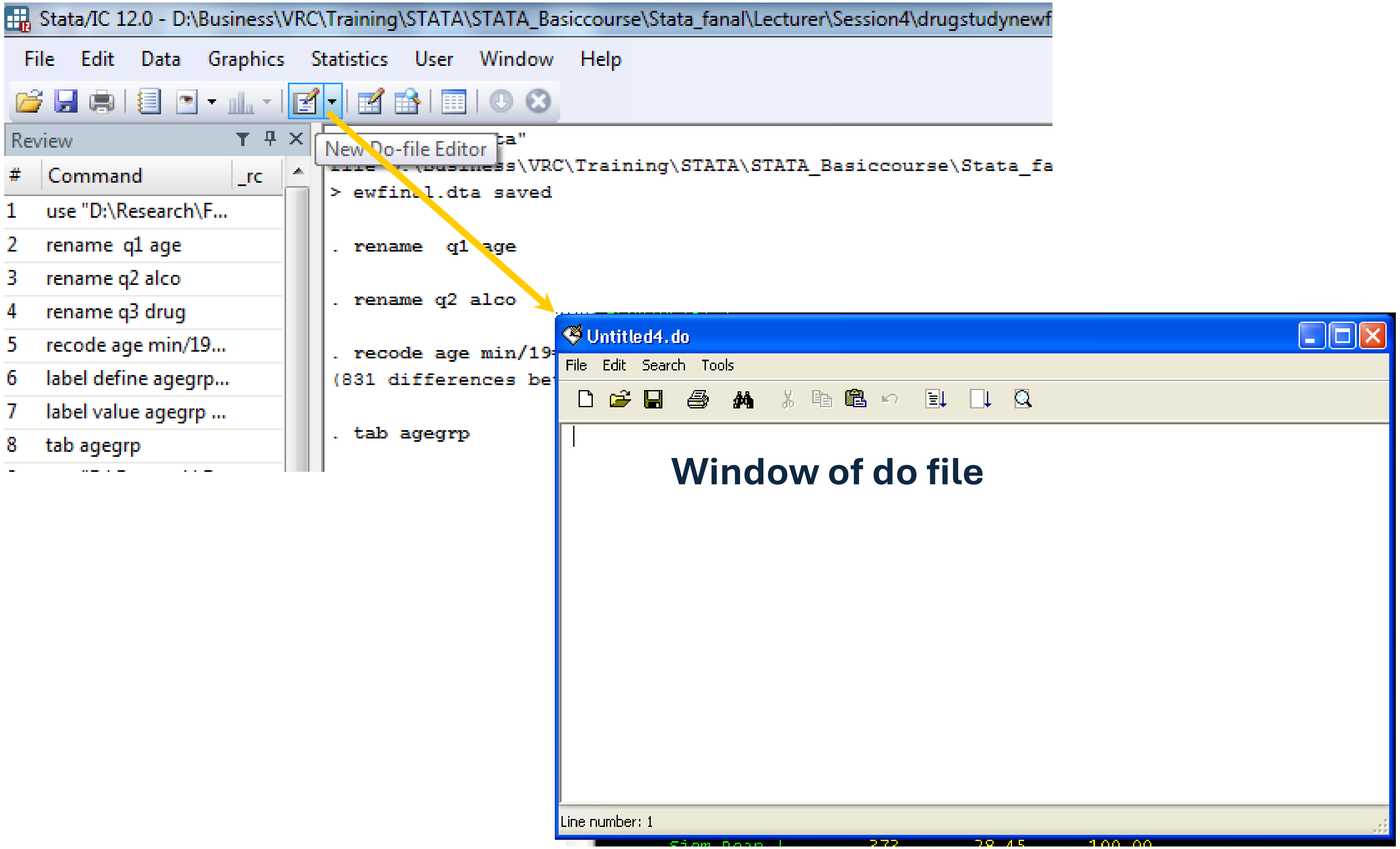
Use do file

- Should use ***do file*** for all your data processing and analysis, this ensures you have a record of all the work you do and you can easily re-run any analysis and correct inevitable errors in recoding, and others can reference and make use of the program
- Issuing a series of commands from a program file.

Creating do file

- Creating do file in the “ Do file editor” from the Window menu at the top of the STATA screen.
- All STATA program file must have the extension .do to be recognized as a program file.

Click on new Do-file Editor



#	Command	_rc
1	use "D:\Research\F...	
2	rename q1 age	
3	rename q2 alco	
4	rename q3 drug	
5	recode age min/19...	
6	label define agegrp...	
7	label value agegrp ...	
8	tab agegrp	

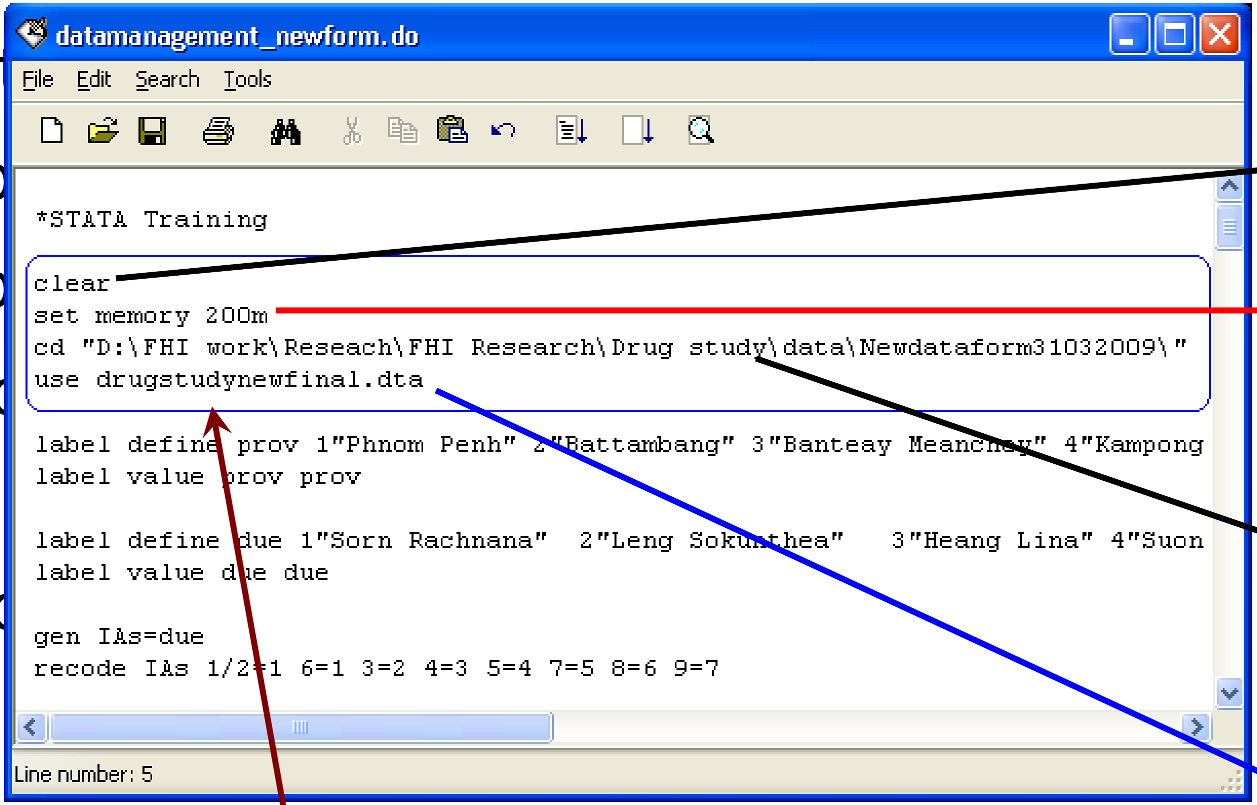
```
                          a"
file D:\Business\VRC\Training\STATA\STATA_Basiccourse\Stata_fa
> ewfinal.dta saved

. rename  q1 age

. rename q2 alco

. recode age min/19
(831 differences be

. tab agegrp
```

Window of do file

Start working with do file

You t

- To
- To data 200M
- To
- To

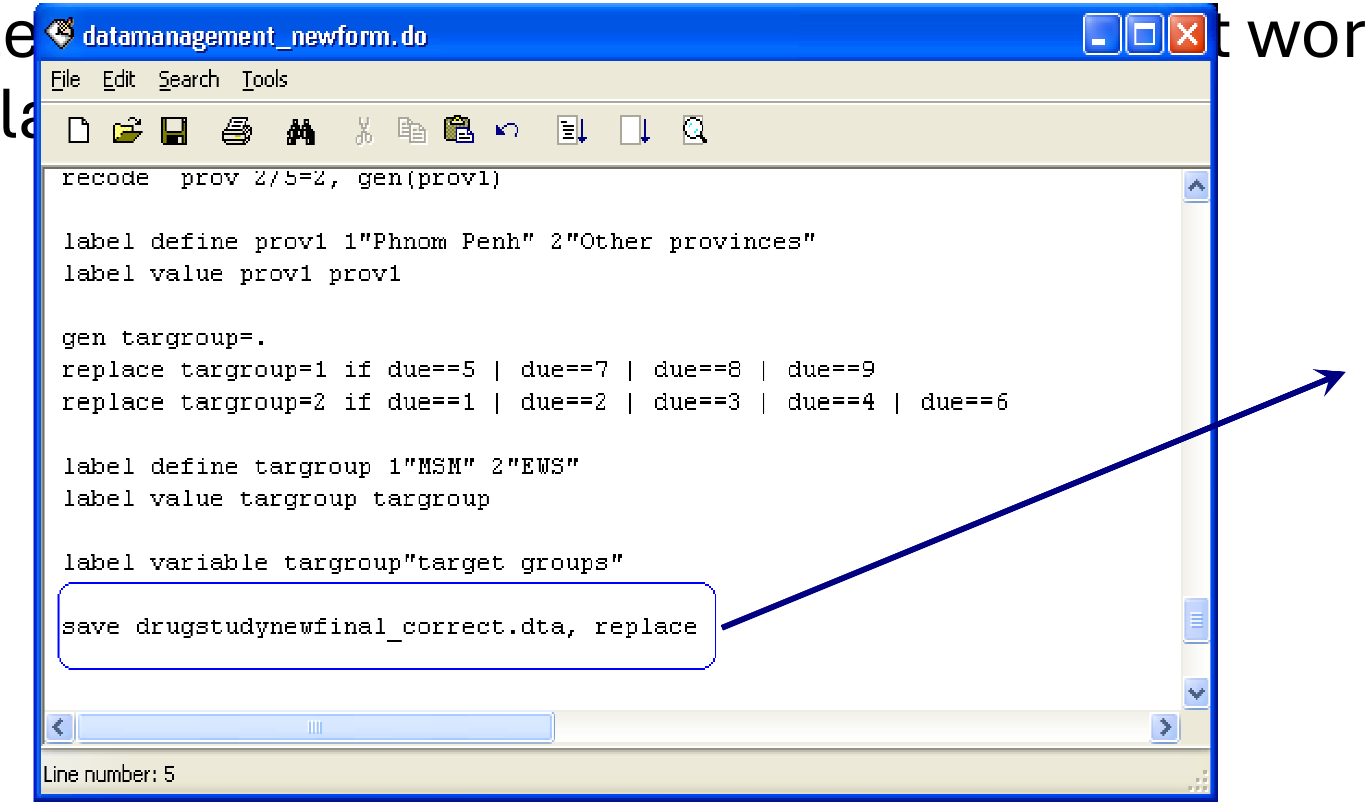


```
*STATA Training

clear
set memory 200m
cd "D:\FHI work\Reseach\FHI Research\Drug study\data\Newdataform31032009\"
use drugstudynewfinal.dta

label define prov 1"Phnom Penh" 2"Battambang" 3"Banteay Meanchey" 4"Kampong
label value prov prov

label define due 1"Sorn Rachnana"  2"Leng Sokunthea"   3"Heang Lina" 4"Suon
label value due due

gen IAs=due
recode IAs 1/2=1 6=1 3=2 4=3 5=4 7=5 8=6 9=7
```

You have to use these commands at the beginning of your do file.

Save dataset by using do file

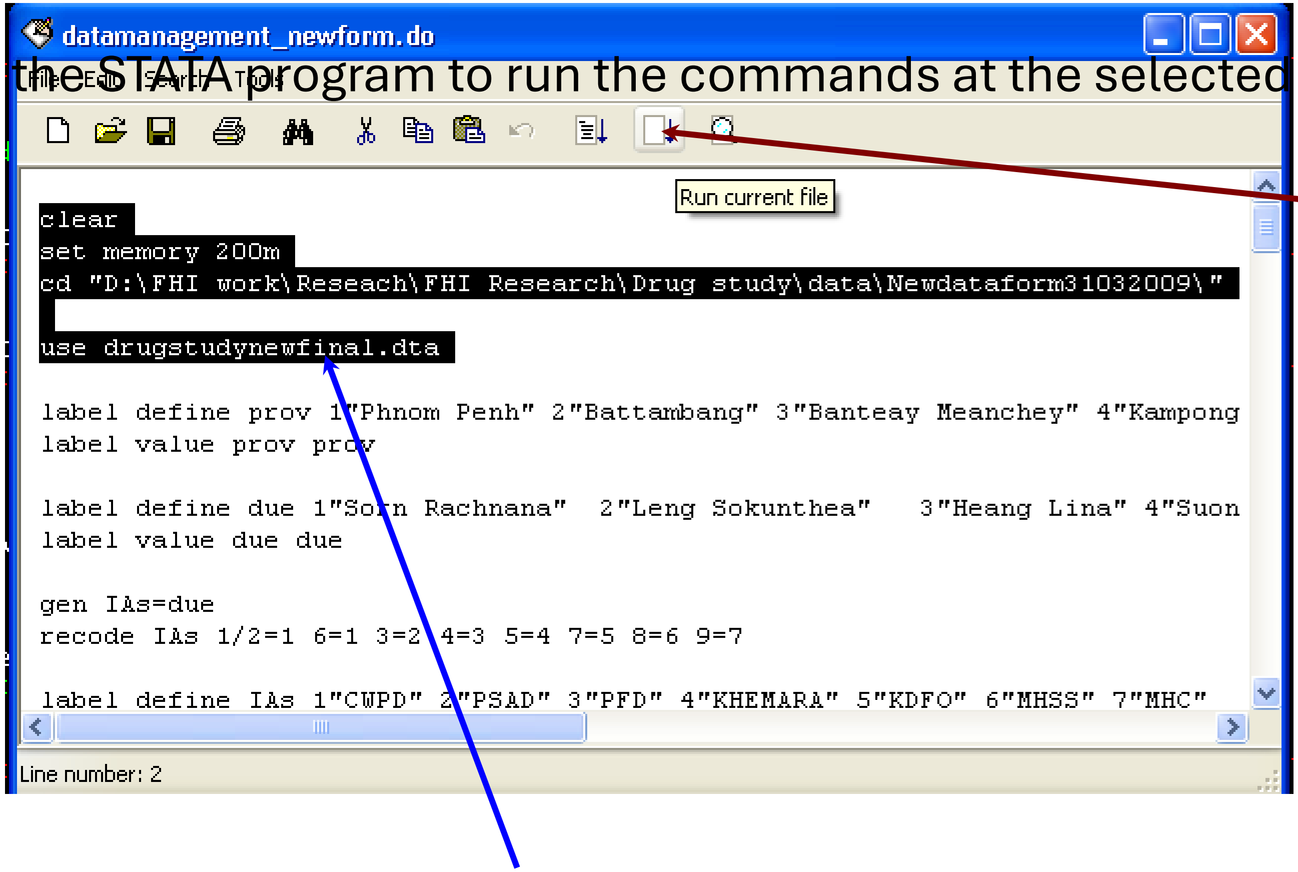
You te t working directory by repla



```
recode  prov 2/5=2, gen(prov1)

label define prov1 1"Phnom Penh" 2"Other provinces"
label value prov1 prov1

gen targroup=.
replace targroup=1 if due==5 | due==7 | due==8 | due==9
replace targroup=2 if due==1 | due==2 | due==3 | due==4 | due==6

label define targroup 1"MSM" 2"EWS"
label value targroup targroup

label variable targroup"target groups"

save drugstudynewfinal_correct.dta, replace
```

Running a ***do file*** by selecting what you want

You tell the STATA program to run the commands at the selected one.



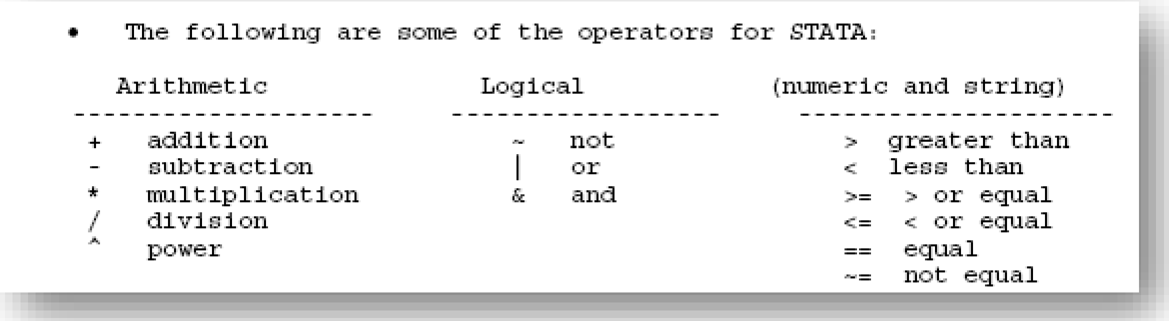
```
clear
set memory 200m
cd "D:\FHI work\Reseach\FHI Research\Drug study\data\Newdataform31032009\"

use drugstudynewfinal.dta

label define prov 1"Phnom Penh" 2"Battambang" 3"Banteay Meanchey" 4"Kampong
label value prov prov

label define due 1"Sorn Rachnana"  2"Leng Sokunthea"   3"Heang Lina" 4"Suon
label value due due

gen IAs=due
recode IAs 1/2=1 6=1 3=2 4=3 5=4 7=5 8=6 9=7

label define IAs 1"CWPD" 2"PSAD" 3"PFD" 4"KHEMARA" 5"KDFO" 6"MHSS" 7"MHC"
```

Where you can click to tell STATA to run a do file for you.

Preparing the data for analysis

1. STATA Operators Used in Data Manipulation

When managing or preparing your dataset, you will nearly always apply a condition in one way or another by using the appropriate STATA operators.



- The following are some of the operators for *STATA*:

Arithmetic		Logical		(numeric and string)	
+	addition	~	not	>	greater than
-	subtraction	|	or	<	less than
*	multiplication	&	and	>=	> or equal
/	division			<=	< or equal
^	power			==	equal
				~=	not equal

Note: *Table above is good to refer when applying logical expressions to your data. See help operators for more information*

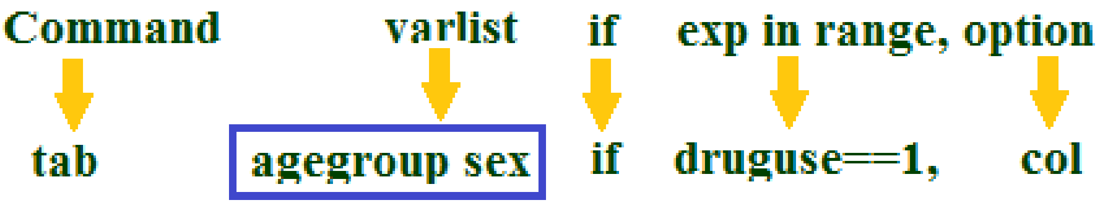
**** You have to remember the general structure of STATA commands ******

Most commands have a common syntax, which we write as

command varlist if exp in range, option

- **Command:** *What STATA is supposed to do, such as tabulation, list, save, etc.*
- **Varlist:** *The variables STATA is to use. varlist is optional in many commands. Example. list agegroup sex*
- **If exp:** *Read this as "if expression". This qualifier specifies the observations STATA is to use. If not specified, that means command is performed on all the observations.*
- **In range** : *This qualifier performs the same task as if exp but specifies the restriction in terms of observation numbers.*
- **Option:** *are features specific to each command.*

E.g. I want to know the percentage of drug users by age group and sex. I knew that agegroup: population by age group, sex: Male and Female, q3=1: drug users. I will do as following;



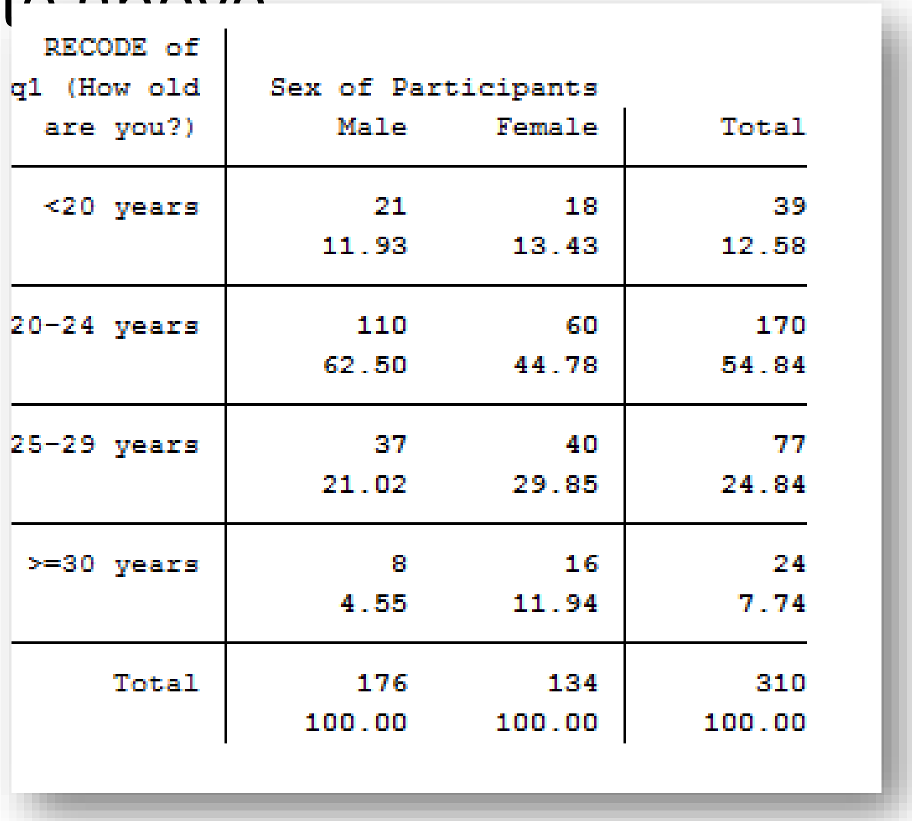
Command → tab

varlist → agegroup sex

if → if

exp in range, option → druguse==1, col

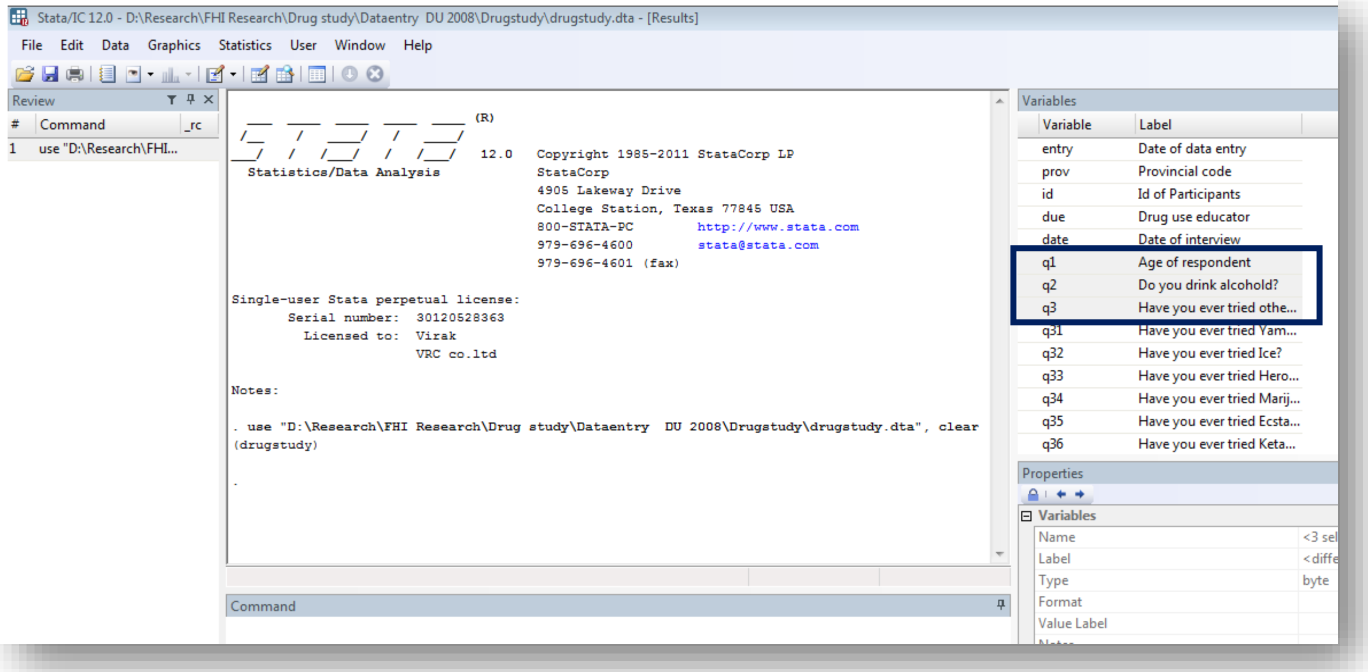
Output Result of the example above



```
 RECODE of
q1 (How old   Sex of Participants
 are you?)       Male     Female      Total

 <20 years         21         18         39
                11.93      13.43      12.58

20-24 years       110         60        170
                62.50      44.78      54.84

25-29 years        37         40         77
                21.02      29.85      24.84

 >=30 years         8         16         24
                 4.55      11.94       7.74

     Total        176        134        310
               100.00     100.00     100.00
```

2. Rename Variable

The names assigned to each variable in the dataset refer to the question number in a survey questionnaire, see picture shown below. During data manipulation and analysis, these names are not so helpful. If you want to change to a more meaningful names, e.g q1 to age, q2 to alco, ..etc.


Stata/IC 12.0 - D:\Research\FHI Research\Drug study\Dataentry DU 2008\Drugstudy\drugstudy.dta - [Results]
File
Edit
Data
Graphics
Statistics
User
Window
Help
Review
Variables
Properties
Command

#	Command	_rc
1	use "D:\Research\FHI...	

```
  ___  ____  ____  ____  ____ (R)
 /__    /   ____/   /   ____/
___/   /   /___/   /   /___/   12.0   Copyright 1985-2011 StataCorp LP
  Statistics/Data Analysis            StataCorp
                                      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)

Single-user Stata perpetual license:
       Serial number:  30120528363
         Licensed to:  Virak
                       VRC co.ltd

Notes:

. use "D:\Research\FHI Research\Drug study\Dataentry  DU 2008\Drugstudy\drugstudy.dta", clear
(drugstudy)

.
```

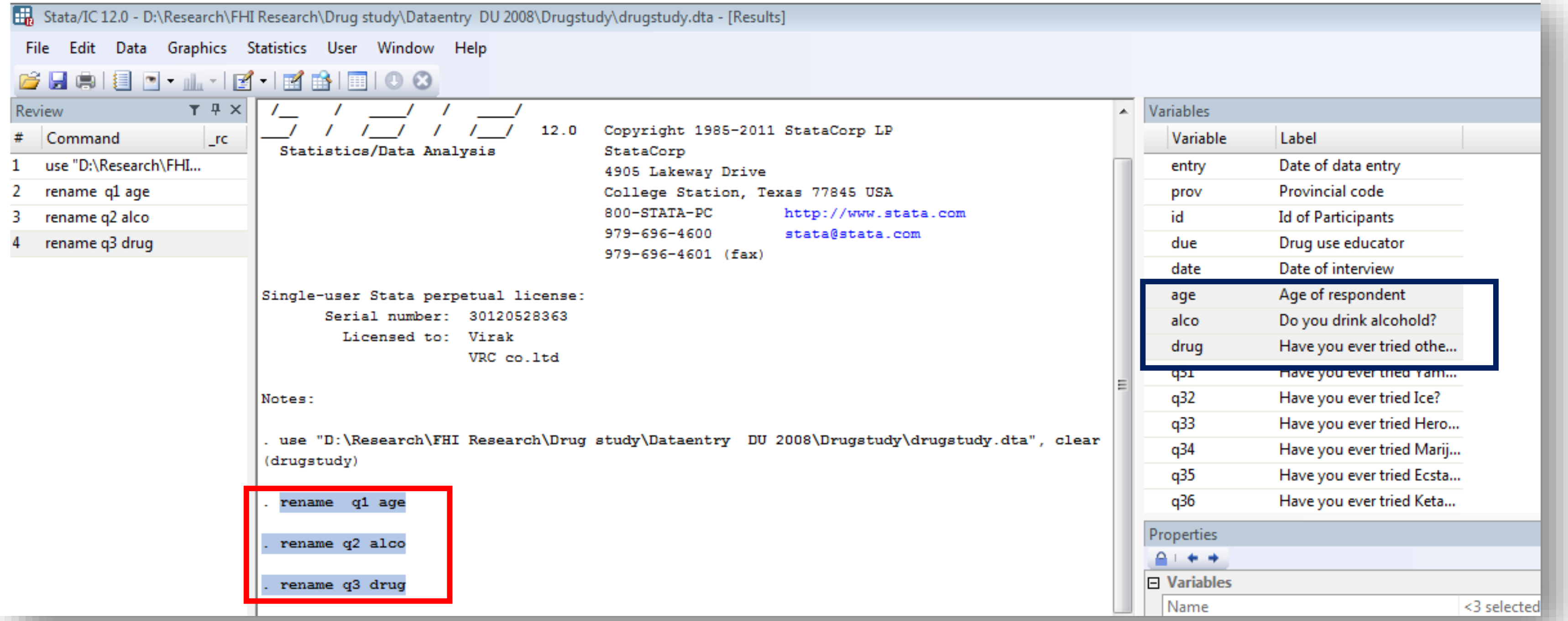
Variable	Label
entry	Date of data entry
prov	Provincial code
id	Id of Participants
due	Drug use educator
date	Date of interview
q1	Age of respondent
q2	Do you drink alcohold?
q3	Have you ever tried othe...
q31	Have you ever tried Yam...
q32	Have you ever tried Ice?
q33	Have you ever tried Hero...
q34	Have you ever tried Marij...
q35	Have you ever tried Ecsta...
q36	Have you ever tried Keta...

⊟ Variables

Name	<3 sel
Label	<diffe
Type	byte
Format	
Value Label	

2. Rename Variable (Cont.)

- To rename a new variable based on an existing variable you simply type rename followed by the original variable name and the new variable name.
- Ex: I renamed variables, q1 to age, q2 to alco, and q3 to drug
- **Type rename q1 age, rename q2 alco, rename q3 drug**



Stata/IC 12.0 - D:\Research\FHI Research\Drug study\Dataentry DU 2008\Drugstudy\drugstudy.dta - [Results]

#	Command	_rc
1	use "D:\Research\FHI...	
2	rename q1 age	
3	rename q2 alco	
4	rename q3 drug	

```
  /__    /   ____/   /   ____/
 ___/   /   /___/   /   /___/   12.0   Copyright 1985-2011 StataCorp LP
  Statistics/Data Analysis             StataCorp
                                       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)

Single-user Stata perpetual license:
       Serial number:  30120528363
         Licensed to:  Virak
                       VRC co.ltd

Notes:

. use "D:\Research\FHI Research\Drug study\Dataentry  DU 2008\Drugstudy\drugstudy.dta", clear
(drugstudy)

. rename  q1  age

. rename q2 alco

. rename q3 drug
```

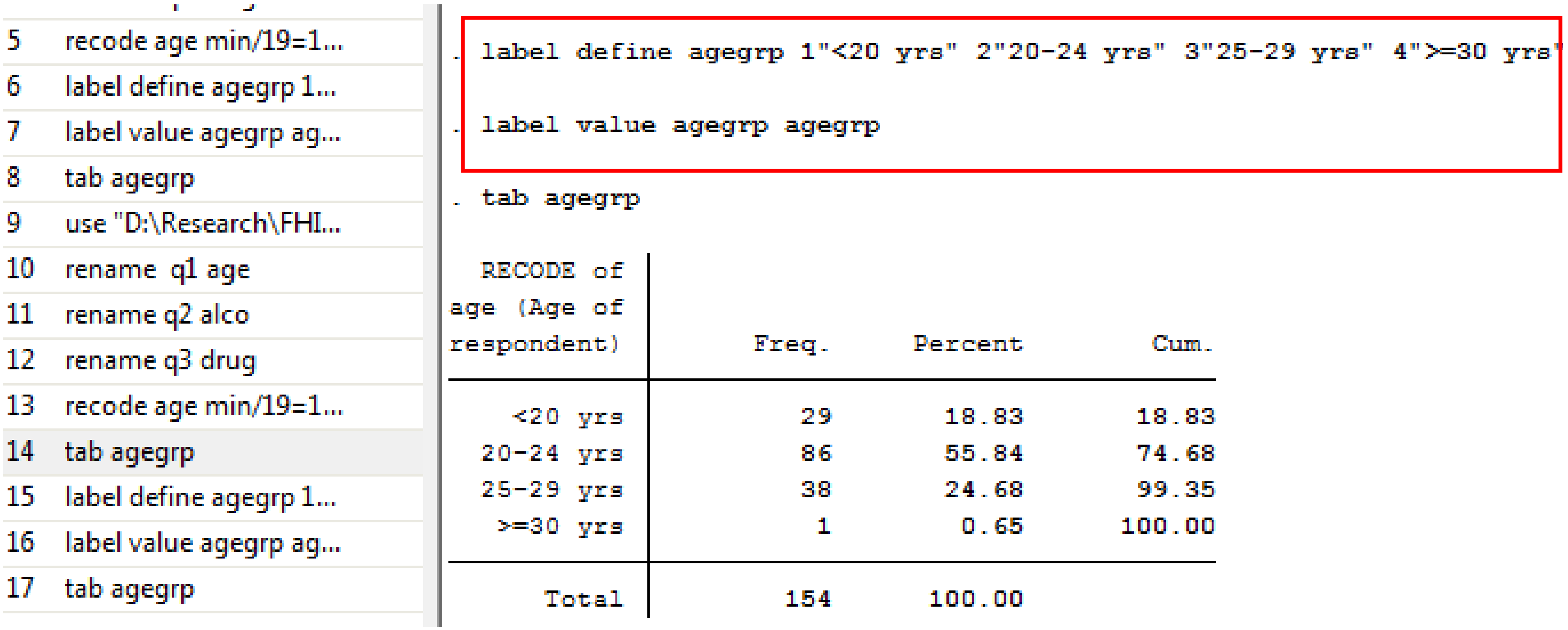
Variable	Label
entry	Date of data entry
prov	Provincial code
id	Id of Participants
due	Drug use educator
date	Date of interview
age	Age of respondent
alco	Do you drink alcohol?
drug	Have you ever tried othe...
q31	Have you ever tried Yam...
q32	Have you ever tried Ice?
q33	Have you ever tried Hero...
q34	Have you ever tried Marij...
q35	Have you ever tried Ecsta...
q36	Have you ever tried Keta...

Variables

Name <3 selected

3. Labeling variable and categories within a variable

- Label value code of agegrp. Type two lines below;
  - label define agegrp 1"<19yrs" 2"20-24yrs" 3"25-29yrs" 4"30-34yrs" 5">34Yrs“
  - label value agegrp agegrp



```
5    recode age min/19=1...
6    label define agegrp 1...
7    label value agegrp ag...
8    tab agegrp
9    use "D:\Research\FHI...
10   rename  q1 age
11   rename q2 alco
12   rename q3 drug
13   recode age min/19=1...
14   tab agegrp
15   label define agegrp 1...
16   label value agegrp ag...
17   tab agegrp
```

```
. label define agegrp 1"<20 yrs" 2"20-24 yrs" 3"25-29 yrs" 4">=30 yrs"

. label value agegrp agegrp

. tab agegrp

 RECODE of |
age (Age of |
respondent) |      Freq.     Percent        Cum.
------------+-----------------------------------
    <20 yrs |         29       18.83       18.83
  20-24 yrs |         86       55.84       74.68
  25-29 yrs |         38       24.68       99.35
   >=30 yrs |          1        0.65      100.00
------------+-----------------------------------
      Total |        154      100.00
```

Data Cleaning

(Initiating Data Exploration)

What should be done for data examination?

Before starting to do data analysis, it is absolutely essential that you have to examine the raw data first in order to find some types of error are included;

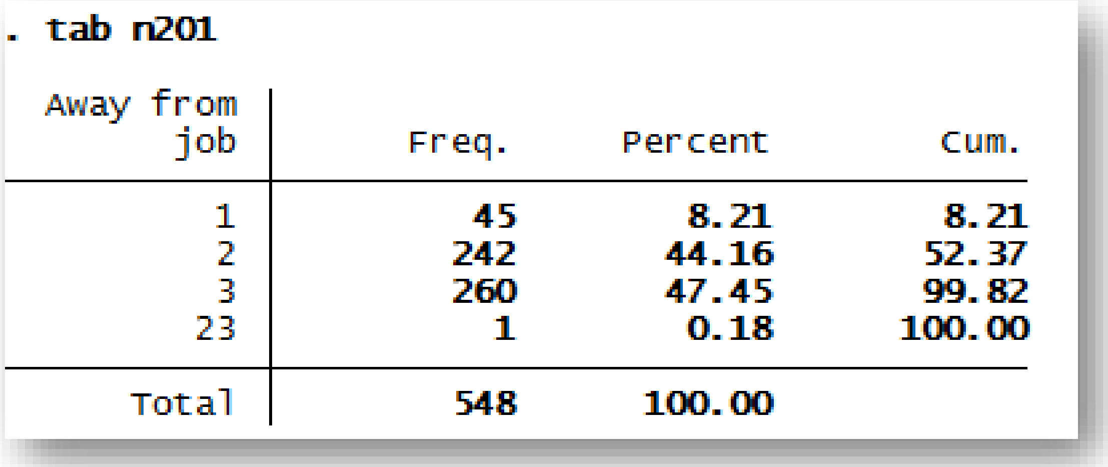
- **Illegal code**: values are not specified in code instruction
- **Omissions**: Do not follow correctly the SKIP instruction
- **Logical consistencies**: current age is less than marriage age
- **Improbabilities**: a woman aged 25 years old with 10 living children.

Where should you start

- First of all, take a look at all the variables you have identified for analysis, and produce simple tabulations for categorical variable or summary statistics for quantitative variable in order **to search Primarily for some errors.**

Tabulating Categorical Variables

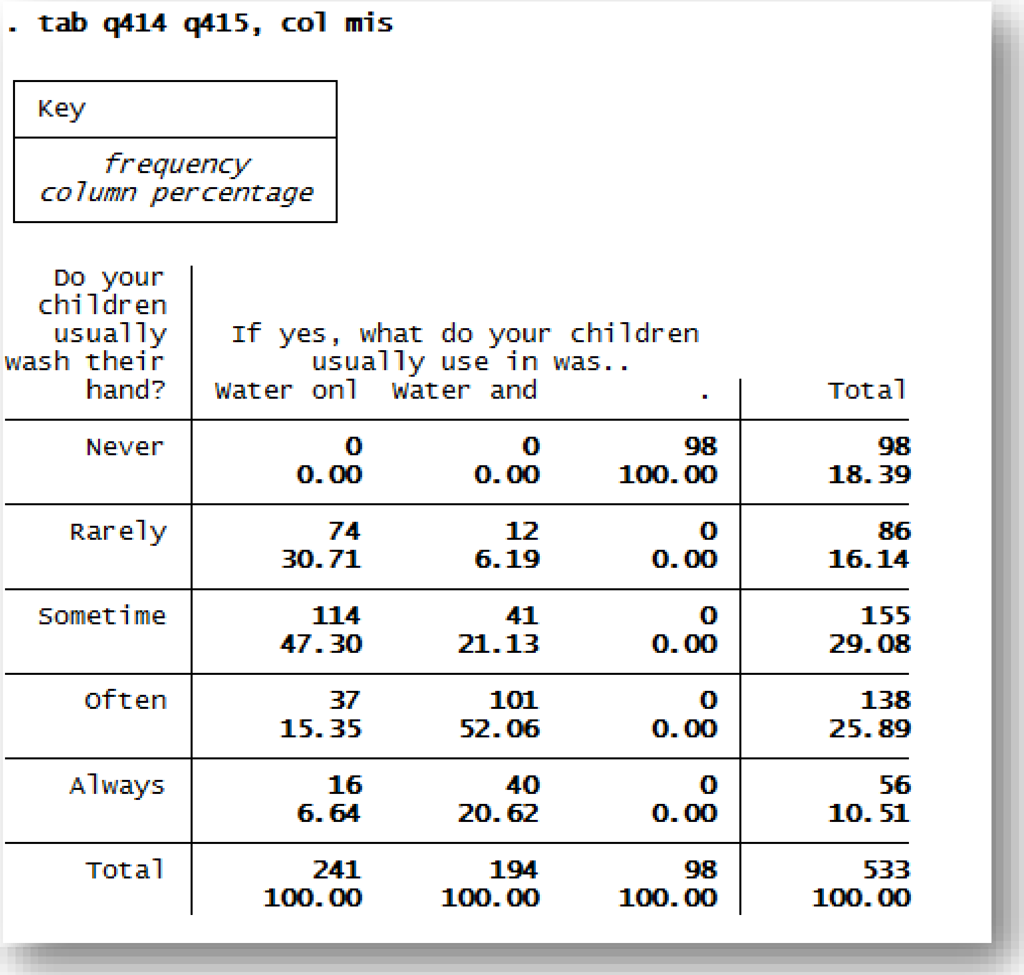
- Look at the distribution of categorical variable using the tabulate command.



```
. tab n201

  Away from |
        job |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |         45        8.21        8.21
          2 |        242       44.16       52.37
          3 |        260       47.45       99.82
         23 |          1        0.18      100.00
------------+-----------------------------------
      Total |        548      100.00
```

Tabulating Categorical Variables

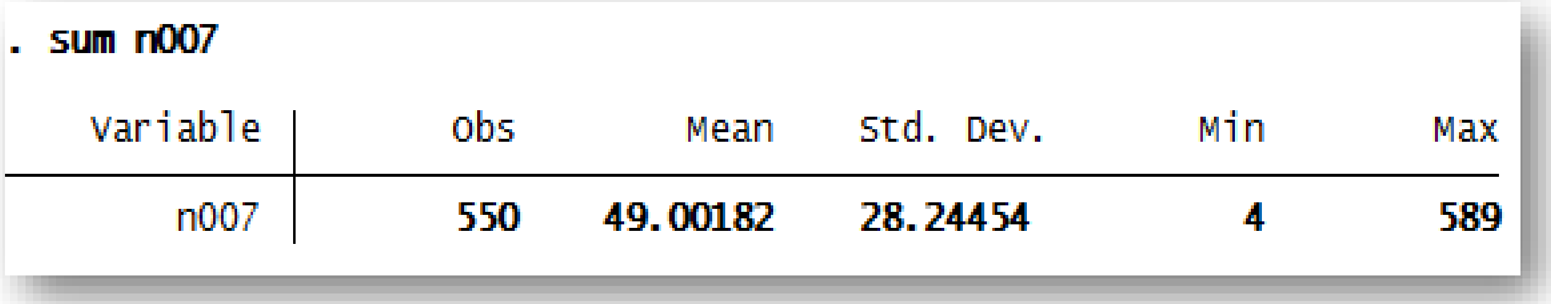
To obtain a two-way table by using command tab with the two variables.



```
. tab q414 q415, col mis

+-------------------+
| Key               |
|-------------------|
|     frequency     |
| column percentage |
+-------------------+

   Do your |
  children |
   usually |  If yes, what do your children
wash their |       usually use in was..
     hand? | Water onl  Water and          . |     Total
-----------+---------------------------------+----------
     Never |         0          0         98 |        98 
           |      0.00       0.00     100.00 |     18.39 
-----------+---------------------------------+----------
    Rarely |        74         12          0 |        86 
           |     30.71       6.19       0.00 |     16.14 
-----------+---------------------------------+----------
  Sometime |       114         41          0 |       155 
           |     47.30      21.13       0.00 |     29.08 
-----------+---------------------------------+----------
     Often |        37        101          0 |       138 
           |     15.35      52.06       0.00 |     25.89 
-----------+---------------------------------+----------
    Always |        16         40          0 |        56 
           |      6.64      20.62       0.00 |     10.51 
-----------+---------------------------------+----------
     Total |       241        194         98 |       533 
           |    100.00     100.00     100.00 |    100.00 
```

Summarizing Quantitative

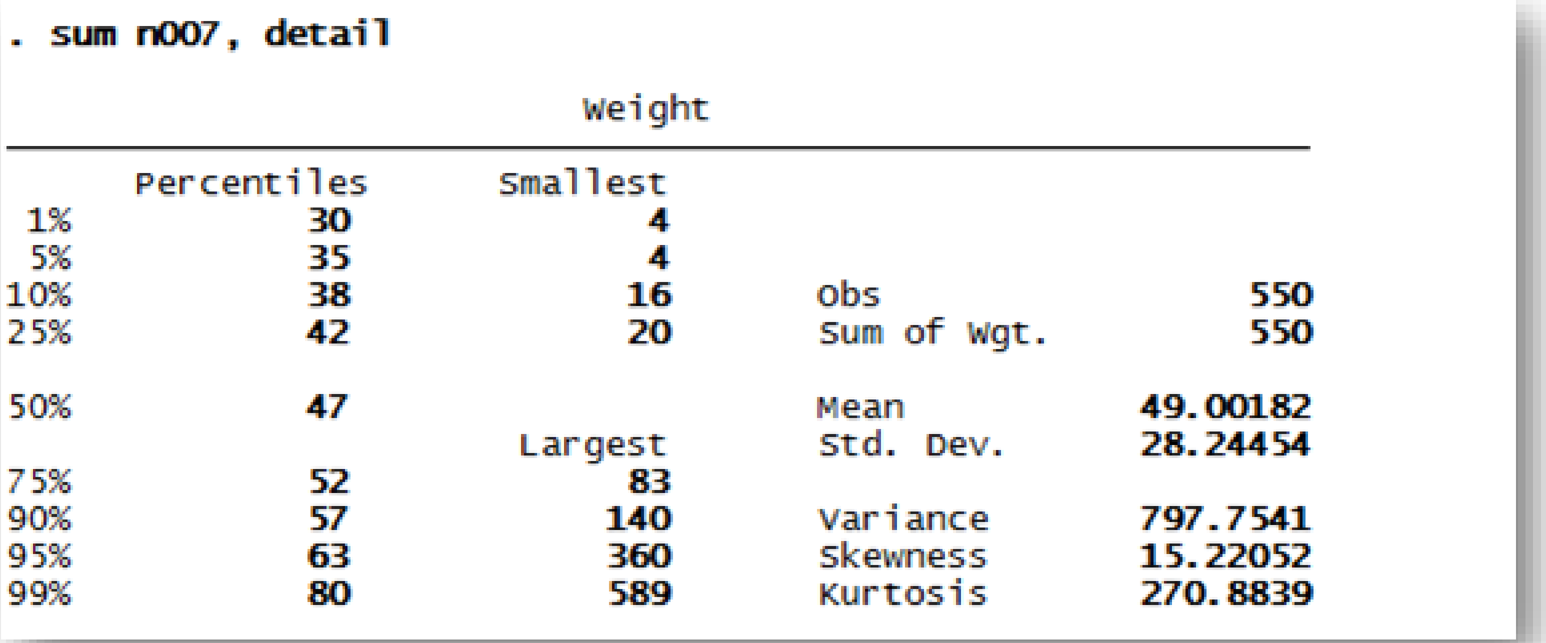
- For Quantitative variable, it is more efficient to look at summary measures (means, median and standard deviations and well as maximum and minimum values) rather than producing a table for each individual value of the variable. E.g. I want to obtain a summary of distribution of weight (n007) of respondents
  - **sum n007**



```
. sum n007

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
        n007 |        550    49.00182    28.24454          4        589
```

- In doing that it can help you to identify incorrect values and outliers.

- Stata can provide additional information by using the detail option with the summarize command.
  - sum n007 (weight variable), detail



```
. sum n007, detail

                           Weight
-------------------------------------------------------------
      Percentiles      Smallest
 1%           30              4
 5%           35              4
10%           38             16       Obs                 550
25%           42             20       Sum of Wgt.         550

50%           47                      Mean           49.00182
                        Largest       Std. Dev.      28.24454
75%           52             83
90%           57            140       Variance       797.7541
95%           63            360       Skewness       15.22052
99%           80            589       Kurtosis       270.8839
```

- Stata would provide you the number observations, the mean, standard deviation, variance, …etc for variable age.

Cross checking variables for error

- As well as tabulating and summarizing data to check for identify errors also cross check related variables.

- Age at first sex cannot be older than current age. An easy way to check this is to create a check variable:
  - gen checkage=currentage – agefirstsex
  - tab checkage… if there are negative values, it means the respondent had reported an age at first sex that is older then his or her age now. These recodes need more investigation.

Correcting data error

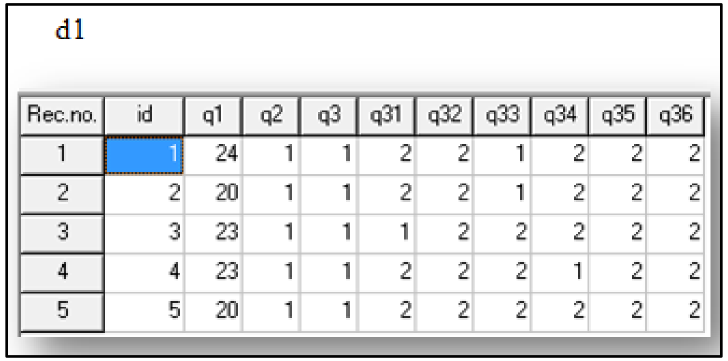
- You should do some basic consistency checking before you think of starting analysis. E.g it is not possible for someone who has never had sex to report an age at first sex etc.

COMBINE DATASETS

The process of adding two datasets into one new dataset.

Appending Data: combining two datasets which have similar or same data structures into one dataset.

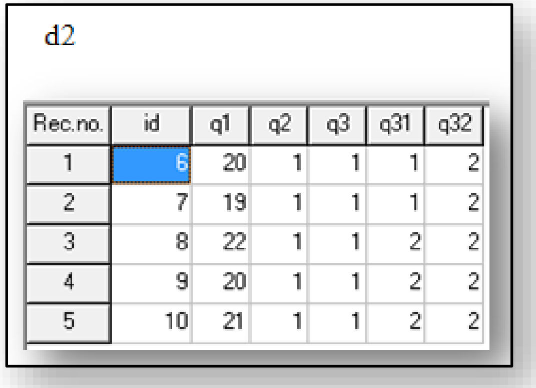
Dataset **d1**



d1

Rec.no.	id	q1	q2	q3	q31	q32	q33	q34	q35	q36
1	1	24	1	1	2	2	1	2	2	2
2	2	20	1	1	2	2	1	2	2	2
3	3	23	1	1	1	2	2	2	2	2
4	4	23	1	1	2	2	2	1	2	2
5	5	20	1	1	2	2	2	2	2	2

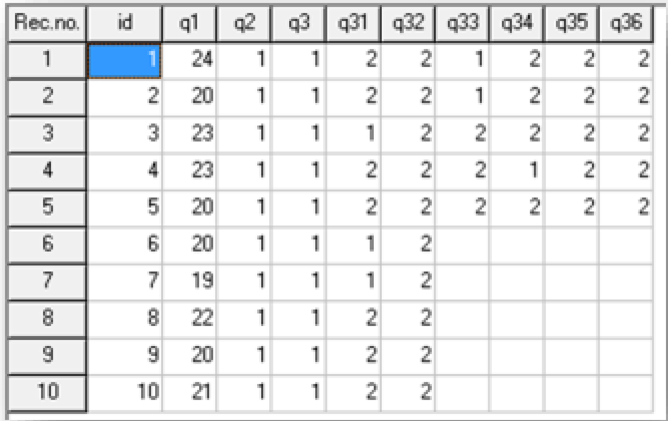
Dataset **d2**



d2

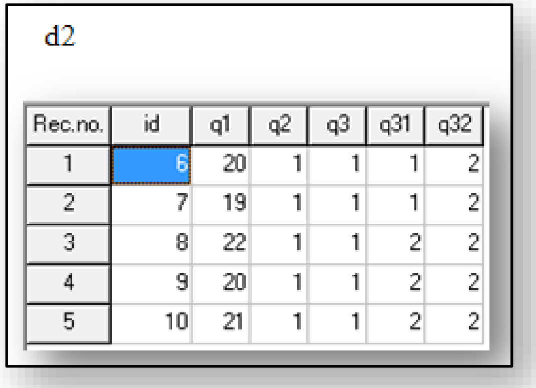
Rec.no.	id	q1	q2	q3	q31	q32
1	6	20	1	1	1	2
2	7	19	1	1	1	2
3	8	22	1	1	2	2
4	9	20	1	1	2	2
5	10	21	1	1	2	2

Dataset **d1** appended with **d2**



Rec.no.	id	q1	q2	q3	q31	q32	q33	q34	q35	q36
1	1	24	1	1	2	2	1	2	2	2
2	2	20	1	1	2	2	1	2	2	2
3	3	23	1	1	1	2	2	2	2	2
4	4	23	1	1	2	2	2	1	2	2
5	5	20	1	1	2	2	2	2	2	2
6	6	20	1	1	1	2				
7	7	19	1	1	1	2				
8	8	22	1	1	2	2				
9	9	20	1	1	2	2				
10	10	21	1	1	2	2				

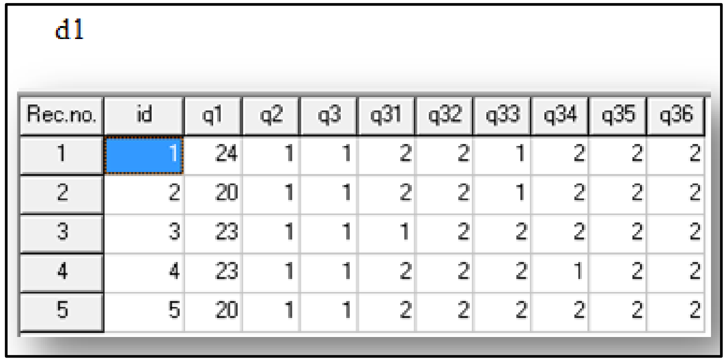
Dataset **2**



d2

Rec.no.	id	q1	q2	q3	q31	q32
1	6	20	1	1	1	2
2	7	19	1	1	1	2
3	8	22	1	1	2	2
4	9	20	1	1	2	2
5	10	21	1	1	2	2

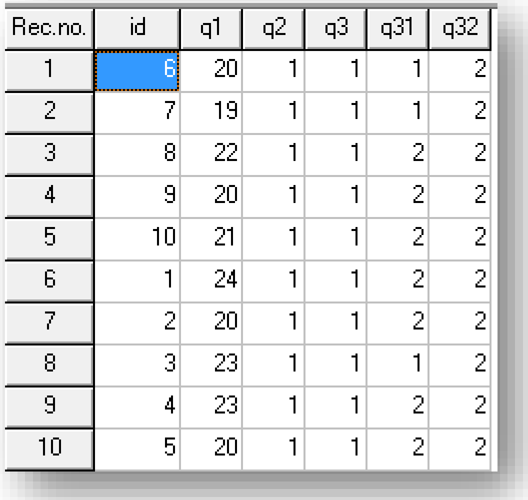
Dataset **2**



d1

Rec.no.	id	q1	q2	q3	q31	q32	q33	q34	q35	q36
1	1	24	1	1	2	2	1	2	2	2
2	2	20	1	1	2	2	1	2	2	2
3	3	23	1	1	1	2	2	2	2	2
4	4	23	1	1	2	2	2	1	2	2
5	5	20	1	1	2	2	2	2	2	2

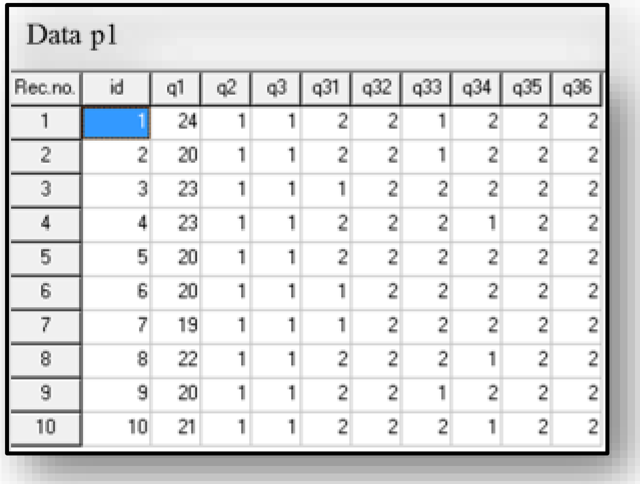
Dataset **d2** appended with **d1**



Rec.no.	id	q1	q2	q3	q31	q32
1	6	20	1	1	1	2
2	7	19	1	1	1	2
3	8	22	1	1	2	2
4	9	20	1	1	2	2
5	10	21	1	1	2	2
6	1	24	1	1	2	2
7	2	20	1	1	2	2
8	3	23	1	1	1	2
9	4	23	1	1	2	2
10	5	20	1	1	2	2

Merging Data: combine two datasets which have different data structures into one dataset.

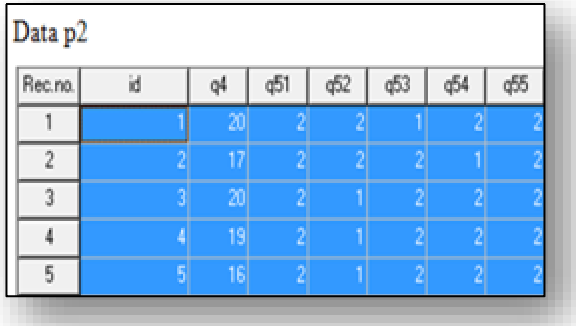
Dataset **p1**



Data p1

Rec.no.	id	q1	q2	q3	q31	q32	q33	q34	q35	q36
1	1	24	1	1	2	2	1	2	2	2
2	2	20	1	1	2	2	1	2	2	2
3	3	23	1	1	1	2	2	2	2	2
4	4	23	1	1	2	2	2	1	2	2
5	5	20	1	1	2	2	2	2	2	2
6	6	20	1	1	1	2	2	2	2	2
7	7	19	1	1	1	2	2	2	2	2
8	8	22	1	1	2	2	2	1	2	2
9	9	20	1	1	2	2	1	2	2	2
10	10	21	1	1	2	2	2	1	2	2

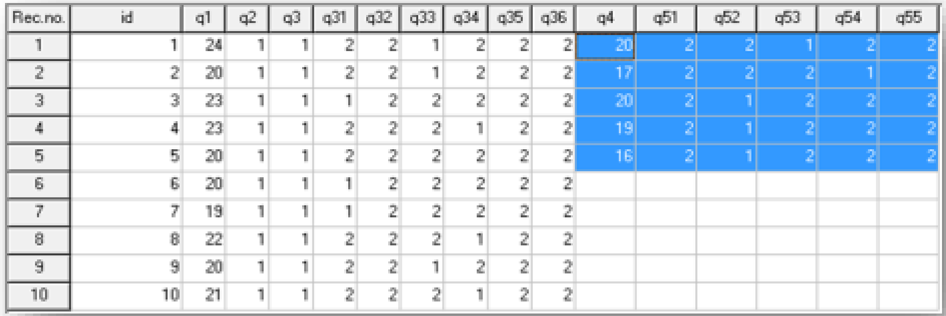
Dataset **p2**



Data p2

Rec.no.	id	q4	q51	q52	q53	q54	q55
1	1	20	2	2	1	2	2
2	2	17	2	2	2	1	2
3	3	20	2	1	2	2	2
4	4	19	2	1	2	2	2
5	5	16	2	1	2	2	2

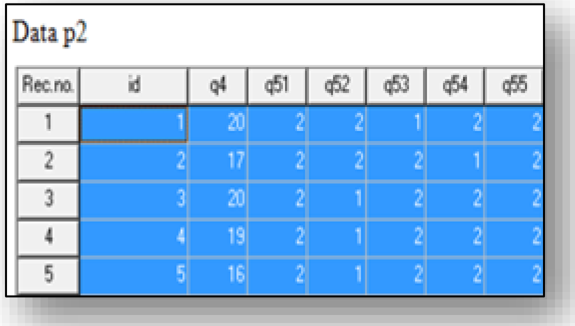
Dataset **p1** merged with **p2**



Rec.no.	id	q1	q2	q3	q31	q32	q33	q34	q35	q36	q4	q51	q52	q53	q54	q55
1	1	24	1	1	2	2	1	2	2	2	20	2	2	1	2	2
2	2	20	1	1	2	2	1	2	2	2	17	2	2	2	1	2
3	3	23	1	1	1	2	2	2	2	2	20	2	1	2	2	2
4	4	23	1	1	2	2	2	1	2	2	19	2	1	2	2	2
5	5	20	1	1	2	2	2	2	2	2	16	2	1	2	2	2
6	6	20	1	1	1	2	2	2	2	2						
7	7	19	1	1	1	2	2	2	2	2						
8	8	22	1	1	2	2	2	1	2	2						
9	9	20	1	1	2	2	1	2	2	2						
10	10	21	1	1	2	2	2	1	2	2						

Merging Data: combine two datasets which have different data structures into one dataset.

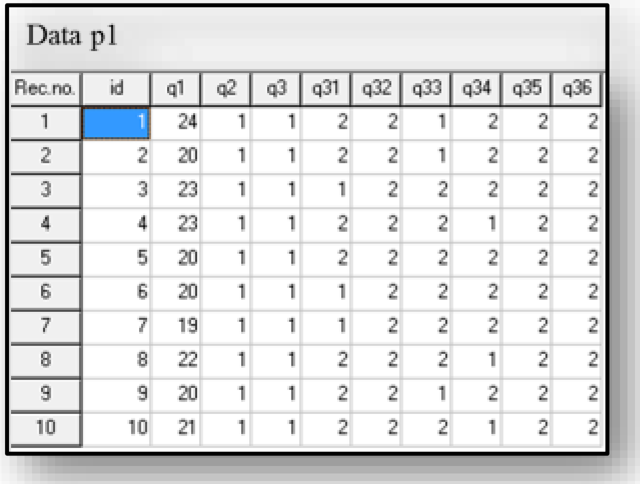
Dataset **p2**



Data p2

Rec.no.	id	q4	q51	q52	q53	q54	q55
1	1	20	2	2	1	2	2
2	2	17	2	2	2	1	2
3	3	20	2	1	2	2	2
4	4	19	2	1	2	2	2
5	5	16	2	1	2	2	2

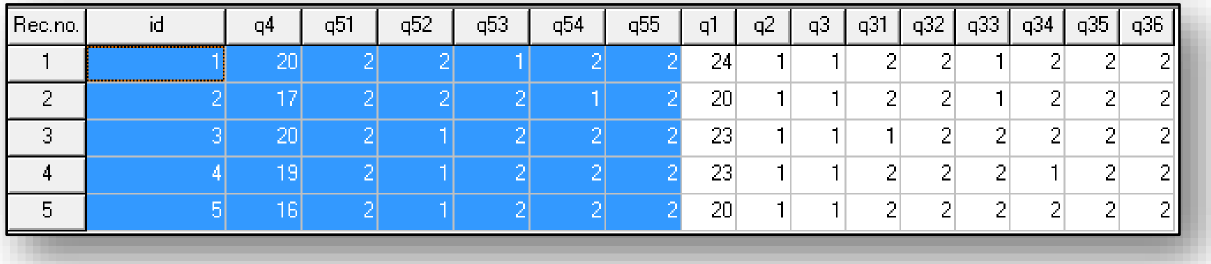
Dataset **p1**



Data p1

Rec.no.	id	q1	q2	q3	q31	q32	q33	q34	q35	q36
1	1	24	1	1	2	2	1	2	2	2
2	2	20	1	1	2	2	1	2	2	2
3	3	23	1	1	1	2	2	2	2	2
4	4	23	1	1	2	2	2	1	2	2
5	5	20	1	1	2	2	2	2	2	2
6	6	20	1	1	1	2	2	2	2	2
7	7	19	1	1	1	2	2	2	2	2
8	8	22	1	1	2	2	2	1	2	2
9	9	20	1	1	2	2	1	2	2	2
10	10	21	1	1	2	2	2	1	2	2

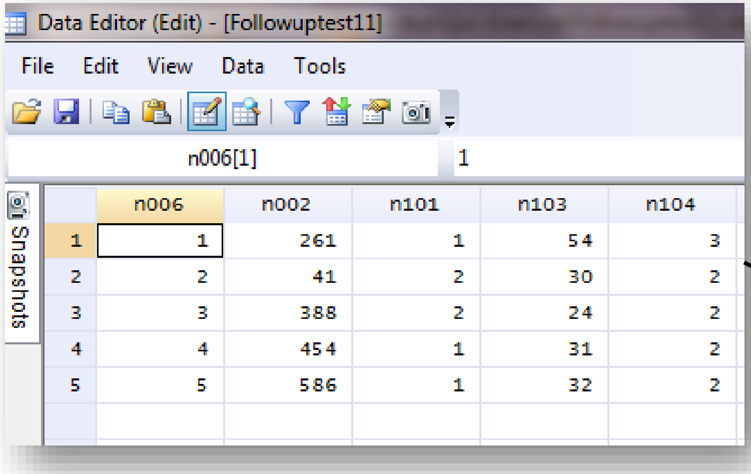
Dataset **p2** merged with **p1**



Rec.no.	id	q4	q51	q52	q53	q54	q55	q1	q2	q3	q31	q32	q33	q34	q35	q36
1	1	20	2	2	1	2	2	24	1	1	2	2	1	2	2	2
2	2	17	2	2	2	1	2	20	1	1	2	2	1	2	2	2
3	3	20	2	1	2	2	2	23	1	1	1	2	2	2	2	2
4	4	19	2	1	2	2	2	23	1	1	2	2	2	1	2	2
5	5	16	2	1	2	2	2	20	1	1	2	2	2	2	2	2

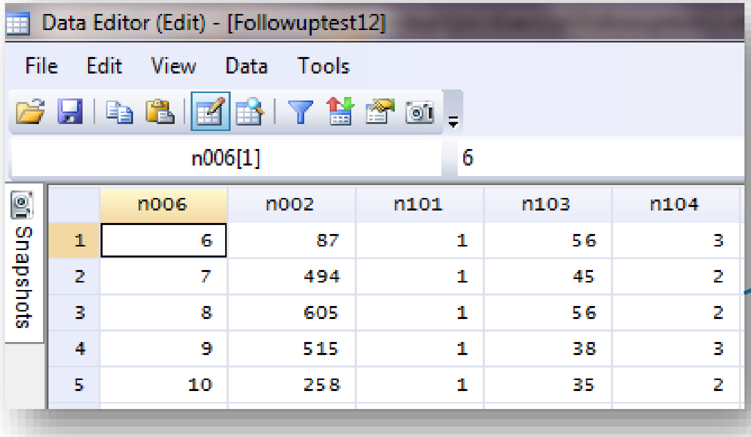
Combine datasets: by Adding observations

```
use “C:\Users\virak\Desktop\Final_STATA\Exercise\Followuptest11.dta”

append using "C:\Users\virak\Desktop\Final_STATA\Exercise\Followuptest12.dta"

save "C:\Users\virak\Desktop\Final_STATA\Exercise\Followuptest11&12.dta”, replace
```



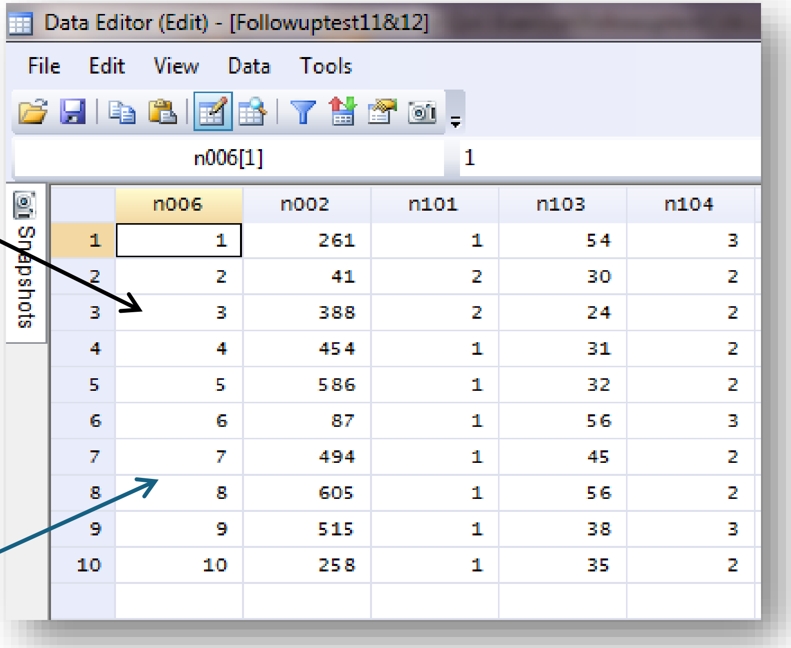
n006[1] 1

	n006	n002	n101	n103	n104
1	1	261	1	54	3
2	2	41	2	30	2
3	3	388	2	24	2
4	4	454	1	31	2
5	5	586	1	32	2



n006[1] 6

	n006	n002	n101	n103	n104
1	6	87	1	56	3
2	7	494	1	45	2
3	8	605	1	56	2
4	9	515	1	38	3
5	10	258	1	35	2

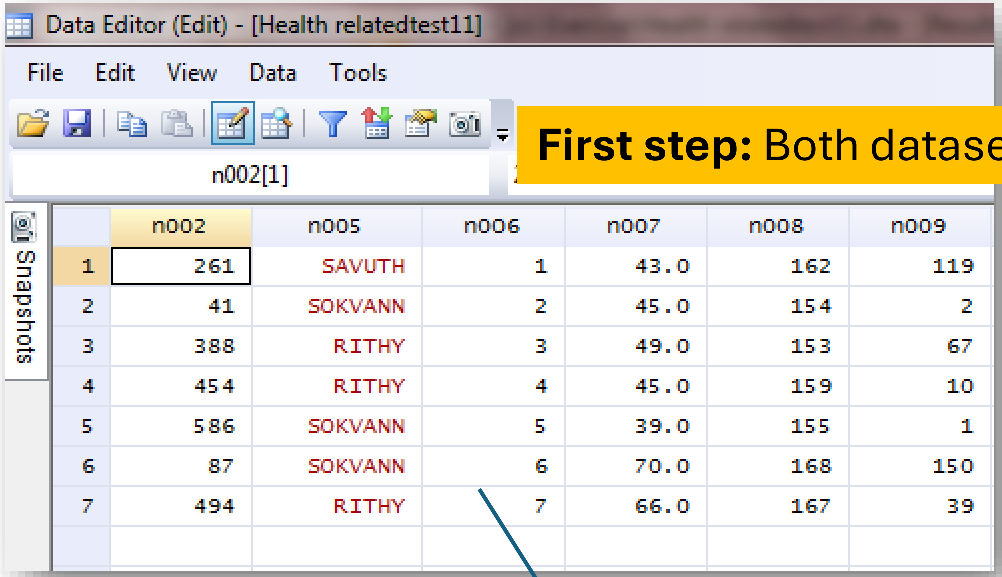


n006[1] 1

	n006	n002	n101	n103	n104
1	1	261	1	54	3
2	2	41	2	30	2
3	3	388	2	24	2
4	4	454	1	31	2
5	5	586	1	32	2
6	6	87	1	56	3
7	7	494	1	45	2
8	8	605	1	56	2
9	9	515	1	38	3
10	10	258	1	35	2

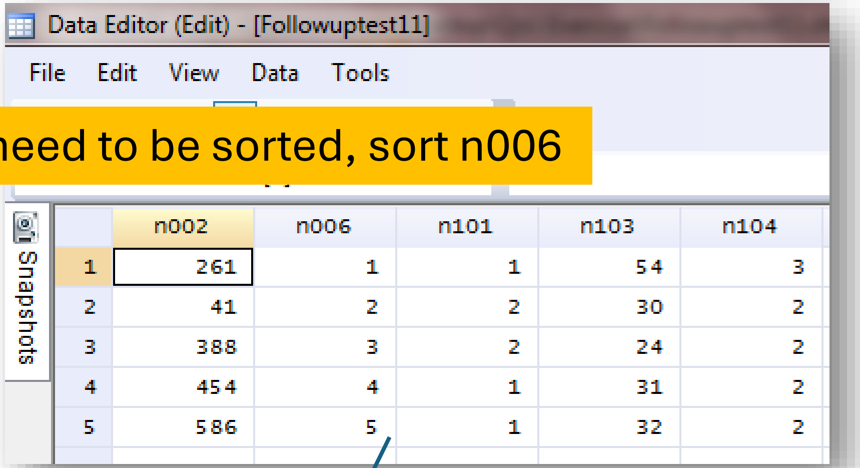
```
use "C:\Users\virak\Desktop\Final_STATA\Exercise\Healthrelatedtest11.dta"
merge 1:1 n006 using "C:\Users\virak\Desktop\Final_STATA\Exercise\Followuptest11.dta“
save "C:\Users\virak\Desktop\Final_STATA\Exercise\Merge_follow11&health11.dta"
```

First step: Both datasets need to be sorted, sort n006

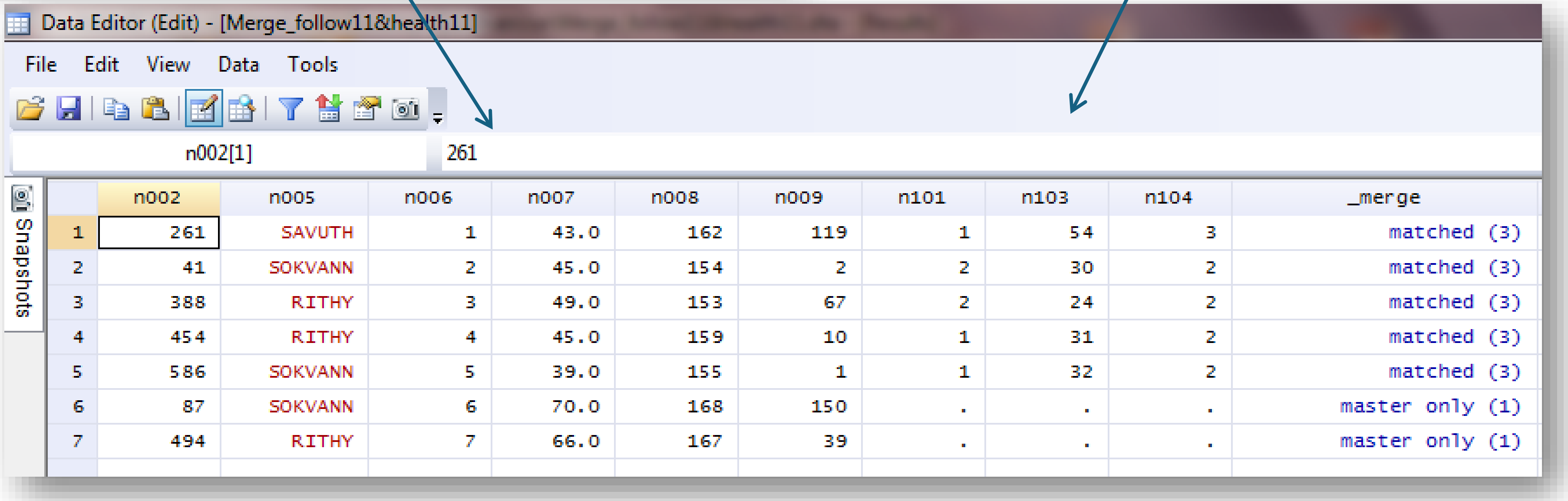

Data Editor (Edit) - [Health relatedtest11]
File Edit View Data Tools
Snapshots

n002[1]

	n002	n005	n006	n007	n008	n009
1	261	SAVUTH	1	43.0	162	119
2	41	SOKVANN	2	45.0	154	2
3	388	RITHY	3	49.0	153	67
4	454	RITHY	4	45.0	159	10
5	586	SOKVANN	5	39.0	155	1
6	87	SOKVANN	6	70.0	168	150
7	494	RITHY	7	66.0	167	39


Data Editor (Edit) - [Followuptest11]
File Edit View Data Tools
Snapshots

	n002	n006	n101	n103	n104
1	261	1	1	54	3
2	41	2	2	30	2
3	388	3	2	24	2
4	454	4	1	31	2
5	586	5	1	32	2


Data Editor (Edit) - [Merge_follow11&health11]
File Edit View Data Tools
Snapshots

n002[1] 261

	n002	n005	n006	n007	n008	n009	n101	n103	n104	_merge
1	261	SAVUTH	1	43.0	162	119	1	54	3	matched (3)
2	41	SOKVANN	2	45.0	154	2	2	30	2	matched (3)
3	388	RITHY	3	49.0	153	67	2	24	2	matched (3)
4	454	RITHY	4	45.0	159	10	1	31	2	matched (3)
5	586	SOKVANN	5	39.0	155	1	1	32	2	matched (3)
6	87	SOKVANN	6	70.0	168	150	.	.	.	master only (1)
7	494	RITHY	7	66.0	167	39	.	.	.	master only (1)