

# Introduction to Statistical Computing in Clinical Research

---

BIOSTAT 212, Summer 2017



Kristen Aiemjoy  
Department of Epidemiology and Biostatistics  
University of California San Francisco

# Outline today

---

## ◇ Course details

## ◇ Lecture 1

- Why STATA?
- Orientation to STATA
- Finding help
- Command syntax
- Working directories
- .do files
- log files
- Opening data sets
- Exploring data (basics)

# Introductions

---

## ◊ Teaching Assistants

- Phil Perera
- Alyssa Mooney
- Rae Wannier
- Sankari Ayyaluru
- David Hoskins

# Course Objectives

---

- ◇ Introduce you to using STATA for:
  - Data management
  - Basic statistical and epidemiologic analysis
  - Turning raw data into presentable tables, figures and other research products
- ◇ Prepare you for fall courses
- ◇ Start analyzing your own data

# Course Objectives

---

- ◊ NOT statistical theory
  - Biostat200 in the fall is the first biostats class

# Course details

---

- ◇ 1 Unit Course
- ◇ 7 Lectures
  - Tuesdays 1:15 – 2:45
- ◇ 7 Computer labs
  - Tuesdays 3:00 – 4:00
- ◇ 5 Assignments
- ◇ Final project

# Course details

---

## ◊ Office hours

- Online: Wednesdays 10:00-11:00
- In person: Thursdays 11:00 – 12:00 [MH 2800]

# Course Details

---

## ◇ Course Website

- <https://courses.ucsf.edu/course/view.php?id=3998>

## ◇ Recorded Lectures

- Audio + video of lecturer + video of screen
- Available same day for viewing
- Links posted on website syllabus

# Course details

---

- ◇ Online forum
  - Please post questions here
- ◇ TA's will staff forum, reply by end of day on weekdays
- ◇ Before you post, see if it's already there and answered
- ◇ Settings for email notifications
  - <https://learningtech.library.ucsf.edu/-/?s=forum+email+notifications>

# Course details

---

## ◇ Course Requirements

- Hand in all 5 assignments (even if late)
- Satisfactory final project

# Course details

---

## ◊ Grading (not relevant for all students)

- Letter grades: Standard cutoffs

|                                        |   |
|----------------------------------------|---|
| ○ 90-100%                              | A |
| ○ 80-89%                               | B |
| ○ 70-79%                               | C |
| ○ 60-69%                               | D |
| ○ <60% or Course Requirements not met: | F |

- Satisfactory/Unsatisfactory

|        |              |
|--------|--------------|
| ○ >80% | Satisfactory |
|--------|--------------|

- TICR Grading policy

- [http://www.epibiostat.ucsf.edu/courses/masters/grading\\_policy.html](http://www.epibiostat.ucsf.edu/courses/masters/grading_policy.html)

# Lecture Topics

---

1. Introduction to Stata
2. Exploring and describing data
3. Data management and cleaning
4. Graphs and figures
5. Commands for statistical analyses
6. Epi commands and tables
7. Automating, programming and generating synthetic data

# Computer labs

---

- ◇ Mini lectures
- ◇ Labs (these do not need to be turned in)
  - Live coding
- ◇ Time to work on you assignments
- ◇ Time to ask questions and get help

# Assignments

---

## ◇ 5 Assignments

- ◇ Assignments are due 1 week after assigned
  - Please upload before lecture starts. **By 1:00 PM**
- ◇ Assignments turned in late will receive only half credit
- ◇ ALL assignments must be turned in to pass the class

# Lecture 1: Introduction to Stata

---

Introduction to Statistical Computing in Clinical Research  
BIOSTAT 212, Summer 2016



Kristen Aiemjoy  
Department of Epidemiology and Biostatistics  
University of California San Francisco

# Lecture 1: Outline

---

- **Why STATA?**
- Orientation to STATA
- Finding help
- Command syntax
- Working directories
- .do files
- log files
- Opening data sets
- Exploring data (basics)

# Why use STATA?

---

- ◇ Quick start, user friendly, but full-featured and powerful
- ◇ Command line and menu-driven option
- ◇ Good? graphics
- ◇ Good manuals, help menu, customer service
- ◇ UCSF/ TICR

# Why not STATA?

---

- ◇ \$\$\$
- ◇ R and Epi-info are free
- ◇ Programming more intuitive in R
- ◇ Graphics are really good/flexible in R
- ◇ In STATA, you can only have one data set open at a time

# STATA vs. other software packages

---

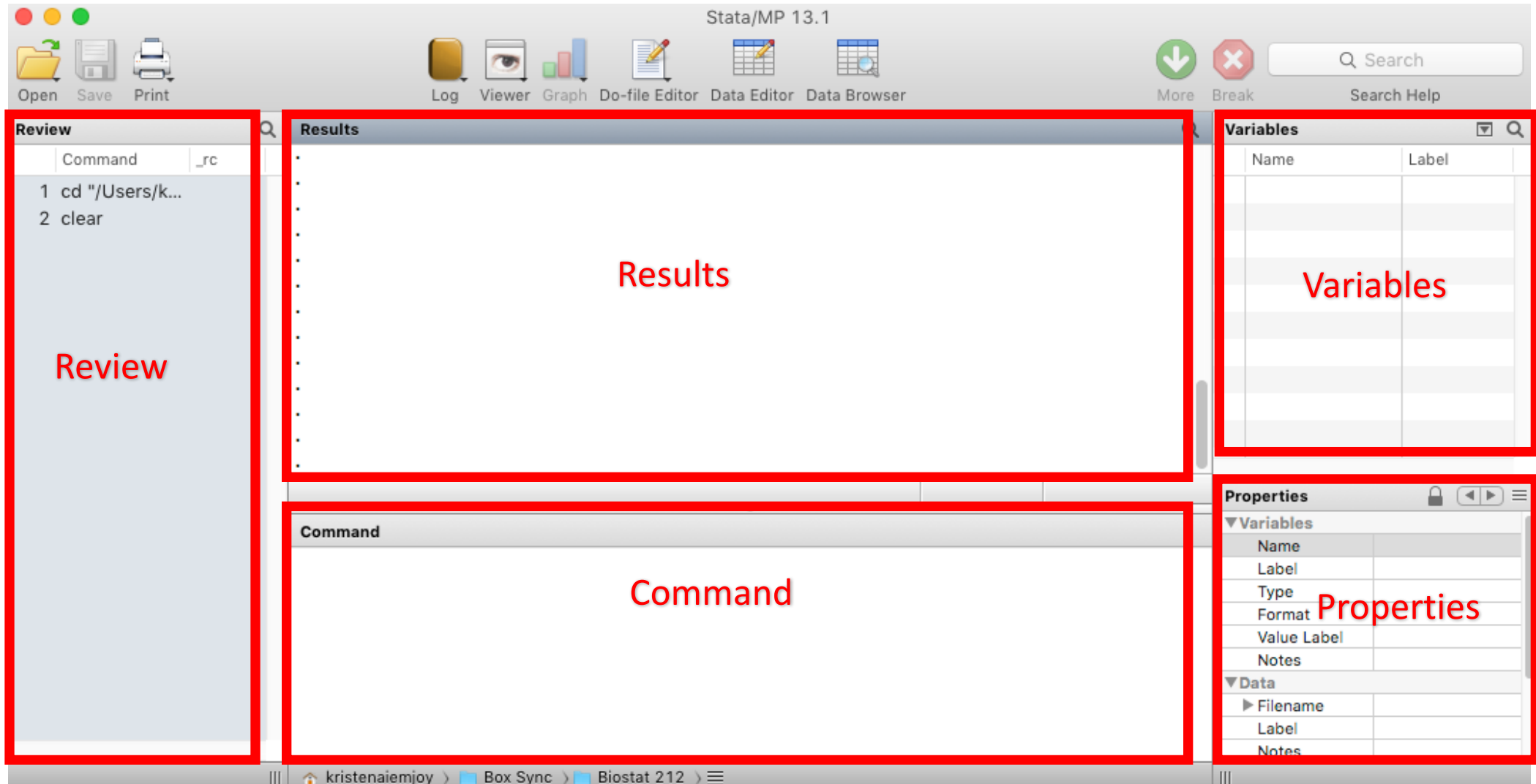
| Features          | Stata                                                    | SPSS                                                                | SAS                        | R                  |
|-------------------|----------------------------------------------------------|---------------------------------------------------------------------|----------------------------|--------------------|
| Learning curve    | Moderate                                                 | Gradual/flat                                                        | Pretty steep               | Pretty steep       |
| User interface    | Programming/point-and-click                              | Mostly point-and-click                                              | Programming                | Programming        |
| Data manipulation | Very strong                                              | Moderate                                                            | Very strong                | Very strong        |
| Data analysis     | Powerful                                                 | Powerful                                                            | Powerful/versatile         | Powerful/versatile |
| Graphics          | Very good                                                | Very good                                                           | Good                       | Excellent          |
| Cost              | Affordable (perpetual licenses, renew only when upgrade) | Expensive (but not need to renew until upgrade, long term licenses) | Expensive (yearly renewal) | Open source        |

# Lecture 1: Outline

---

- Why STATA?
- **Orientation to STATA**
- Finding help
- Command syntax
- Working directories
- .do files
- log files
- Opening data sets
- Exploring data (basics)

# Windows

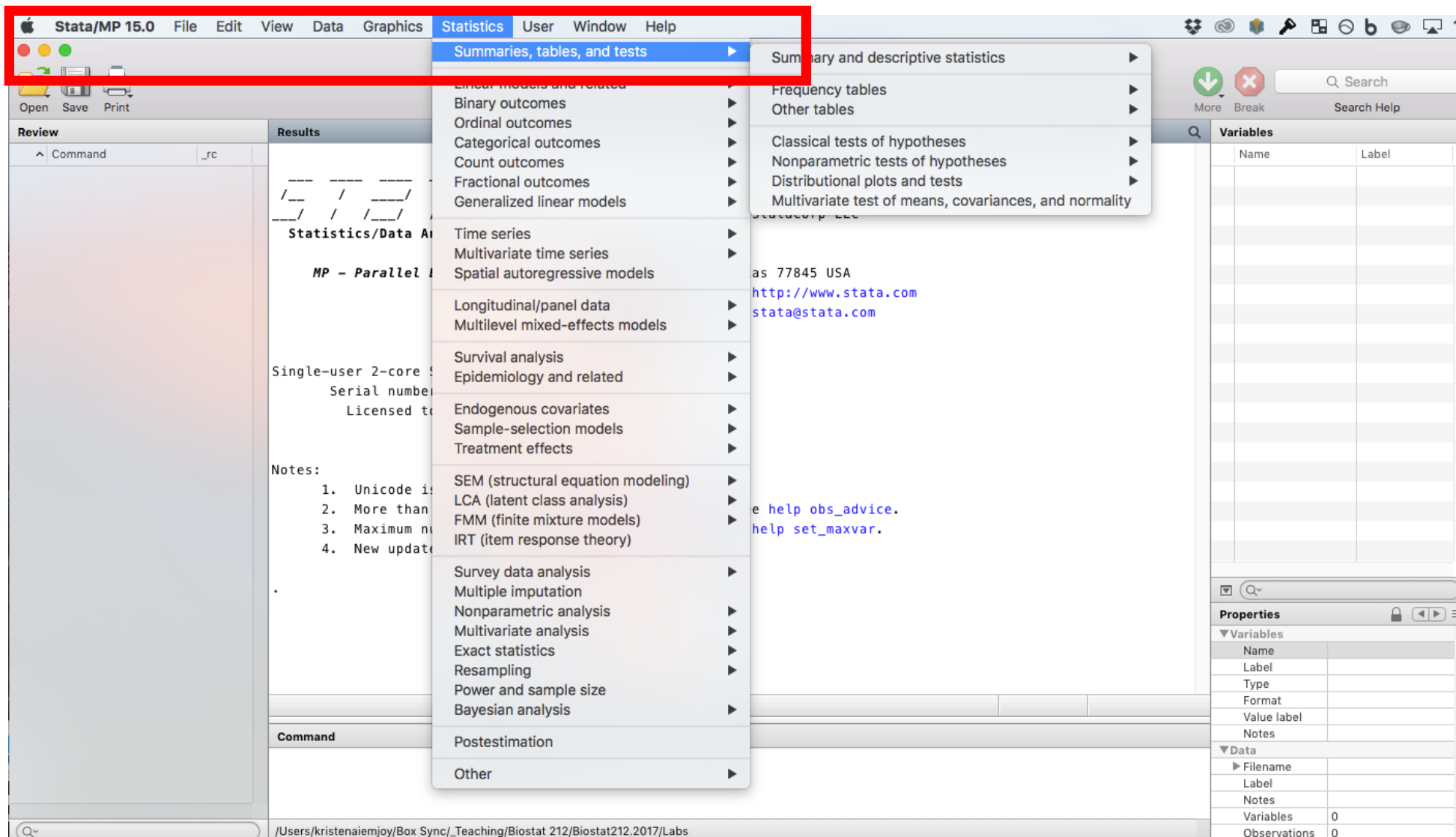


The screenshot shows the Stata/MP 15.0 application window. The 'Graphics' menu is open, displaying a list of plotting options. The main window area shows a command window with the following text:

```
MP -
Single-use
Ser
L
Notes:
1. 
2. 
3. 
4. New update available; type -update all-
```

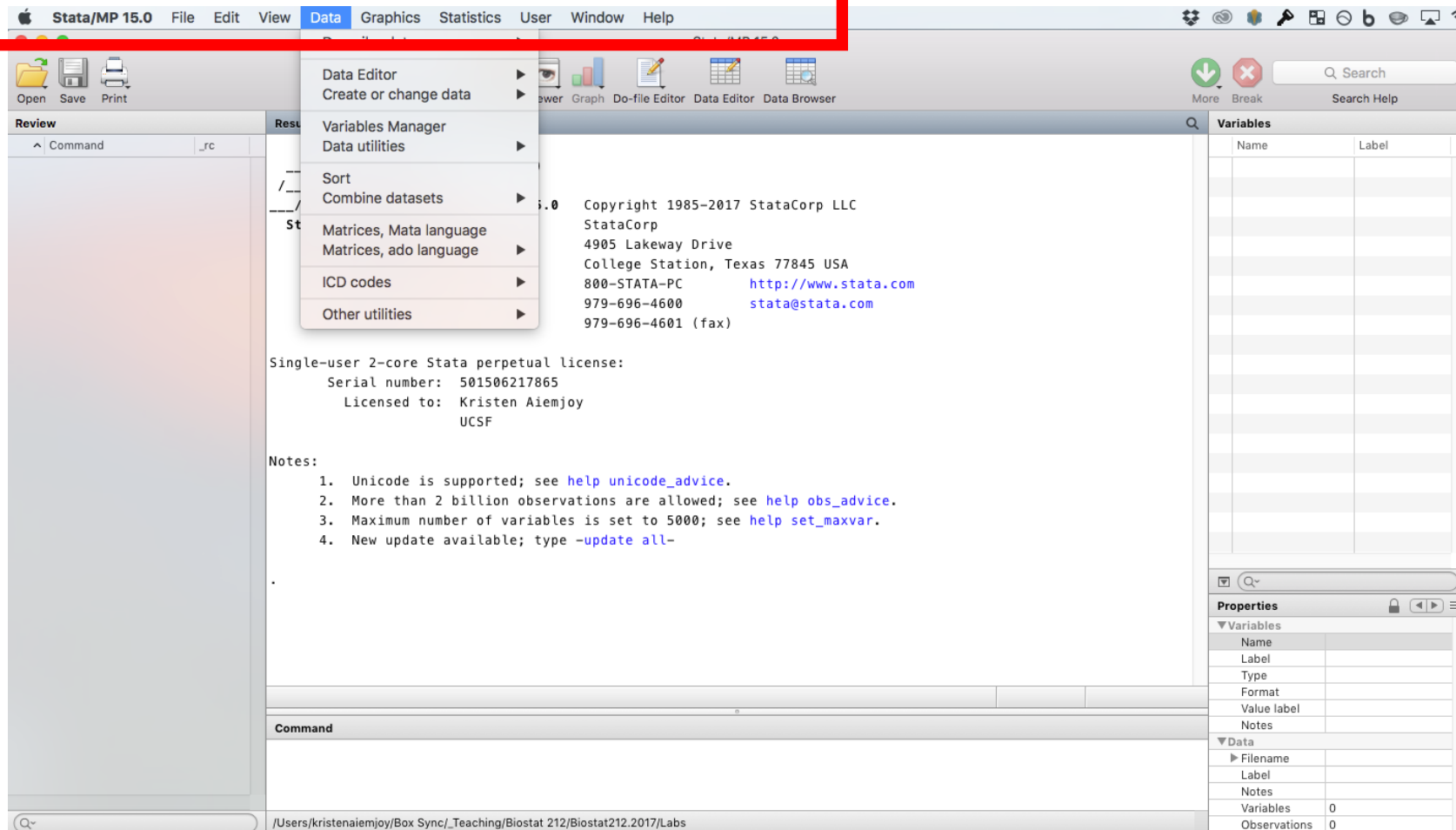
The bottom status bar indicates the current file path: `/Users/kristenaiemjoy/Box Sync/_Teaching/Biostat 212/Biostat212.2017/Labs`.

# The Statistics menu





A screenshot of the Stata/MP 15.0 menu bar. The menu items are: Apple logo, Stata/MP 15.0, File, Edit, View, Data (highlighted in blue), Graphics, Statistics, User, Window, and Help. Below the menu bar, the first three window control buttons (red, yellow, green) are visible.



# The Data Editor

Data Editor (Edit) — cancer.dta

Edit Browse Filter Variables Properties Snapshots

studytime[1] 1

|    | studytime | died | drug | age | _st | _d | _t | _t0 |
|----|-----------|------|------|-----|-----|----|----|-----|
| 1  | 1         | 1    | 1    | 61  | 1   | 1  | 1  | 0   |
| 2  | 1         | 1    | 1    | 65  | 1   | 1  | 1  | 0   |
| 3  | 2         | 1    | 1    | 59  | 1   | 1  | 2  | 0   |
| 4  | 3         | 1    | 1    | 52  | 1   | 1  | 3  | 0   |
| 5  | 4         | 1    | 1    | 56  | 1   | 1  | 4  | 0   |
| 6  | 4         | 1    | 1    | 67  | 1   | 1  | 4  | 0   |
| 7  | 5         | 1    | 1    | 63  | 1   | 1  | 5  | 0   |
| 8  | 5         | 1    | 1    | 58  | 1   | 1  | 5  | 0   |
| 9  | 8         | 1    | 1    | 56  | 1   | 1  | 8  | 0   |
| 10 | 8         | 0    | 1    | 58  | 1   | 0  | 8  | 0   |
| 11 | 8         | 1    | 1    | 52  | 1   | 1  | 8  | 0   |
| 12 | 8         | 1    | 1    | 49  | 1   | 1  | 8  | 0   |
| 13 | 11        | 1    | 1    | 50  | 1   | 1  | 11 | 0   |
| 14 | 11        | 1    | 1    | 55  | 1   | 1  | 11 | 0   |
| 15 | 12        | 1    | 1    | 49  | 1   | 1  | 12 | 0   |
| 16 | 12        | 1    | 1    | 62  | 1   | 1  | 12 | 0   |
| 17 | 15        | 1    | 1    | 51  | 1   | 1  | 15 | 0   |
| 18 | 17        | 1    | 1    | 49  | 1   | 1  | 17 | 0   |
| 19 | 22        | 1    | 1    | 57  | 1   | 1  | 22 | 0   |
| 20 | 23        | 1    | 1    | 52  | 1   | 1  | 23 | 0   |
| 21 | 6         | 1    | 2    | 67  | 1   | 1  | 6  | 0   |
| 22 | 6         | 0    | 2    | 65  | 1   | 0  | 6  | 0   |
| 23 | 7         | 1    | 2    | 58  | 1   | 1  | 7  | 0   |

Vars: 8 Order: Dataset Obs: 48

**Variables**

| Name                                          | Label                |
|-----------------------------------------------|----------------------|
| <input checked="" type="checkbox"/> studytime | Months to de...      |
| <input checked="" type="checkbox"/> died      | 1 if patient di...   |
| <input checked="" type="checkbox"/> drug      | Drug type (1...      |
| <input checked="" type="checkbox"/> age       | Patient's age...     |
| <input checked="" type="checkbox"/> _st       | 1 if record is...    |
| <input checked="" type="checkbox"/> _d        | 1 if failure; 0 i... |
| <input checked="" type="checkbox"/> _t        | analysis time...     |
| <input checked="" type="checkbox"/> _t0       | analysis time...     |

**Properties**

▼ Variables

| Name        | studytime             |
|-------------|-----------------------|
| Label       | Months to death or... |
| Type        | byte                  |
| Format      | %8.0g                 |
| Value label |                       |
| Notes       |                       |

▼ Data

|              |                          |
|--------------|--------------------------|
| ► Filename   | cancer.dta               |
| Label        | Patient Survival in D... |
| ► Notes      |                          |
| Variables    | 8                        |
| Observations | 48                       |
| Size         | 384                      |
| Memory       | 64M                      |

Filter: Off

# Menu versus Command Line

---

## ◊ Menu advantages

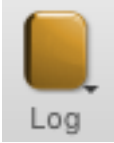
- Browse for commands you don't know already
- See the options for each command in dialog boxes
- Good way to learn syntax for complex commands

## ◊ Command line advantages

- MUCH faster
- ONLY way to write “do” files
  - Document and repeat analyses

# Buttons

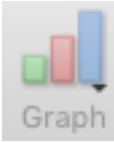
---



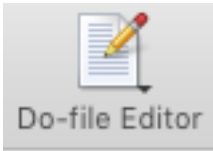
**Log Begin/Close/Suspend/Resume** Starts or closes a file to record results



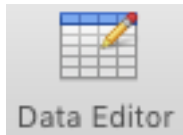
**Viewer** Opens a help window



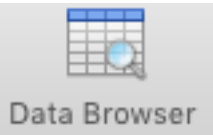
**Graph** Brings Graph Window to Front



**Do-file Editor** Opens a new .do (command) file



**Data Editor** Opens data window and allows edits



**Data Browser** Opens data window to browse but not edit

# Break and More

---

When there are more results than will fit in your result window use:



Or the

Space

To see more



To stop

\*note: you can also use the command `set more on`  
`set more off`

# Lecture 1: Outline

---

- Why STATA?
- Orientation to STATA
- **Finding help**
- Command syntax
- Working directories
- .do files
- log files
- Opening data sets
- Exploring data (basics)

# Help!

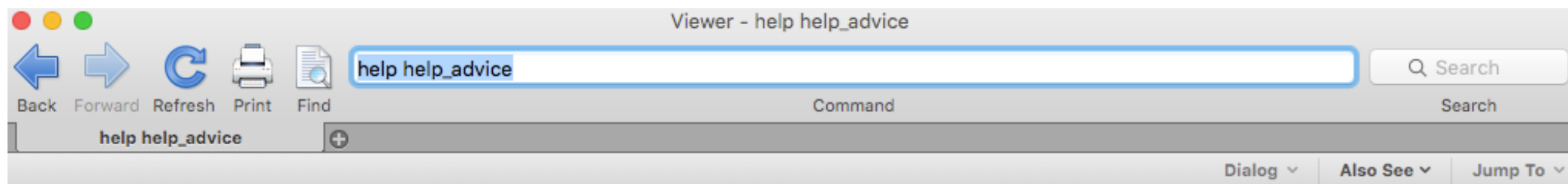
---

.

Command

help tabluate

# Help!



## Advice on finding help

Let's say you are trying to find out how to do something. With over 2,000 help files and 11,000 pages of PDF documentation, we have probably explained how to do whatever you want. The documentation is filled with worked examples that you can run on supplied datasets. Whatever your question, try the following first:

1. Select **Help** from the Stata menu and click on **Search...**
2. Type some keywords about the topic that interests you, say, "logistic regression".
3. Look through the resulting list of available resources, including help files, FAQs, Stata Journal articles, and other resources.
4. Select the resource whose description looks most helpful. Usually, this description will be a help file and will include "(help <xyz>)", or, as in our example, perhaps "(help [logistic](#))". Click on the blue link "logistic".
5. Let's assume you have selected the "(help [logistic](#))" entry. You are probably not interested in the syntax at the top of the file, but you would like to see some examples. Select **Jump To** from the Viewer menu (in the top right corner of the help file now on your screen), and click on **Examples**. You will be taken to example commands that you can run on example datasets. Simply cut and paste those commands into Stata to see the results.
6. If you are new to the **logistic** command and want both an overview and worked examples with discussion, from the **Also See** menu, click on [\[R\] logistic](#) with the PDF icon. Or at the top of the help file, click on the blue title of the

# Help!

[All](#) [Images](#) [Videos](#) [News](#) [Shopping](#) [More ▾](#) [Search tools](#)

About 125,000 results (0.50 seconds)

**Stata | FAQ: Creating percent summary variables**  
[www.stata.com](http://www.stata.com) › [Resources & support ▾](#) [Stata ▾](#)  
How can I create variables containing **percent** summaries? ... sysuse auto (1978 Automobile Data) .  
`tabulate rep78 foreign, row +-----+ ....` For **example**,

**[PDF] tabulate twoway - Stata**  
[www.stata.com/manuals13/rtabulatetwoway.pdf](http://www.stata.com/manuals13/rtabulatetwoway.pdf) ▾ [Stata ▾](#)  
Unix(console) users can instead use the `set linesize` **command** to take advantage of .... `tabulate` can  
calculate column **percentages** and cell **percentages**, as well.

**st: RE: Percentages in cross-tabulation - Stata**  
[www.stata.com/statalist/archive/2002-09/msg00432.html](http://www.stata.com/statalist/archive/2002-09/msg00432.html) ▾ [Stata ▾](#)  
Sep 24, 2002 - The > problem is how do I get column **percentages** of income > without transferring the  
**tabulation** results to XL? See the on-line **help**. . `tab sex ...`

# Lecture 1: Outline

---

- Why STATA?
- Orientation to STATA
- Finding help
- **Command syntax**
- Working directories
- .do files
- log files
- Opening data sets
- Exploring data (basics)

# Command syntax

---

1. I will use `courier` font for commands
2. Words that are NOT meant to be typed in but instead are placeholders for variable names or numbers will be `orange`.
3. Parts of the command that are optional will be placed in `[brackets]`.
  - *When you execute the command with the option you do not need to include the bracket*
4. Shortcuts for commands are indicated by underlining
  - summarize

# Command syntax

---

tabulate varname1 varname2, [row]

`tab male smoke, row`

| male  | smoke          |                | Total            |
|-------|----------------|----------------|------------------|
|       | 0              | 1              |                  |
| 0     | 2,723<br>65.00 | 1,466<br>35.00 | 4,189<br>100.00  |
| 1     | 3,539<br>57.24 | 2,644<br>42.76 | 6,183<br>100.00  |
| Total | 6,262<br>60.37 | 4,110<br>39.63 | 10,372<br>100.00 |

# Command syntax

---

`varlist`

1 or more variables

`varname`

One variable

`varname1, varname2`

2 variables

`#`

A number

# Command syntax

---

- ◇ STATA is case sensitive
  - Commands are in lower case
  - If your variables have upper case letters, you need to remember this when typing them into the command line
    - Often easier to use all lower case

# Lecture 1: Outline

---

- Why STATA?
- Orientation to STATA
- Finding help
- Command syntax
- **Working directories**
- .do files
- log files
- Opening data sets
- Exploring data (basics)

# Working Directories

- ❖ Important! Where all your files will be saved/ retrieved from
- ❖ Create sub folders for different projects
- ❖ To see current working directory:
  - Command line
    - `pwd` displays the path of the current working directory

```
.
. pwd
/Users/kristenaiemjoy/Box Sync
```

Command

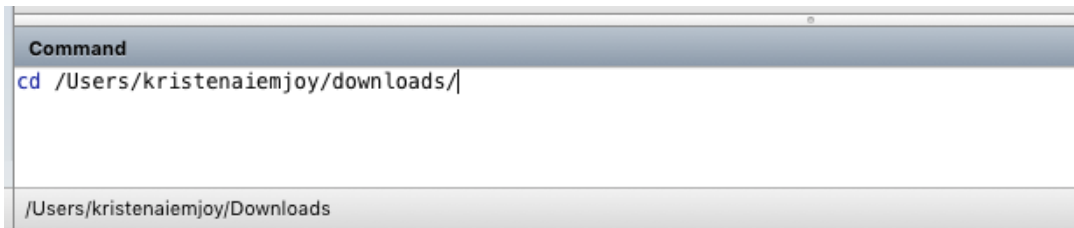
# Change working Directories

## 1. Drop-down menu

- File < Change Working Directory

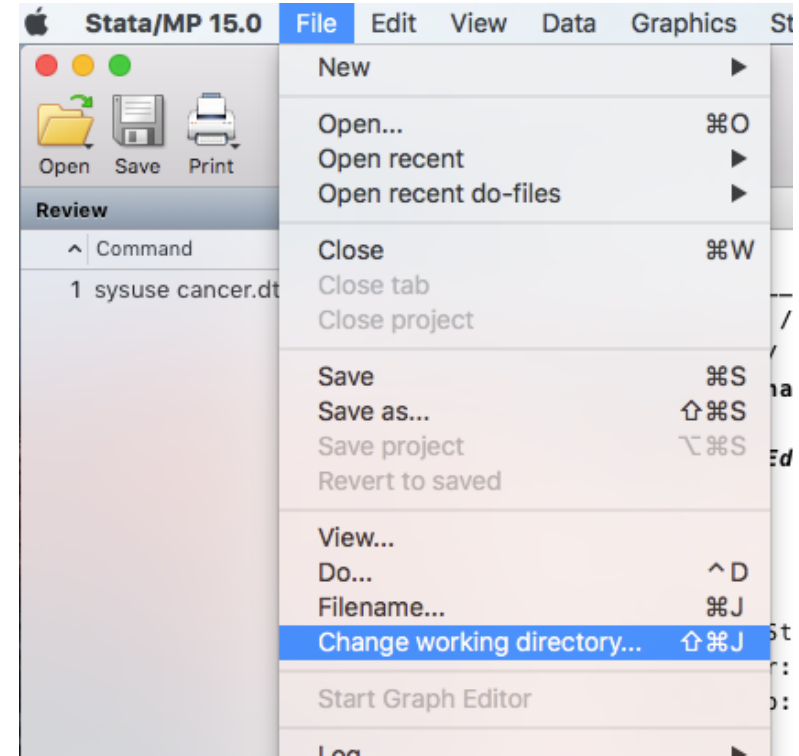
## 2. Command line

- `cd` change current working directory
  - `dir` view contents of directory



```
Command
cd /Users/kristenaiemjoy/downloads/

/Users/kristenaiemjoy/Downloads
```



- *Unix users .. Gets you to the parent folder*
- *cd alone gets you to home directory*

# Working directories: PC vs. Mac

---

## PC Users

"C:\"

"C:\data\"

"C:\data\biostat212"

cd "C:\data\biostat212\"

## Mac Users equivalent

Home directory (also specified by  
"/users/yourusername")

"~/Documents" – your Documents folder in your  
home directory

Folder called "biostat212" in your Documents  
folder

cd "~/Documents/biostat212"

- ◇ *The “~” indicates the use of the Home Directory so you don’t have to specify “/users/yourusername”.*
- ◇ *Mac uses / windows uses \*

# Lecture 1: Outline

---

- Why STATA?
- Orientation to STATA
- Finding help
- Command syntax
- Working directories
- **.do files**
- log files
- Opening data sets
- Exploring data (basics)

# .do files

---

- ◇ A 'recipe' book for your analysis
- ◇ Add comments remind you what is occurring at each stage of analysis
- ◇ Both comments and commands
- ◇ Allows you to repeat analyses quickly
- ◇ Cut and paste
- ◇ Find and replace

# .do file template

---

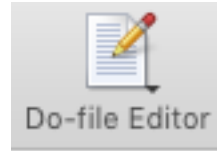
```
1  cd "C:\data\biostat212\"    // state the working directory where your files are
2
3  capture log close           //close any log files that may be open
4  log using "Lab2_yourname.smcl", replace
5
6  set more off                // tells stata not to pause or display the -more- message
7
8  version 13                  // assumes you are using Stata 13
9
10
11 /*  Lab 2
12
13 */
14
15 use "Coronary Calcium.dta", clear
16
17 ***** Analysis
18 tab dm
19 tab smoke
20 tab chol
21 *****
22
23 log close
24
25 save "Coronary Calcium_NEW.dta", replace
26
```

# Open a .do file

---

Three different ways:

1. Do file button



2. CTRL/CMND + 9

# .do files: best practices

??

```
table phototherapy, ///
  c(min birth_wt p5 birth_wt p50 birth_wt p95 birth_wt max birth_wt) ///
  format(%4.2f)
twoway ///
  (kdensity birth_wt if phototherapy==1, bw(0.2) lpattern(solid)) ///
  (kdensity birth_wt if phototherapy==0, bw(0.2) lpattern(longdash)), ///
  ytitle("Density") xtitle("Birth weight (kg)") ///
  legend(order(1 "Phototherapy" 2 "Untreated")) ///
  name(bw_overlap, replace)
foreach x in bwcat male qual_TSB gest_age age_days {
  tab `x' phototherapy, col nokey
}
logistic over_thresh i.phototherapy ib4.qual_TSB ib40.gest_age ///
  ib1.age_days male birth_wt, vce(cluster hospital) cformat(%5.2f) baselevel
scalar b0 = _b[1.phototherapy]
estat gof, group(10)
logistic over_thresh i.phototherapy ib4.qual_TSB ib40.gest_age ///
  ib1.age_days male bw_csp*, vce(cluster hospital)
testparm bw_csp*
testparm bw_csp2-bw_csp4
label list bwcat // reminder of what the categories are
logistic over_thresh i.phototherapy ib4.qual_TSB ib40.gest_age ib1.age_days male ///
  ib3.bwcat, vce(cluster hospital)
testparm i.bwcat
```

\*/Insider tip: use

\*\*\*\*\*

To separate sections/\*

## Labels help organize your analysis

```
3 set more off
4 clear all
5
6 /* Modify the cd statement to give the path to where you have put the phototherapy
7 data and this do-file. Note: "~/" is a MAC shortcut for "/Users/username/". */
8
9 cd "~/work/mb/atcr/215/labs"
10
11 /* Prefacing a command with "capture" or "cap" prevents Stata from stopping
12 when it detects an error, as it would in the next line if no log were open. */
13
14 capture log close
15
16 /* If you prefer the log as a text rather than a formatted smcl file, which only
17 looks right in a Stata viewer, remove the two forward slashes before the
18 option "text" in the next line. (Stata ignores anything after the two slashes.) */
19
20 log using "lab1", replace // text
21
22 /* You may need to modify the next line to read in the data. If you put the
23 dataset in your current working directory, the next line should work. */
24
25 cap use "phototherapy.dta", clear
26
27 /* Otherwise, you may need to specify the whole path, as this line does for
28 TN's computer. */
29
30 cap use "/Users/thomasnewman/Box Documents/phototherapy.dta"
31
32 ***** PART 1: checking for covariate overlap *****
```

# .do files: making comments

---

- ◇ “Comments” are a way to document your analysis – here are the options

\* Anything after “\*” is comment (entire line)

/\* Anything until you reach the reciprocal symbol  
is comment even on multiple lines\*/

//anything after “//” is a comment (same line)

# .do files: running commands

---

◇ Highlight line or block of code you want to run

1. hit [Do] button in the do file editor



2. Keyboard shortcuts

- CMND+shift+d (MAC)

- Control+D (Windows)

# .do files: running entire .do file at once

---

1. hit [Do] button in the do file editor



2. Command line: `do nameofdofile`

1. Must be in the working directory where you are

3. File < do

4. Keyboard shortcut (within the do file)

1. CMND+shift+d (MAC)

2. Control+D (Windows)

# Lecture 1: Outline

---

- Why STATA?
- Orientation to STATA
- Finding help
- Command syntax
- Working directories
- .do files
- **log files**
- Opening data sets
- Exploring data (basics)

# Log files

---

- ◇ A record of everything of all your analyses Time/date
  - Commands + output
    - Including commands run from the .do file and command line
- ◇ Documents analysis on data *as it was at that time* (.do files don't do this)
- ◇ Two file types
  - **.log**: plain text, easy to open share
  - **.smcl**: native Stata format, difficult to make changes (advantage)
- ◇ **Do not edit**

# Log files

---

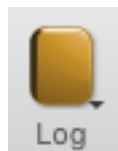
```
log using yourlogname.log  
    , append  
    , replace  
log close  
log off  
log on
```

```
open  
add to end  
replace  
close  
pause  
resume
```

# Open a log file

1. File <log <begin/append

2. Log button (begin/append)

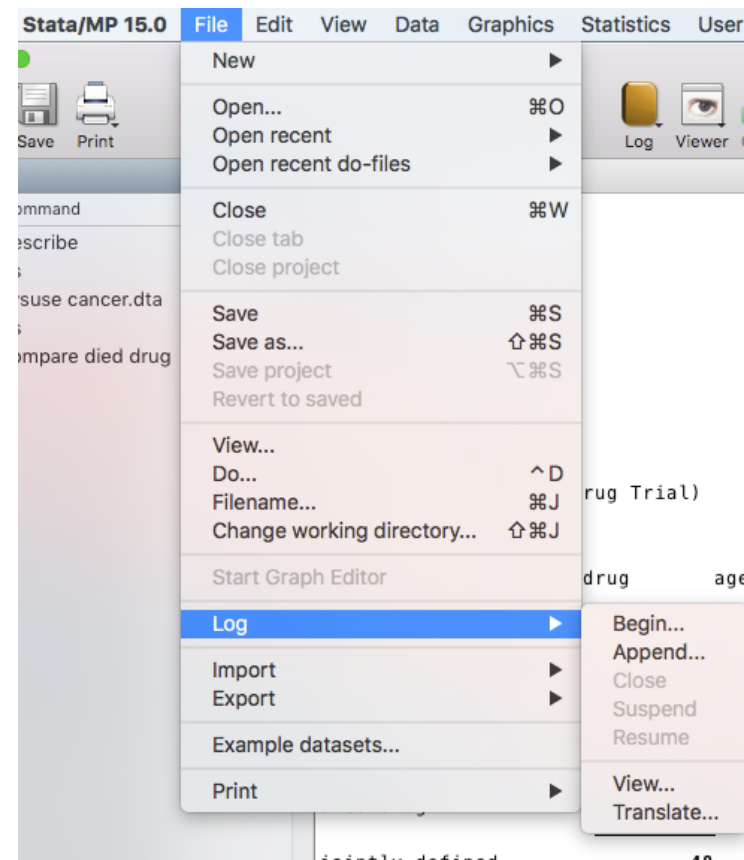


3. Command line

- log using `yourlogname.smcl`, replace

4. From .do file **\*\*recommended**

- log using `yourlogname.smcl`, replace



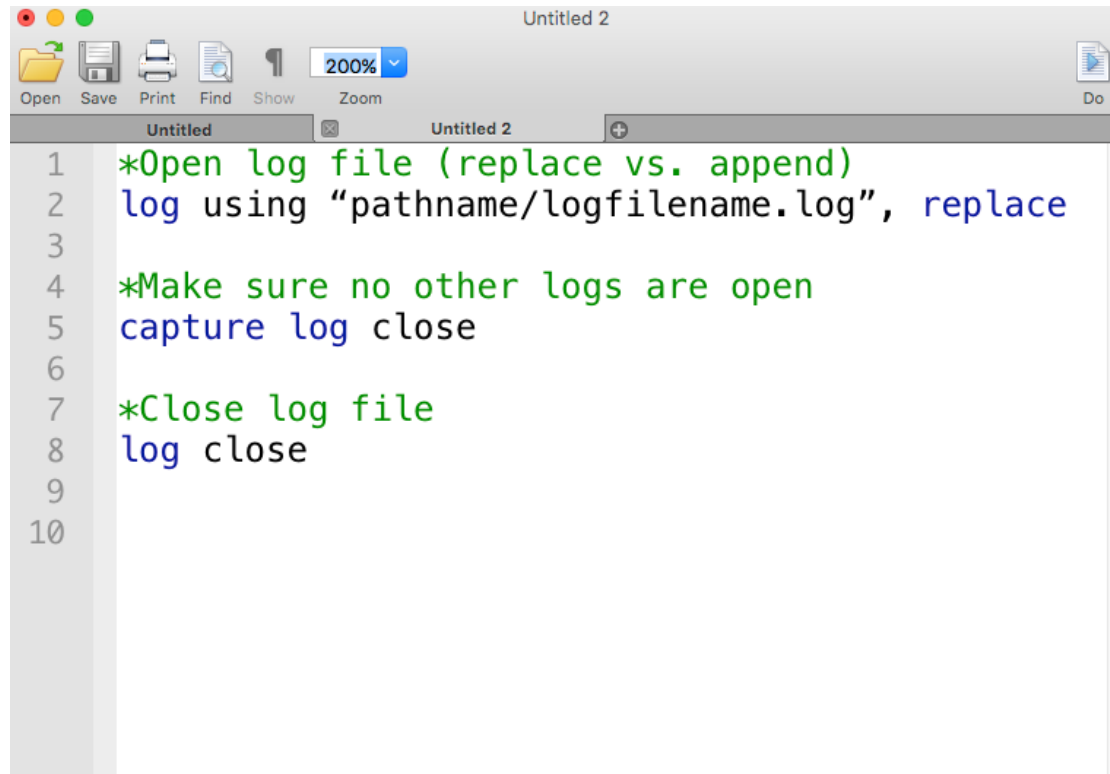
# .do files and log files together

---

- ◇ Open/close the log file within the .do file
  - Everything documented every time
  - Easily repeatable

# .do files and log files together

---



```
1  *Open log file (replace vs. append)
2  log using "pathname/logfilename.log", replace
3
4  *Make sure no other logs are open
5  capture log close
6
7  *Close log file
8  log close
9
10
```

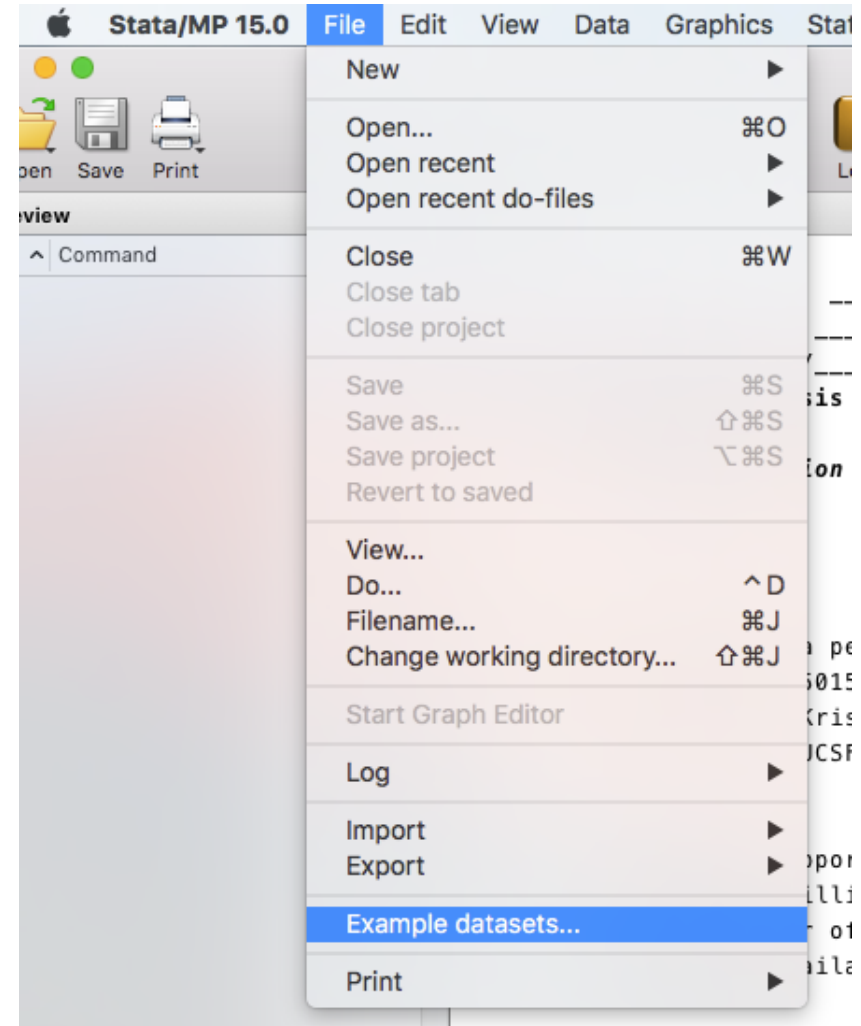
# Lecture 1: Outline

---

- Why STATA?
- Orientation to STATA
- Finding help
- Command syntax
- Working directories
- .do files
- log files
- **Opening data sets**
- Exploring data (basics)

# Opening an example 'system' data set

1. Menu option:
  - File < Example datasets...
2. Command line:
  - `sysuse cancer.dta`



# Lecture 1: Outline

---

- Why STATA?
- Orientation to STATA
- Finding help
- Command syntax
- Working directories
- .do files
- log files
- Opening data sets
- **Exploring data (basics)**

# Explore data: browse

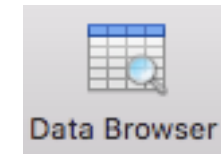
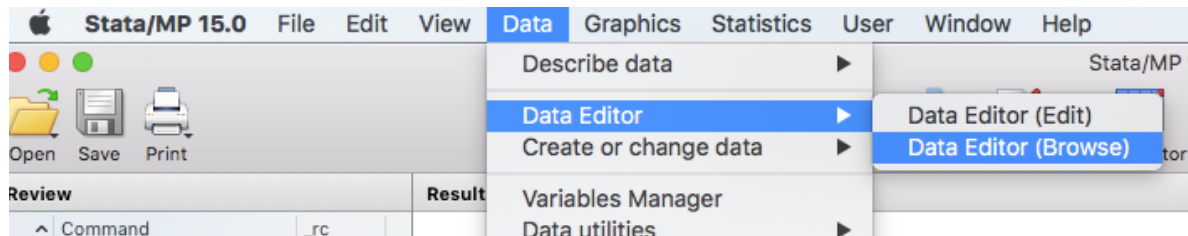
---

browse

- Opens browser window

browse varlist

- Opens browser window and displays specific variables



# Explore data: describe

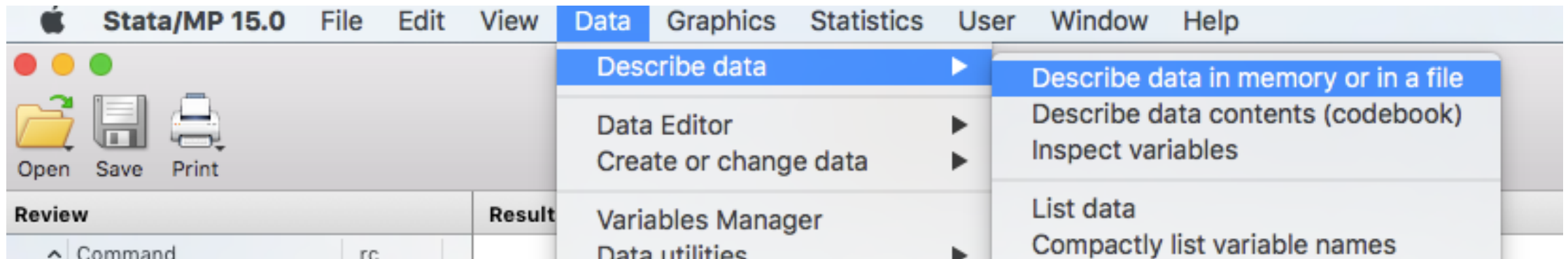
---

`_describe`

- Displays variable names, types, labels for all variables

`_describe varlist`

- Describe 1 or more variables



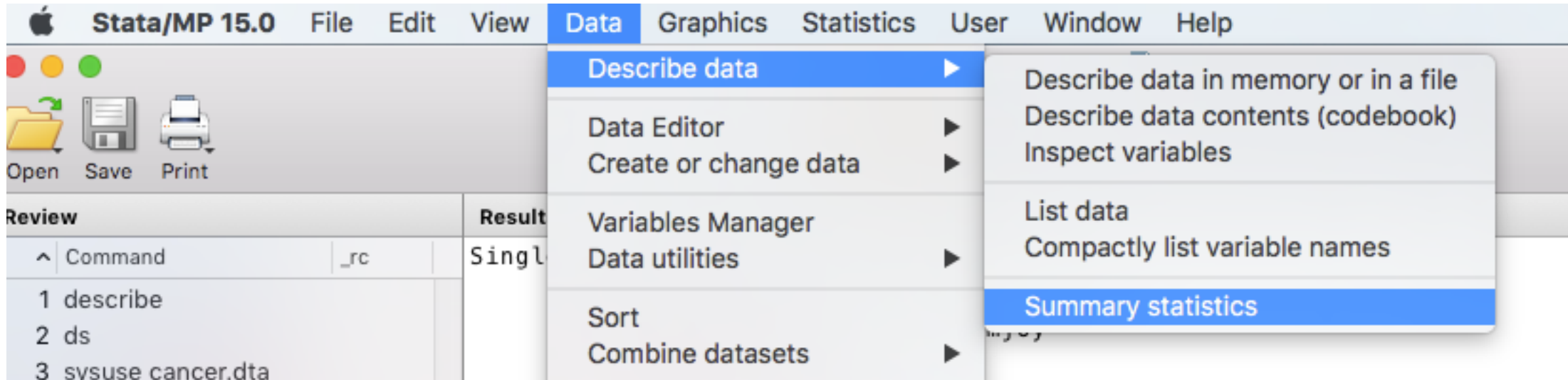
# Explore data: summarize

summarize

- Displays summary statistics (mean, standard deviation, min, max) for all variables

summarize **varlist**

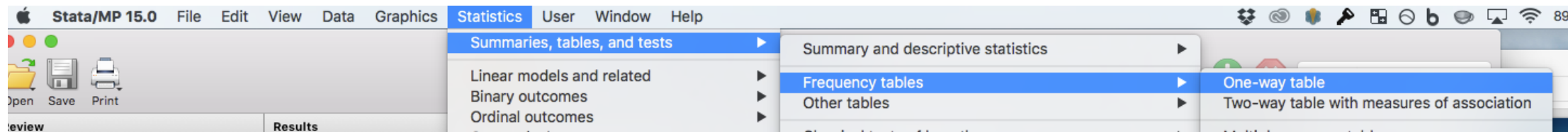
- Summarize 1 or more variables



# Explore data: tabulate

tabulate varname

– One-way table of frequencies



tabulate varname1 varname2

– Two-way table of frequencies

