Lab 2 Notes

# Creating one-way[[1]](#footnote-1) tables in Stata

## Review

*Can you perform all Stata actions referred to below? If you cannot, you need to work through Lab 1 notes again:*

1. *Can you find in Stata the Results Window, the Variables Window, the Data Editor and Variables Manager? Can you explain what their purpose is?*
2. *Can you open/close a new do-file and a new dataset? Can you open/close an existing do-file or a dataset?*
3. *Can you create and explain the purpose of a) Variable names b) Variable Labels and c) Value Labels?*

## Learning objectives for Lab 2

*After this session you should know:*

1. *How to open and save a dataset using a do-file*
2. *How and when to use* tabulate, tab1 *with or without ,*miss
3. *How to get Stata to display numeric codes and value labels in tables*
4. *How to use Stata’s help menu*
5. *What to do with missing data and the so-called ‘non-committals’ when preparing tables for your assignments.*

## Opening a data set using a do-file

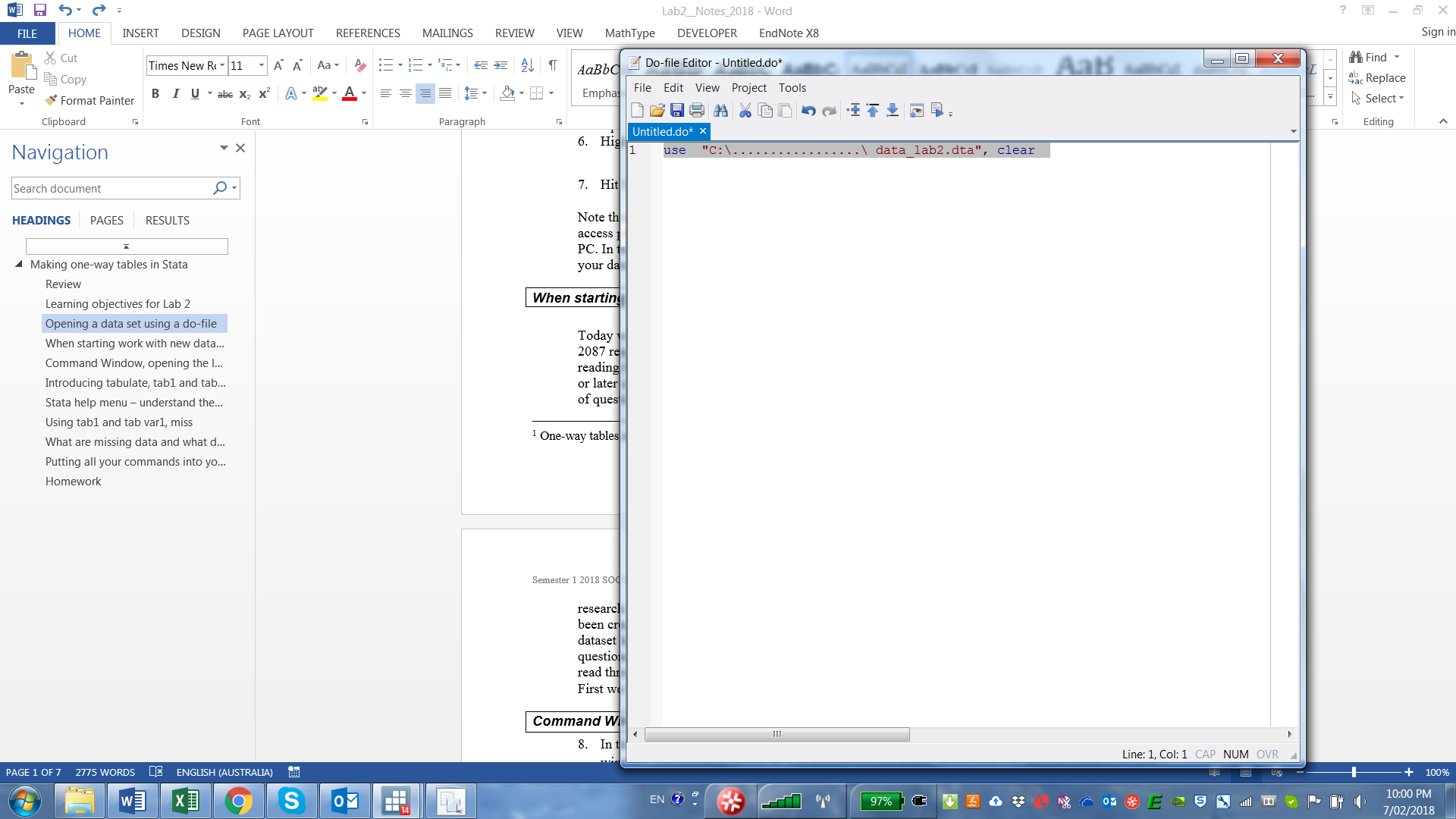
First, let us open an already created dataset. It is a larger, real dataset from the questionnaire we used last week.

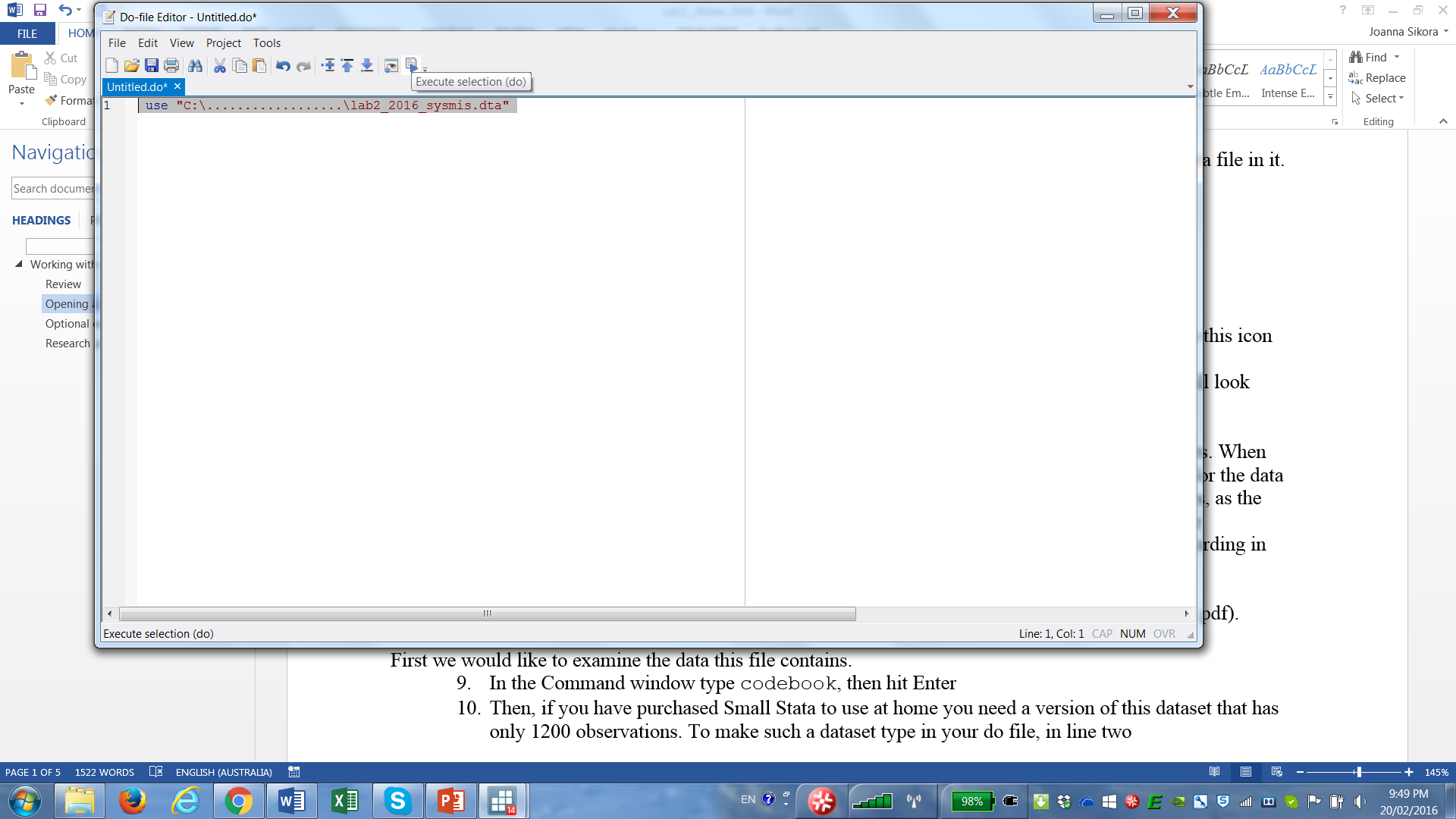
1. Download from Wattle the file called data\_lab2.dta Open it.

Note that once you have opened your file, Stata writes, in the Results a command what looks like this

use "H:\.................\ data\_lab2.dta" , clear

1. From your Results Window, open a new do-file (Ctrl-9 or click Window ->Do-file Editor -> New Do-file Editor) and copy this line from your Results Window into your do-file. Add , clear at the end if it is not there.
2. Save your new do-file on your Home Drive, where you *will be* able to find it later. Give it a name you like.
3. Close Stata.
4. Open Stata again. Stata opens but there is no data file in it. Hit Ctrl-9.From the window that opens, open your saved do-file.
5. Highlight the text of your saved command



1. Hit Ctrl D or click the Execute Do-file icon  (we introduced you to this icon last week).

In the Variables window you will see that you have now opened your data file by executing a do-file command. Note also that your access path will look different to the one in the image above, depending where the data file is on your PC. In this course the do-file you submit for Assignment 2 should start from use followed by the File Path in Windows i.e. use "C:\........\ xxx.dta" , clear

The xxx.dta must replace with the name of one of the assignment data files given to you in Wattle.

## To work with new data first read the questionnaire and explore the derived variables

Today we will work with real data which are an extract from the 2003 AuSSA survey conducted with 2087 respondents. When starting work with a new dataset you should always begin by carefully reading the questionnaire for your data file. Not acquainting yourself with the questionnaire will sooner or later cause problems, as the information in the dataset is mere approximation of the wording of questions and you do need to know **exactly** what questions were asked. Some data sets will also have **derived variables** which are not in the questionnaire but have been created by data owners. You need to understand these by referring to survey documentation. Our lab data set has no derived variables (but your Assignment datasets most likely will). The questionnaire for this survey is in Wattle in Files for Lab 2. (AussA2003\_excerpt\_touse.pdf). Open and read the questionnaire at home (we will not have the time to do this in class).

First, we would like to examine the data this file contains.

## Command Window, opening the log file and running the codebook command

1. In the Command Window type codebook, then hit Enter. Examine the information in the Results window. Ask questions if you do not understand something. While you can execute commands from the Command Window, they are not kept in your do-file. Therefore, it is better to switch to your do-file and execute codebook from there. The amount that Stata writes in the Results window in response to codebook is very large. We have some information about every variable in this file. Try codebook pwrbigb and then try codebook pwrbigb pwrfedg
2. Every time we work with Stata it is a good idea to keep a log of everything we do, just to back up in case we lose our do-file or need to prove to someone that we did the work. A log file keeps all the results of our work plus adds a time stamp. You will need to submit a log file in this course (plus a doctors’ certificate), if you miss a lab due to an illness and want to make up your participation credit. To start a new log file click File>Log>Begin and save your log file where you can find it. Stata will save all of your activity in this log file until the end of your current session. Log files are \*.smcl or \*.log files. The former can be only opened in Stata, the latter is an ASCII text file.

## Introducing tabulate, tab1 and tabulate var, miss

One-way tables enable us to examine **univariate** distributions of variables (*univariate* means involving one variable only). You should never do any complex analyses without first looking at one-way tables of your variables of interest. Today, we will run some commands in Stata, appending the commands to our do-file which later can be used to re-run them or modify them to do something else.

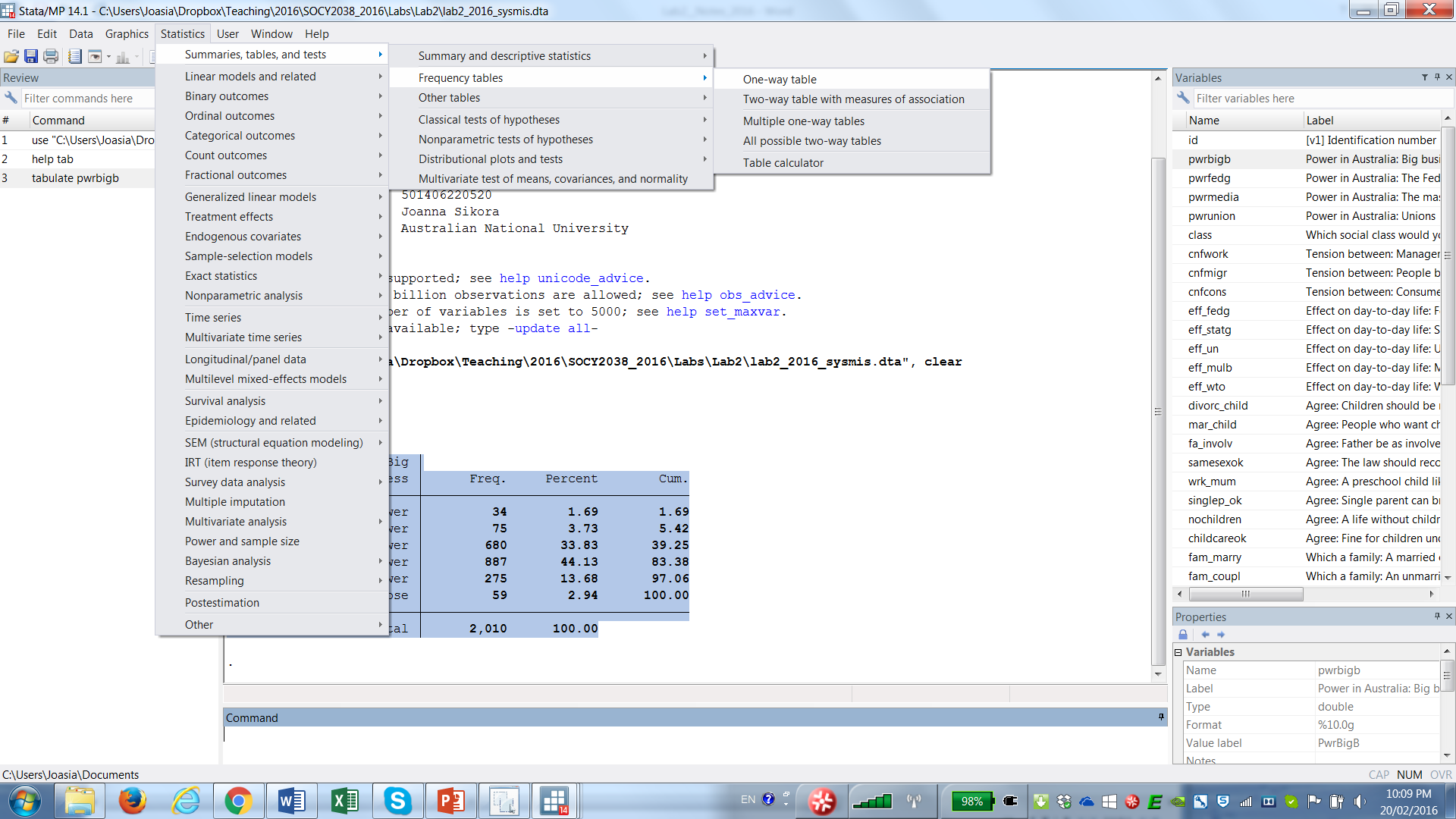
tabulate is a Stata command that returns a one-way table of frequencies (i.e. it puts together respondents’ answers to a specific question.

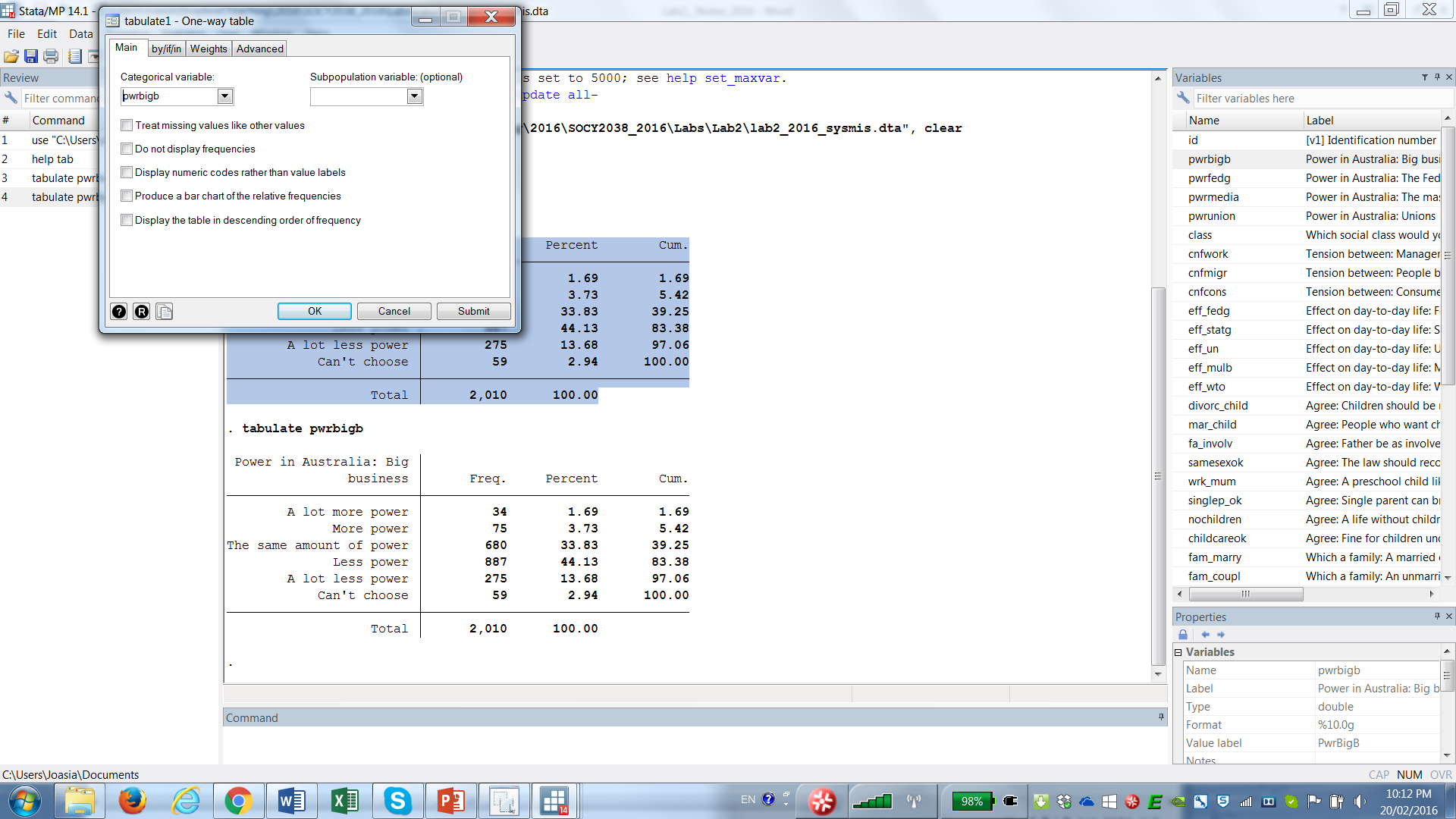
1. In the Do-file type tabulate pwrbigb
2. Note that I actually never type variable names. I copy their names from the Variables pane by right-clicking and copying. Run tabulate pwrbigb This is what you should see in your Results Window:

|  |  |
| --- | --- |
|  |  |

1. (To explore at home) As an alternative to steps 10 through 11 you can use the pull-down menu

Statistics->Summaries, tables and tests, Frequency tables -> One-way table:



Then select pwrbigb and hit OK

1. Examine your Results Window (i.e. what you can see in Step 11).

You should see that only 34 people thought that big business should have more power than what it held – this is under Freq. (frequencies, counts or number of cases or observations). These people made up 1.69 Percent. You should be able to work out where 1.69 came from (34 divided by 2010 – can you work out where other percentage came from?). Finally, Cum. (Cumulative percent) tells us what proportion of the sample was in any given category plus all categories with lower values than that category.

Last week, to make a data set we first entered numeric values and then attached some words to them. But in this dataset for this variable do we have some numeric values? Yes, this variable is numeric (run codebook pwrbigb)which means that it is stored as numeric values. But now, how do we know which words go with which values? Stata by default displays either values or labels but not both. To see values run this command from your do-file:

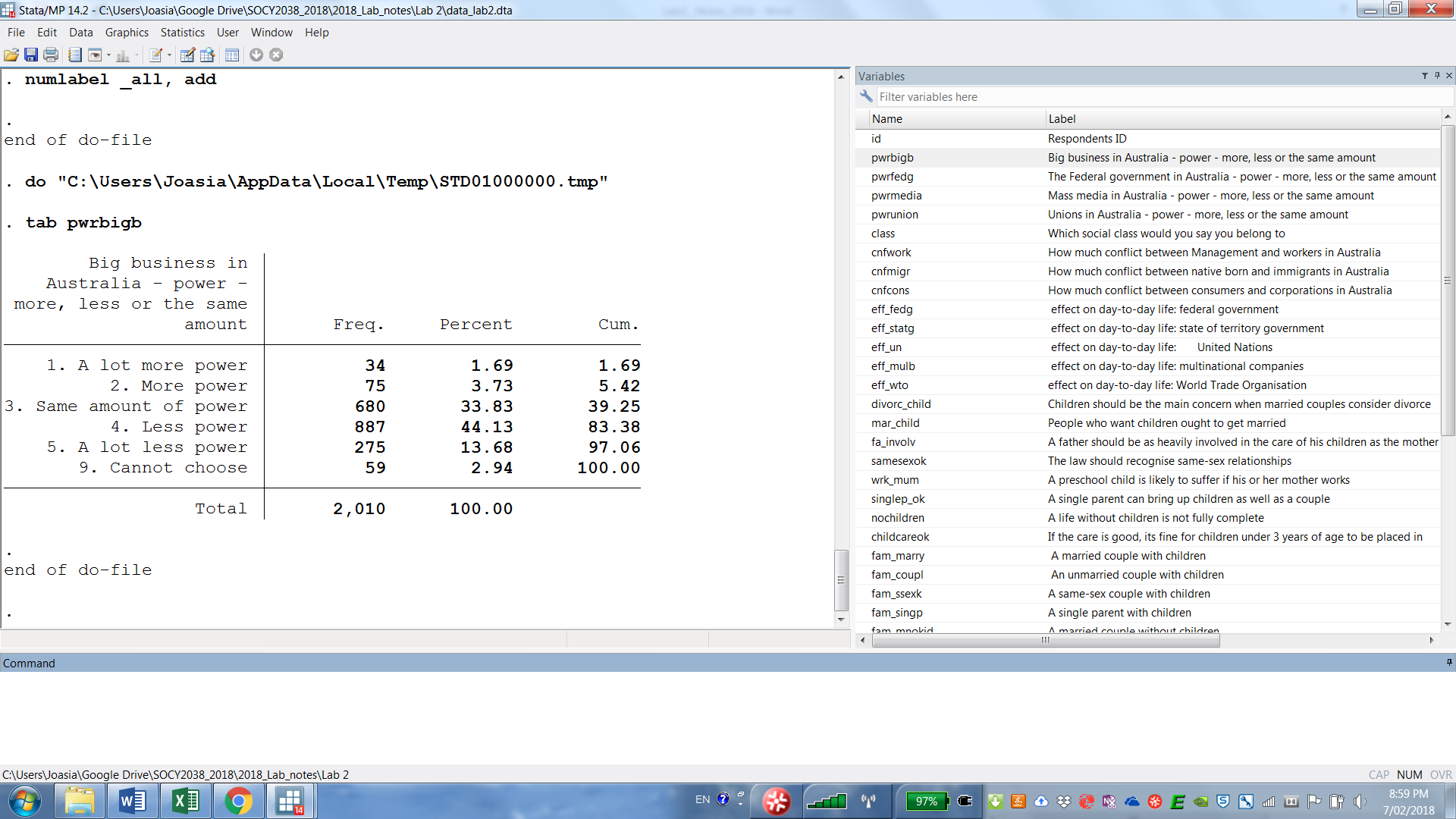
1. tab pwrbigb , nolabel

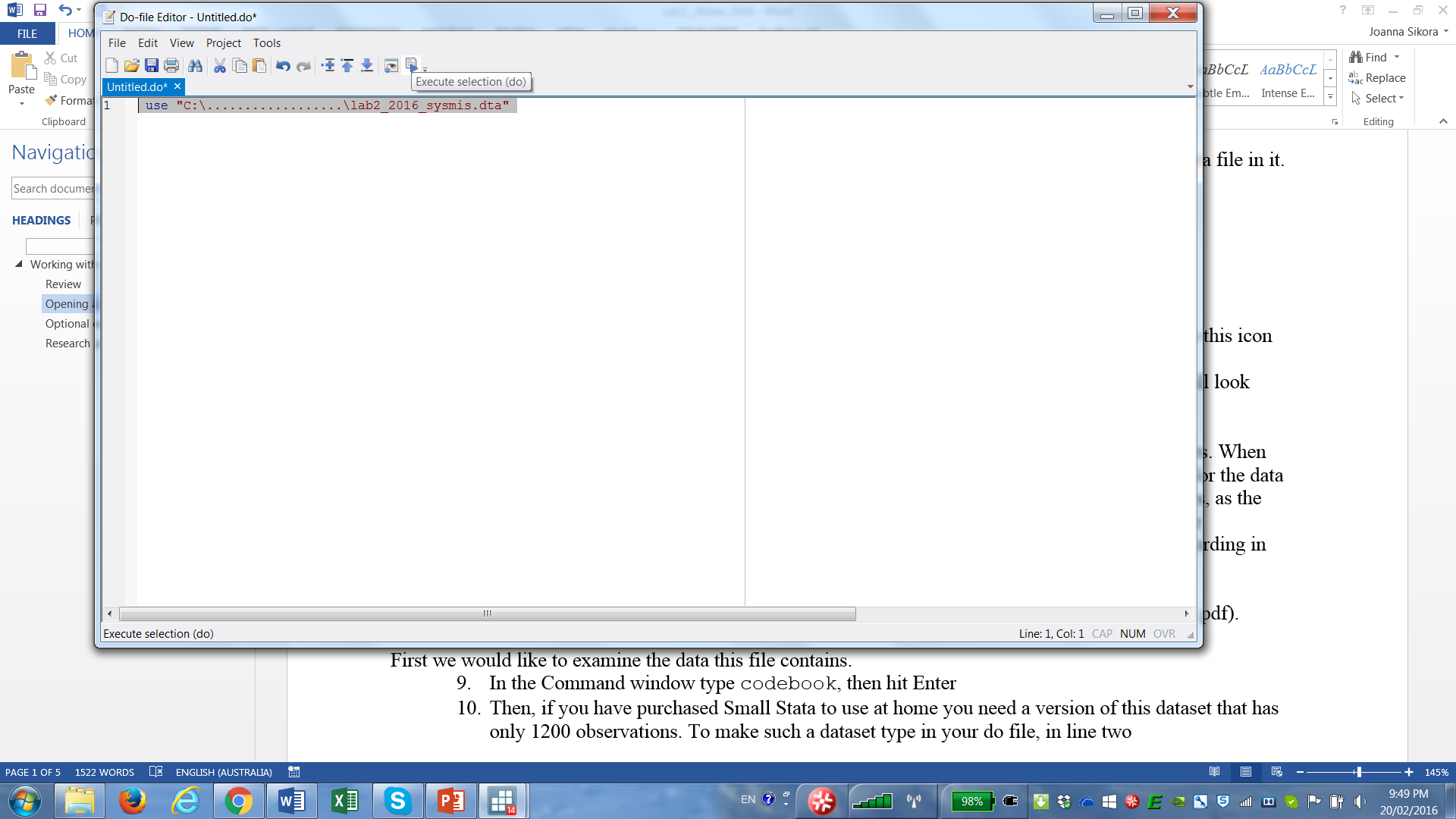
It should show you your variable’s values. But you want to see both values and labels, not either or. To see them, we must attach values to our labels as part of the label. This can be done for this variable using this command (the command comprises the command word followed by the nickname – look the nickname up in Variables Manager - and the option ,add ):

1. numlabel power, add

If all your variables are numeric (check by running codebook or in Variables Manager), you can use

1. numlabel \_all, add from your do-file window
2. Now run tabulate pwrbigb from the Command Window (you can hit the PageUp button to find it if you used it before). This what you should see now



1. Now run tabulate pwrfedg from your do-file window by highlighting and hitting Ctrl-D or the execute icon 
2. Your Results Window should now show:

The Federal government |

in Australia - power - |

more, less or the same |

amount | Freq. Percent Cum.

------------------------+-----------------------------------

1. A lot more power | 90 4.47 4.47

2. More power | 383 19.02 23.49

3. Same amount of power | 1,108 55.01 78.50

4. Less power | 322 15.99 94.49

5. A lot less power | 61 3.03 97.52

9. Cannot choose | 50 2.48 100.00

------------------------+-----------------------------------

Total | 2,014 100.00

## Stata help menu – understand the Stata language

1. Okay, we know that we should not try to type variable names but copy and paste them. To learn more about the Stata syntax, we must learn to use Stata Help menu. Most commands in Stata can be looked up by typing help followed by the command in the Command Window. Let us learn more about tabulate
2. To begin, type help tabulate in your Command window, then click tabulate oneway

Syntax

One-way table

tabulate varname [if] [in] [weight] [, tabulate1\_options]

Note that the letters underlined in tabulate are the abbreviations you can use. To Stata it makes no difference whether you type ta pwrfedg or tabulate pwrfedg

In the command syntax shown in the help menu varname refers to your variable i.e. you are meant to use your variable name instead of varname. For now we will ignore if in and weight but note that we have already used one of the options that follows a comma, i.e. , nolabel Find this option and read its description in the help menu. The help menu enables us to teach ourselves the Stata language to become better Stata programmers.

Note that tabulate varname allows to enter only one variable but

tab1 varlist allows to use several variables in one command. Let us try that.

## Using tab1 varlist and tab varname, miss

1. Switch back to your do-file and tabulate tab1 pwrbigb pwrfedg

Note that we have 2010 people who answered pwrbigb and 2014 who answered pwrfedg

This means that not everyone answers all questions. Let us use another option for tabulate, which will show how many people did not answer, i.e. the missing data:

tab1 pwrbigb pwrfedg, missing

This is what you should see in your Results Window:

tab1 pwrbigb pwrfedg, miss

-> tabulation of pwrbigb

Big business in |

Australia - power - |

more, less or the same |

amount | Freq. Percent Cum.

------------------------+-----------------------------------

1. A lot more power | 34 1.63 1.63

2. More power | 75 3.59 5.22

3. Same amount of power | 680 32.58 37.81

4. Less power | 887 42.50 80.31

5. A lot less power | 275 13.18 93.48

9. Cannot choose | 59 2.83 96.31

. | 77 3.69 100.00

------------------------+-----------------------------------

Total | 2,087 100.00

-> tabulation of pwrfedg

The Federal government |

in Australia - power - |

more, less or the same |

amount | Freq. Percent Cum.

------------------------+-----------------------------------

1. A lot more power | 90 4.31 4.31

2. More power | 383 18.35 22.66

3. Same amount of power | 1,108 53.09 75.75

4. Less power | 322 15.43 91.18

5. A lot less power | 61 2.92 94.11

9. Cannot choose | 50 2.40 96.50

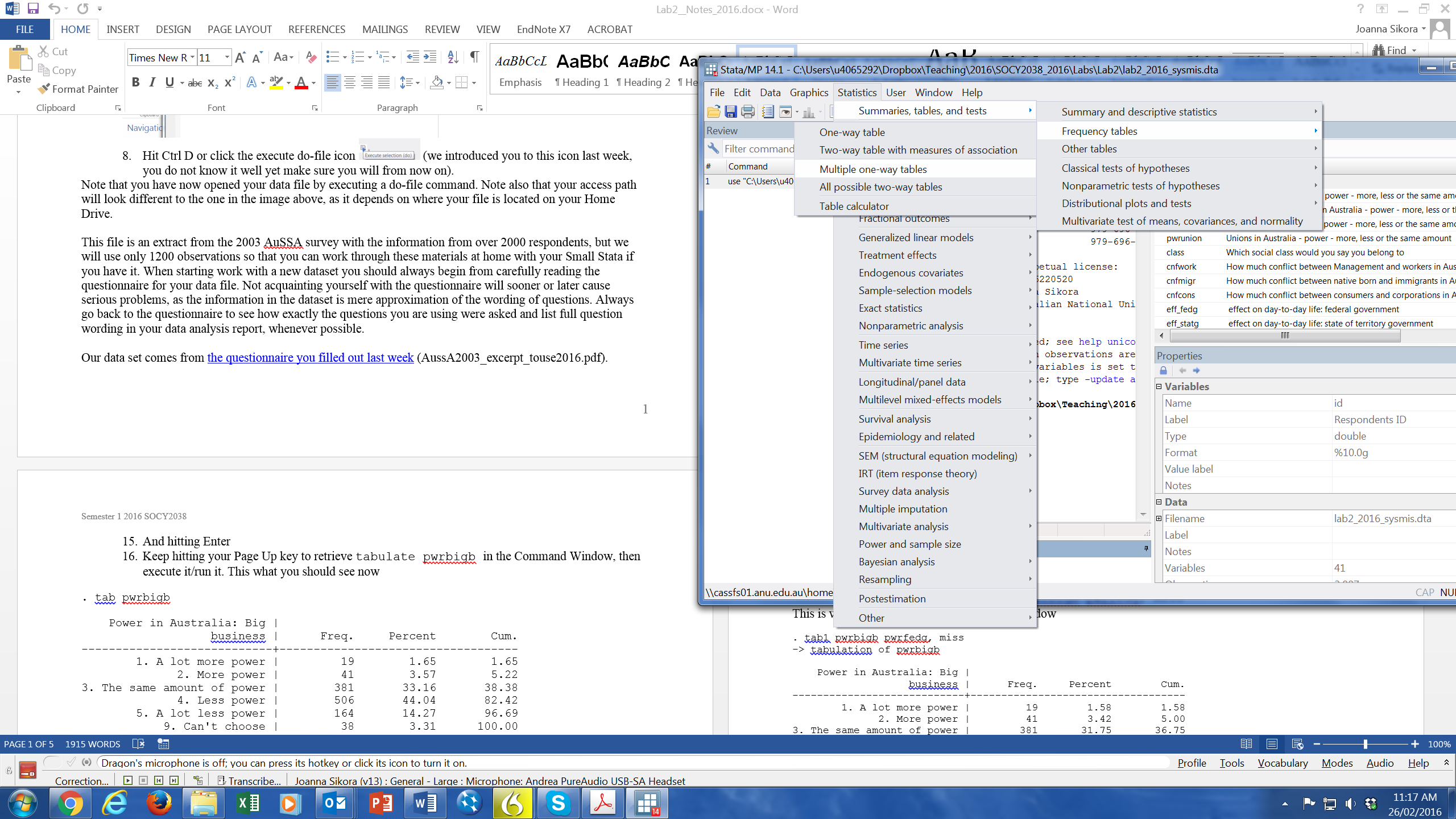
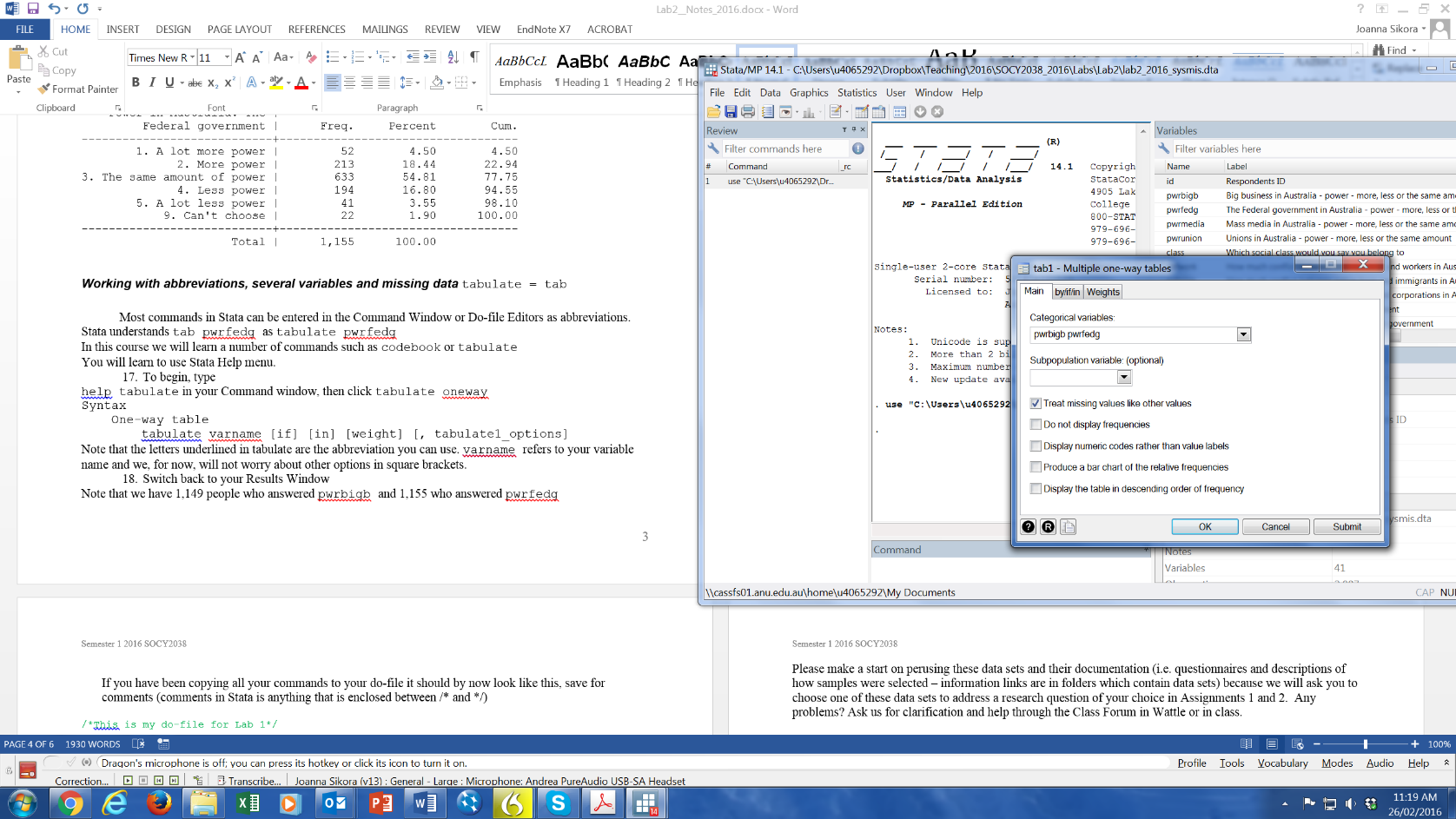
. | 73 3.50 100.00

------------------------+-----------------------------------

Total | 2,087 100.00

By typing tab1 rather than tab you have told Stata that you want it to run univariate frequencies of two variables in a sequence. By typing: ,missing at the end of your command you told Stata that you wanted to see how many people skipped/did not answer these two questions. These people are shown by dots. These are missing data. 77 or 3.69% of people did not answer the first question and 73 or 3.5% of people did not answer the second question (make sure you can explain how these percentages are calculated).

You can get Stata to produce the same tabulations by using drop-down menus



and then choosing two variables and ticking the appropriate option from the menu below

## What are missing data and what to do with them for assignments in this course

A dot is the Stata symbol for system missing values. We call them system missing because there are other types of missing values, i.e. user missing values (when, for instance we do not expect an answer from a particular respondent because the question is not relevant to them etc). In future labs we will learn about other types of missing values. However, please bear in mind that:

In this course, in assignments we will **exclude** missing data and the so-called non-committal answers (i.e. “cannot choose”) from calculating percentages for our tables. Percentages must be calculated using only the data from respondents who provided valid responses. For instance, the correct way of presenting pwrbigb is the table on the left below. If you worry, that once you do that, your table will not add up to 100% (some people worry about that because they have not worked with percentages for some time) see FAQ in Wattle.

For your assignment tables, including Assignment 1, you must convert all user missing data and non-commitals to system missing and then use tab varname without the miss option. See example of edited tables for pwrbigb below (you must have edited tables for Assignment 1, i.e. change the layout and font so that your tables look like the tables below, not like Courier font Stata output on the previous page). Your percentages should be calculated without missing data or non-commitals.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Table 1: Big business should have🗸correct | | |  | Table 1: Big business should have 🗴incorrect | | |
|  |  | Percent |  |  |  | Percent |
|  |  |  |  |  |  |  |
| A lot more power | | 2% |  | A lot more power | | 2% |
| More power | | 4% |  | More power | | 4% |
| Same amount of power | | 35% |  | Same amount of power | | 34% |
| Less power | | 45% |  | Less power | | 44% |
| A lot less power | | 14% |  | A lot less power | | 14% |
|  |  |  |  | Cannot choose | | 3% |
| N |  | 1,951 |  | N |  | 2,010 |

Data: AuSSA 2003

We will learn how to convert user missing values and non-committals into system missing values next week, so do not worry you cannot yet produce the table on the left. You will be able to make one next week.

## Putting all your commands into your do-file to re-run them later

If you have been copying all your commands to your do-file it should by now look like this, save for the comments (comments in Stata is anything that is enclosed between /\* and \*/ or follows \* in one line)

/\*This is my do-file for Lab 2\*/

use "C:\........\data\_lab2.dta", clear

log using "C:\..........\lab2.log”, replace

\*I am examining my data

codebook

/\*I am running one-way table for pwrbigb\*/

tabulate pwrbigb

tabulate pwrbigb , nolabel

/\*I am adding numeric codes to pwrbigb to do that I need to look up what nickname pwrbigb has for its labels the label nickname is power

\*numeric values will be added to all variables that have this nickname \*/

numlabel power, add

/\*I am adding numeric codes the labels of all numeric variables\*/

numlabel \_all, add

/\*I am practising the use of tab instead of tabulate\*/

tab pwrbigb

tab pwrfedg

/\*I am running one-way tables for two variables and I ask for missing data\*/

tab1 pwrbigb pwrfedg, miss

1. Check that you know how to put your comments in your do file, save your do-file and know how to reopen it later.
2. Stata will close your log file when you exit.
3. Do not save your dataset.

## Homework

In preparation for your independent research project i.e. Assignments 1 and 2 examine some data in one chosen dataset available in Wattle under DATA FOR RESEARCH ASSIGNMENTS

Complete the following steps:

1. Start a new Stata session by opening the downloaded dataset.
2. Open a log file to keep a record of what you are doing for yourself.
3. Find the questionnaire for your dataset in Wattle.
4. Use the questionnaire to choose questions that interest you and run some one-way tables on them (one at a time please!).
5. Try to use the commands we have learned today.
6. If any problems arise, please write them up and post a query to the Class Forum in Wattle.

If running tabulate in the data for research assignment data turns out to be too hard, open data\_lab2.dta and the questionnaire for this data set (which you know better) and see if you can manage to use the commands learned today with understanding of what you are doing. Follow steps 1- 6 above and

make sure you have run one-way tables for at least five variables, understand what they show and can describe them (try writing a better description of one of your tables than the description of Table 1 given below – just bear in mind that we always refer to percentages in our descriptions, not to counts i.e. numbers of people. Counts and frequencies are equivalent terms that we will use in this course and you should know that counts are not percentages and vice versa).

|  |  |  |
| --- | --- | --- |
| Table 1: Big business should have | | |
|  |  | Percent |
|  |  |  |
| A lot more power | | 2% |
| More power | | 4% |
| Same amount of power | | 35% |
| Less power | | 45% |
| A lot less power | | 14% |
|  |  |  |
| N |  | 1,951 |

Data: AuSSA 2003

*Table 1 shows how a representative sample of adult Australians answered, in 2003, the following question: “Thinking about the amount of power organizations have in Australia today, please say whether you think big business should have more power, less power or the same amount.” The largest group of respondents, 45%, felt that big business should have less power, with 14% feeling even more strongly, that big business should have a lot less power. Few people thought that large corporations in Australia ought to wield more power than they did. Only 2% said “A lot more power” while 4% opted for “More power”. 35% of respondents thought that the status quo was best and big businesses should have just as much sway as they did at the time. All in all, Australians tended to think that large corporations were too powerful for their liking.*

1. One-way tables are also called one-way tabulations or frequency distributions. One-way tables contain only one variable per table. We will learn later about two-way tables and three-way tables. [↑](#footnote-ref-1)