

Exercises for Stata

Svend Juul, June 2011

The purpose of these exercises is to learn Stata by doing. Use Svend Juul and Morten Frydenberg's Stata book (ISHR3) to look up commands, etc., during the exercises.¹

From <http://www.biostat.au.dk/teaching/senior> you can download datasets for exercises to your hard-disk; at the computer-lab download them to **D:\StataCourse\your name**.² For this exercise we will use the **smoke.dta** dataset. Further instruction will be given at the course.

In the exercise questions I often give a hint about the procedure to be used (e.g. **describe**). Look up the command syntax in ISHR3.

The **smoke.dta** dataset includes the following variables:

Variable	Meaning	Possible values
id	ID number	1-250
sex	Sex	1 male 2 female
age	Age in years	18-100
weight	Weight in kg	40-150
height	Height in cm	120-220
smoker	Smoker?	0 no 1 current smoker 2 former smoker
cigaret	Cigarettes/day	0-60
cheroot	Cigars or cheroots per day	0-20
pipe	Packs of pipe tobacco per week	0-10

1. In the Command window enter:

cd "D:\StataCourse"	to make D:\StataCourse the current file path
mkdir "your name"	to create the folder D:\Statacourse\your name
cd "your name"	to make D:\StataCourse\your name the current file path
use "smoke.dta"	to load the smoke.dta dataset into Stata

Look at the Variables window; it displays the names of variables in the **smoke.dta** dataset. Look in the lower left corner, just below the Variables window; it displays the current file path

¹ Juul S, Frydenberg M. An introduction to Stata for health researchers, 3rd ed. College Station, TX: Stata Press, 2010. I refer to the book as ISHR3.

² Not "your name", of course, but a folder with your name.

2. Open the Data window with the command:

browse

and take a look.

Now close the Data window. Open it again to see the values of **age** and **weight** for the first 5 observations by:

browse age weight in 1/5

There is a button for **browse**, too, but it does not let you select observations and variables.

3. In the Command window now enter:

summarize

and look at the Results window. **summarize** displays the number of valid observations, mean, standard deviation, and the minimum and maximum values of all variables.

Use the mouse to highlight the table, and print it. (Click the printer button. In the dialog that appears, click the Selection radio button; the rest can be done by intuition).

4. Look at the Review window. It displays the two commands issued so far (and possibly also some errors you made). Now, click the last command (**summarize**) and watch it being pasted to the Command window. In the Variables window click **weight** and watch it being pasted to the Command window, which now displays:

summarize weight

Hit the *Enter* key and see what happens.

5. Try the dialog system to obtain the same result. If you know the command name it is easiest to enter the **db** (dialog box) command:

db summarize

If you are searching a command, but don't know its name, use the menu system; for **summarize** it is:³

Statistics ► Summaries, tables, and tests ► Summary and descriptive statistics ►
Summary statistics

In the dialog that appears, use the drop-down list in the Variables field and select **age**.

The dialogs generate commands. In the case of **summarize age** it obviously was easier to write the command than to create it using the dialog, but for complex commands and graphs the dialogs can be an advantage.

³ Stata's menu system is not Scandinavian design, and it can be difficult to guess which branch to follow to locate a command. If you know the command name, use the **db** command.

6. Click anywhere in the Review window and press *Ctrl+A* to select all commands. Then press *Ctrl+C* to copy the commands to the Windows Clipboard. Next, open a Do-file editor window and paste the commands (*Ctrl-V*). There may be some errors. Edit the file, and put a **clear** command in the first line; after editing it should look like this:

```
clear
use smoke.dta
summarize
summarize weight
```

Save the revised do-file as "**D:\StataCourse\your name\test1.do**". Now, click the Do-file editor's do-button (rightmost button) and see what happens. In the Results window you see an output like this:

```
. do "D:\StataCourse\your name\test1.do"
. clear
. use smoke.dta
. summarize
```

Variable	Obs	Mean	Std. Dev.	Min	Max
id	230	134.0696	68.99164	1	250
sex	230	1.73913	.4400666	1	2
age	230	55.80435	14.27799	21	84
weight	227	64.08811	11.85798	43	110
...					

In the output, each command is preceded by a period (Stata's command prompt), but you should not type it when you enter a command yourself.

You may also execute the do-file from the Command window:

```
do test1.do
```

Rather than remembering the exact path and name of the do-file, it may be easier to locate it by: **File ► Do...** Try that.

7. Get an overview of the dataset with these commands

```
summarize
codebook , compact
describe
label list
tab1 sex smoker cigaret
```

Study the output and make clear to yourself what information each command gives. Print selected parts of the output; you will need it for the following exercises.⁴

⁴ Before attempting to print, study the note on log files at the end of this paper.

8. Look at the tables for **sex** and **smoker**; they display value labels. To see the codes instead:

```
tab1 smoker , nolabel
```

Use the **numlabel** command to include the code in the value label:

```
numlabel , add  
tab1 smoker
```

Would it be wise to give this command?

```
tab1 id
```

9. Examine graphically with a scatterplot the relation between height and weight (use the menu system: **Graphs ► Twoway graph ► scatter**). After having studied the graph, look at the Review window; here you see the graph command; it should be:

```
twoway (scatter weight height)
```

Click the command; now it is copied to the Command window. Add:

```
twoway (scatter weight height)(lfit weight height)
```

To restrict the graph to women only, use the **if** qualifier (ISHR3, section 8.1):

```
twoway (scatter weight height)(lfit weight height) if sex==2
```

10. Create a new variable, **agegrp**, which is a reasonable grouping of **age** (**recode**).

Create a new variable, **tobacco**: tobacco use in grams per day (1 cigarette = 1 g, 1 cigar or cheroot = 2 g, 1 pack of pipe tobacco = 40 g) (**generate**).

Define labels for **agegrp** and **tobacco** (**label variable**, **label define**, **label values**).

Create a new dataset, **smoke1.dta**, including the two new variables (**save**).

Since this leads to a new version of the dataset, it is strongly recommended to do it in a do-file, starting with:

```
cd "D:\StataCourse\your name"  
use smoke.dta , clear
```

and ending with:

```
save smoke1.dta , replace
```

with the calculation commands between the **use** and the **save** command.

Did you save the do-file creating **smoke1.dta**? What name did you give it? (My recommendation: **gen_smoke1.do**; the do-file that generates **smoke1.dta**):

11. From **smoke1.dta**: see a frequency table for **tobacco** (**tab1**). Compare with the frequency tables for cigarettes etc. (from question 7) and decide if the result makes sense. Also, list the first 30 cases to see if calculations gave the result intended (**list** or **browse** with the **in** qualifier). If wrong, redo exercise 9.
12. **agegrp** could have been made with **egen**, using the **cut** function (described in the booklet under **recode**). Try to do that.
13. Describe with a crosstable the joint age and sex distribution of the study population (**tab2**). Would you prefer the original age or the grouped age for this kind of output?

14. Create a new variable, **bmi** (Body Mass Index) = weight/height² (weight in kg, height in m). See a **summarize** table for **bmi** and check that it looks reasonable. See the average **bmi** by sex and age groups (**tabstat**). Test if the **bmi** distribution is different for men and women (**ttest**).
15. Group **bmi** in 3 groups (**recode**). See a crosstable between the grouped BMI and age variables (**tab2**). Add column percentages, a chi-squared and a Fisher's exact test to the crosstable (ISHR3, section 11.3)
16. The new variables **bmi** and **bmigrp** are useful and should be included in the dataset. One option is to include the commands in **gen_smoke1.do** and run it again, another option is to include them in a **gen_smoke2.do**, which generates **smoke2.dta**.
17. Make a list of all whose BMI is larger than 30, showing the variables **id weight height bmi** (**list** with the **if** qualifier).

Look at the list. Any surprises? (Read ISHR3, section 5.4 on missing values).

Using log files

In the computer lab, an output log is created automatically; it contains a copy of everything you saw in the Results window. It is a simple text file, **stata.log**, and it is located in **D:\StataCourse\tmp**.⁵


The following unofficial commands are helpful for inspecting the log:

- nlog** Opens the log file in NoteTab Light, a rather simple text editor. Here you can read it, print it, or print selected part of it. It works much like a word processor, but without the formatting options.⁶
- newlog** deletes the log file and opens a new log.

Besides the general output log described above, you can create a specific output log, typically to preserve the results from an analysis. Imagine that we perform a regression analysis and want to preserve the output from this analysis. This is important, so we do it with a do-file. Note the **capture log close**, the **log using ...**, and the **log close** commands.

```
* regres1_bwt.do
cd C:\docs\proj1
capture log close            // Closes log in case one is open
log using regres1_bwt.log , replace
use lbw1.dta , clear
regress bwt i.race i.smoke lwt
testparm i.race
predict rbwt if e(sample) , residual
histogram rbwt
rvfplot , yline(0)
log close
```

⁵ This is not default behavior by Stata, but something we made for the course. If you want your own computer to generate an output log automatically, study section 1.6 in ISHR3.

⁶ Find a short description of NoteTab Light at <http://www.folkesundhed.au.dk/uddannelse/software>. Note especially that if you want to print a selected part of the text, press the *Shift* key while clicking the Print icon, and select  Selection.