

# Lecture 1: Introduction to Statistical Computing

---

BIOSTAT 212, Summer 2020



Amanda Irish, DVM MPH PhD(c)  
Amanda.Irish@ucsf.edu  
Department of Epidemiology and Biostatistics  
University of California San Francisco

# Live sessions next week

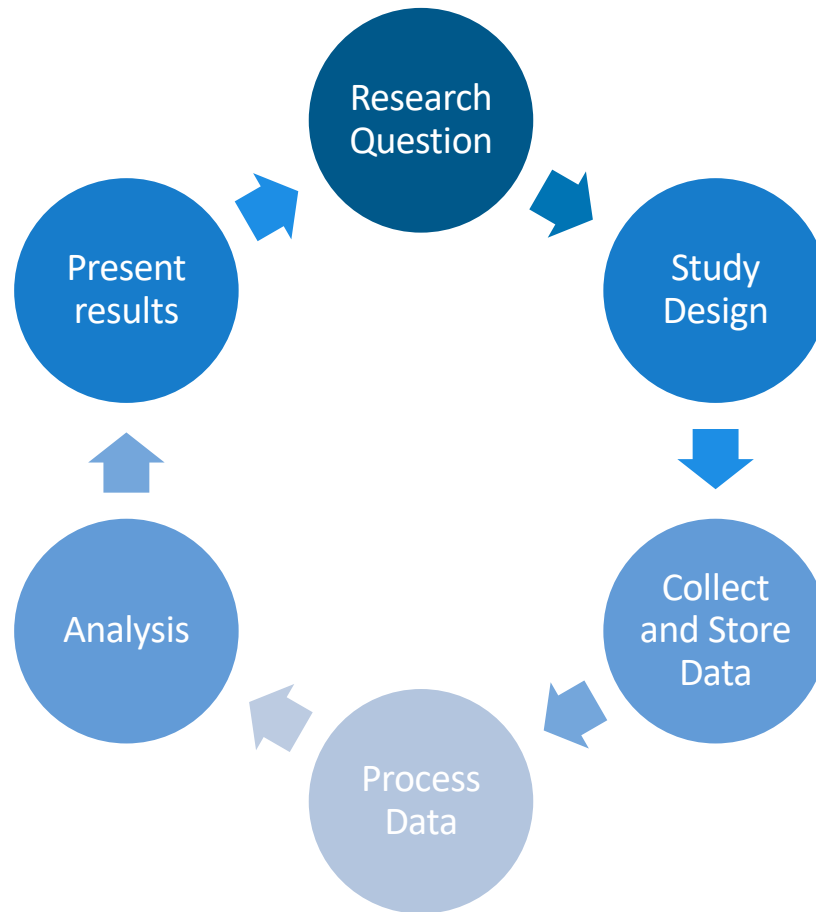
---

## Lecture 1 recap and time for questions (~1hr or shorter)

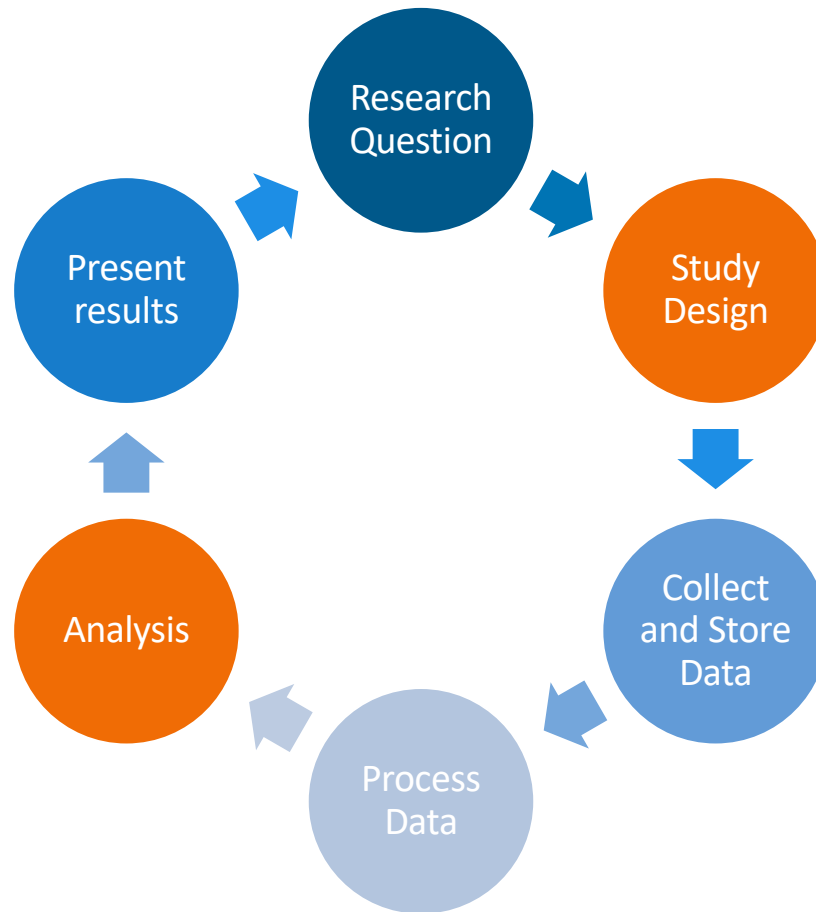
- Anatomy of a clinical research study
- The importance of reproducible research
- Getting started with STATA

## Lab 1 (~1.5 hr)

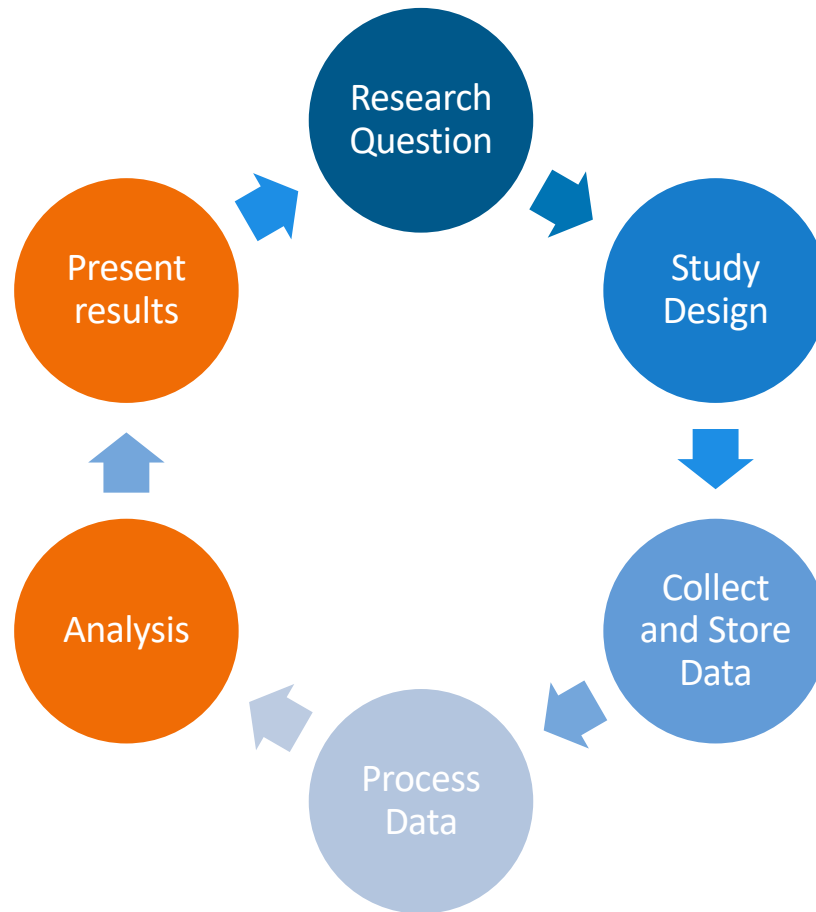
- Install and set up STATA (if you haven't already)
- CLE quiz
- Assignment 1
- Time to ask questions/get help



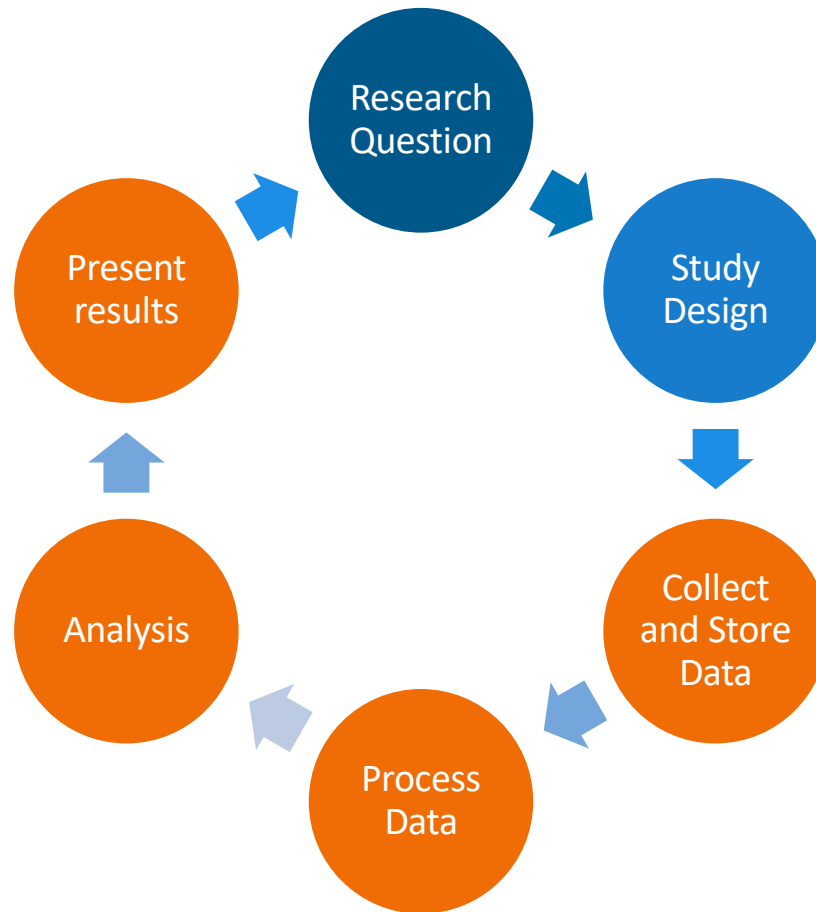
## Life Cycle of a Clinical Research Study



Epidemiology Methods Courses



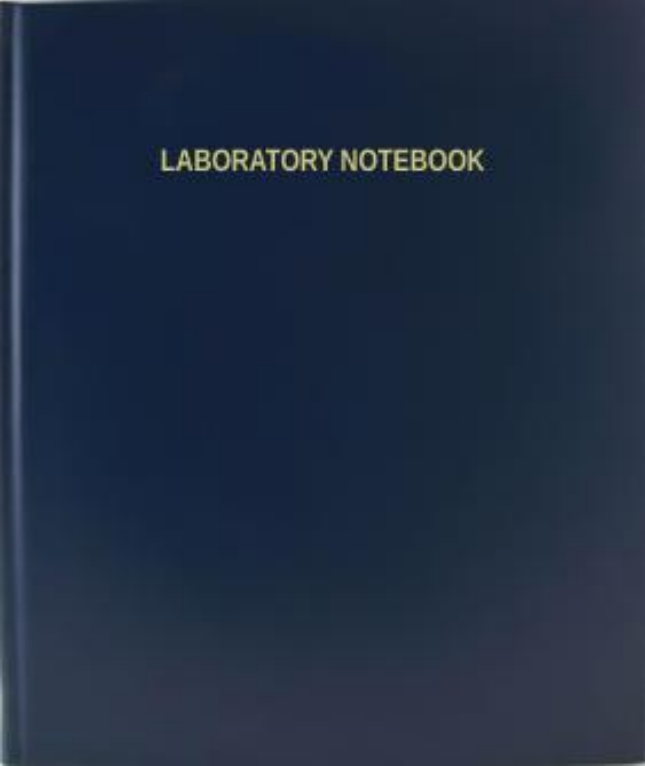
Other Biostats Courses



This Course

A close-up, slightly blurred photograph of a person's hands working on a desk. In the foreground, a hand is pressing a button on a silver and black calculator. Another hand in the background holds a pen over a piece of paper. The desk is cluttered with various papers, some featuring charts and graphs. The overall lighting is warm and the image has a soft, out-of-focus quality.

More time is spent on data  
processing and cleaning than  
conducting analyses



LABORATORY NOTEBOOK

# Reproducible Research

# Reproducible Research

---

The practice of distributing **all data, code, and tools** required to reproduce the results presented in a research publication

# Reproducible Research

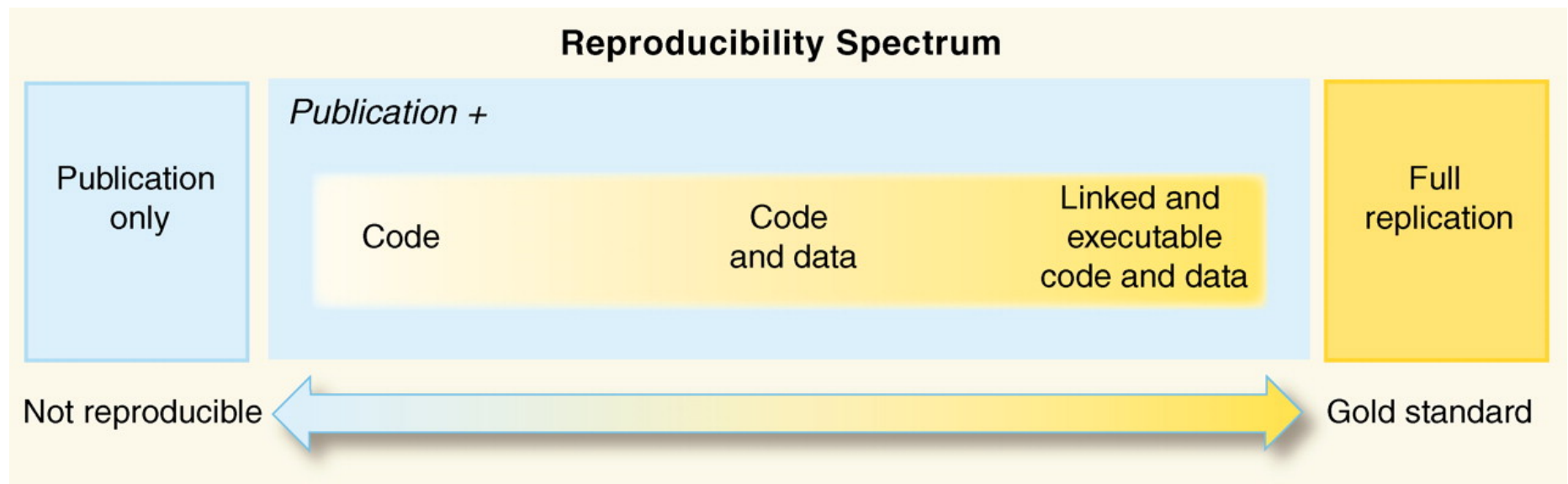
---

Includes **EVERY** step of the scientific process!

**Replication:** Same conclusions from a new study (gold standard)

**Reproducibility:** Same results from the same data and code

Reproducibility vs Replication



From Peng, RD. Reproducible Research in Computational Science, Science 2012

Okay, but how do I do this and  
why are you talking about this  
in an intro to STATA class?

---

# Organization and Documentation

---

- **Ideally** research results should be accompanied by code that can automatically reproduce all of the results given the raw data
- ***At least*** should be able to provide detailed written instructions that allow for reproducing the results from the raw data
- ***This includes all the data cleaning and formatting!***

This makes your life easier too!

## Reproducible code allows you to:

- Automatically adjust analyses if you need to make changes along the way (without having to re-do the whole thing!)
- Easily reply and respond to reviewers
- Re-use code for subsequent analyses
- Doing this by hand takes A LOT of time!



## Why use STATA?

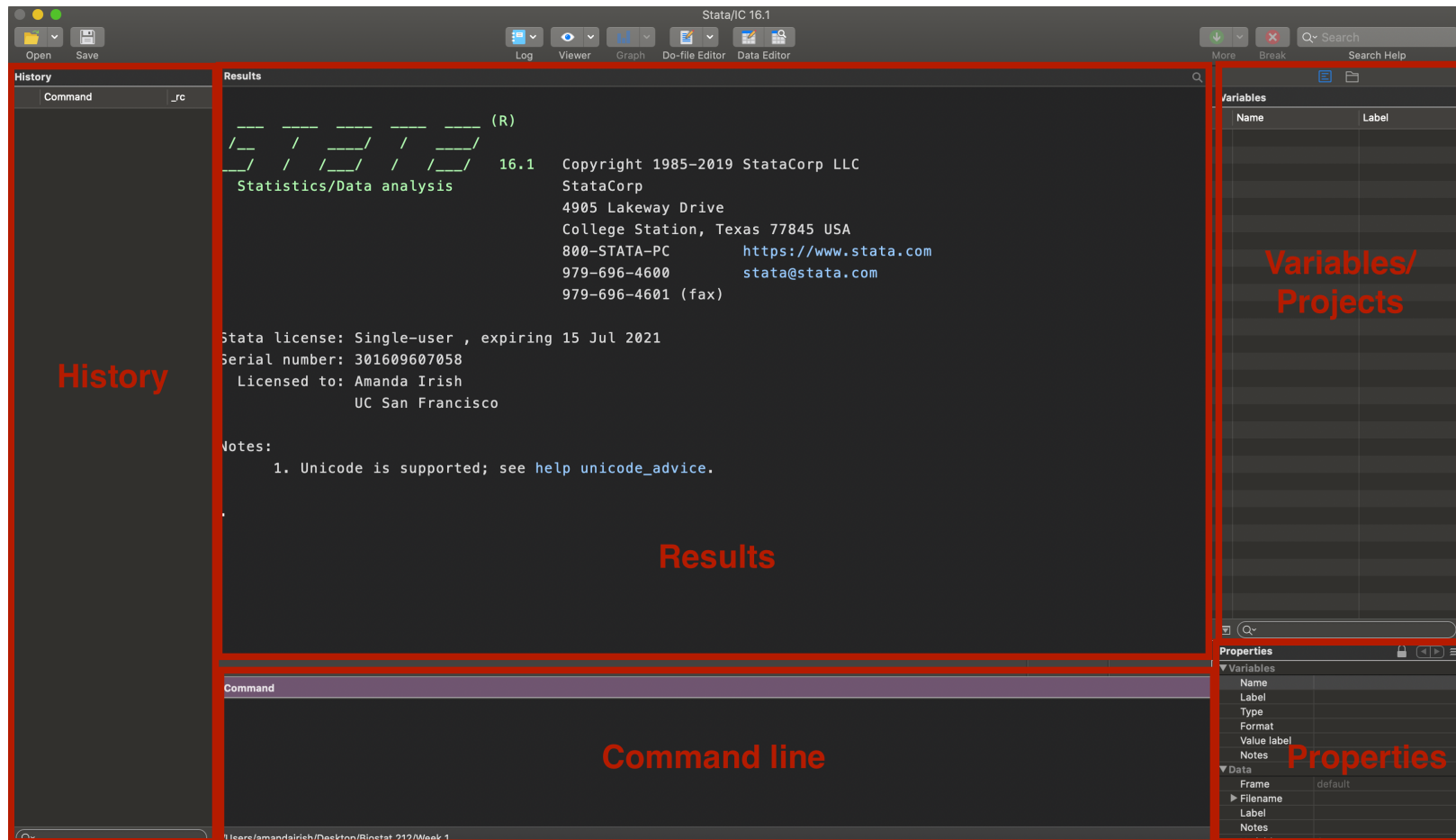
- Quick start, user friendly, but full-featured and powerful
- Command line and menu-driven option
- Good manuals, help menu, help files, customer service
- UCSF/TICR

## STATA vs. other software packages

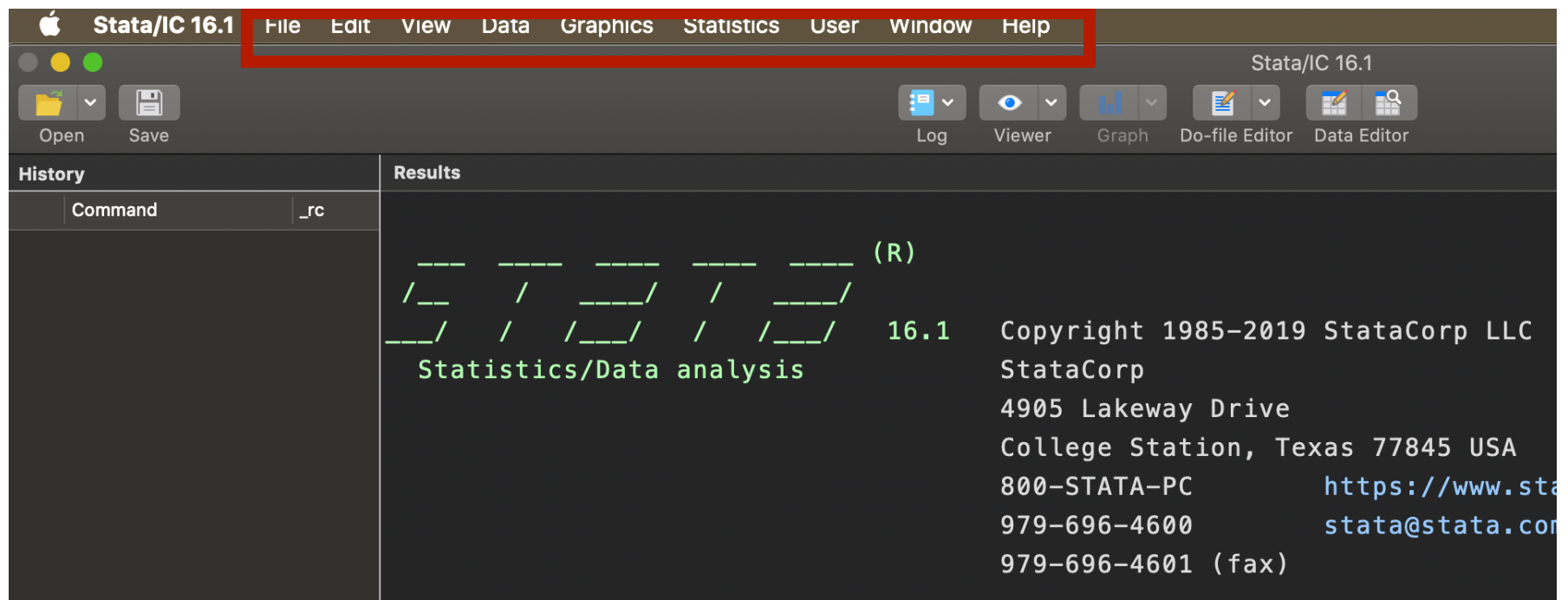
| Features                 | Stata                                                    | SPSS                                                                | SAS                        | R                  |
|--------------------------|----------------------------------------------------------|---------------------------------------------------------------------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

# A tour of STATA...

# Windows

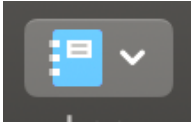


# Menus



# Buttons

---



Log begin/append/etc.

Manages a log file



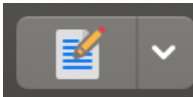
Viewer

Opens a help window



Graph

Brings graph window to the front



Do-file editor

Opens a new or recent do-file



Data editor

Opens data window and allows edits



Data Browser

Opens data window to browse but not edit

# Break

---

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



Or the

Space

To see more



To stop

Note: can also use the commands `set more off`  
`set more on`

# Menu vs. command line

---

- Menu advantages
  - Browse for commands you don't already know
  - 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
    - Reproducible research!
    - Document and repeat your analysis, reuse code, etc.

# Command syntax

---

1. In slides, I will use `courier font` for commands
2. In slides, words that are NOT meant to be typed in but instead are placeholders for variable names or numbers will be `orange`.
3. In slides and STATA help files, parts of the command that are optional will be placed in `[brackets]`.
  - *When you execute the command with the option you SHOULD NOT include the bracket*
4. In slides and STATA help files, shortcuts for commands are indicated by underlining
  - summarize

# Command syntax

---

`tabulate varname1 varname2 [, options]`

`tab male smoke, row`

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

# Help!

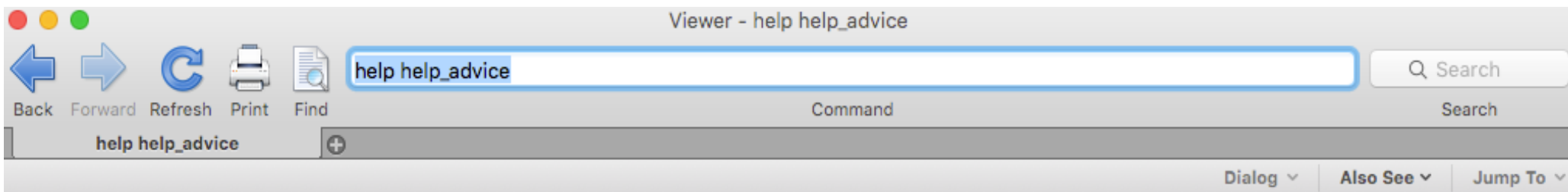
---

- Knowing how to find help is probably the single most important skill for using any programming/stats language!

# Help!– STATA's help files

---

```
.  
  
Command  
help tabulate
```

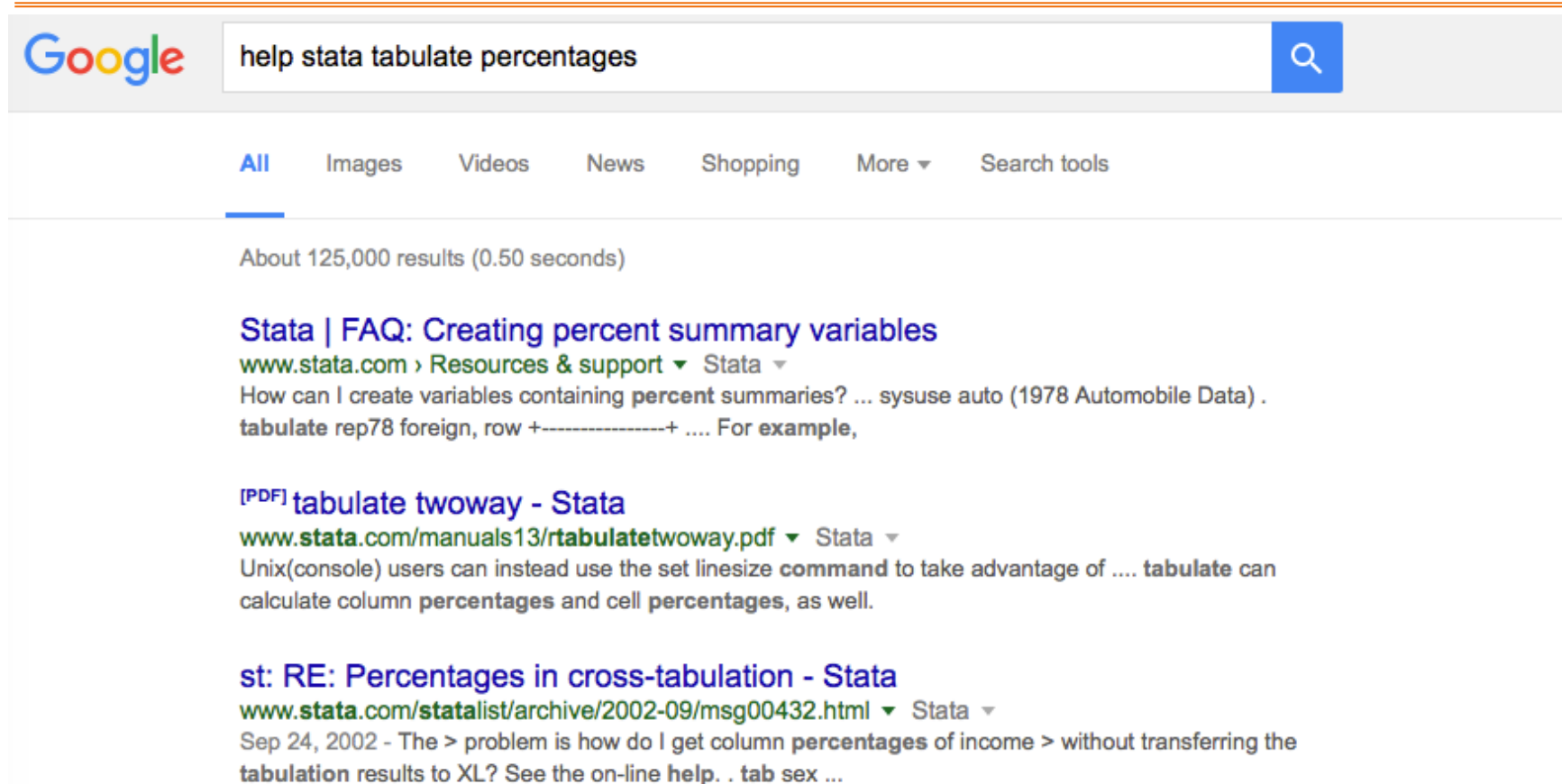


### Advice on finding help

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!– Google is your best friend.



Google

help stata tabulate percentages

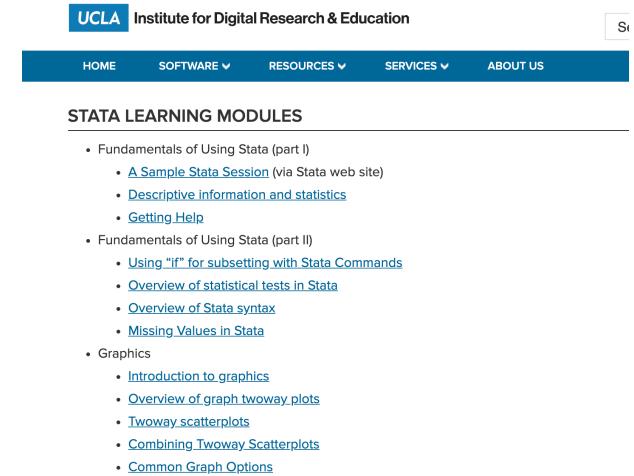
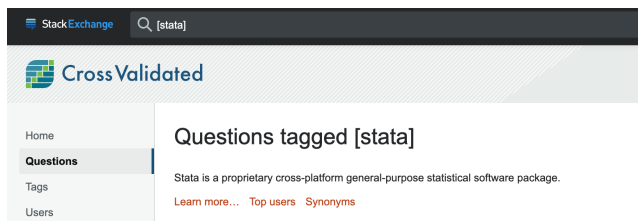
All Images Videos News Shopping More ▾ Search tools

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 ...



# Help!--Other resources

# Working directories

---

- Where all your files will be saved/retrieved from
- Organization of your working directories covered in next lecture
- `pwd`            prints the path of the current working directory
- `cd`            changes the current working directory
  - File > Change working directory from the menu

# Do-files

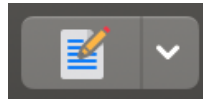
---

- Document all stages of your analysis
    - All the commands you use
  - Comments remind you (or tell someone else) what your code is doing
  - Allows you to repeat analyses quickly
  - Reproducible research!
- 
- More on do-file organization & comments next lecture

# Opening a do-file

---

- Do-file button



- File > New > Do-file
- File > Open recent do-files

# Running commands from a do-file

---

- Highlight the line or block of code you want to run
- Hit the “Do” button in the top right of the do-file editor
  - Can also use command+shift+d (Macs) or control+d (Windows)
- Hitting the “Do” button without highlighting code will run your entire do-file from the start
  - Good way to check for errors



# Log files

---

- A record of your analysis specific to the date/time it was run
  - Commands and output
  - Commands from the do-file and anything you put directly into the command line
- Important complement to do-files because log files document the analysis as it was at that time
  - Do-files are modifiable “living documents” that may change
- Two file types
  - **\*.log**     plain text
  - **\*.smcl**    native Stata format – difficult to change (advantage)

# Commands for log files

---

log using **yourlogname.smcl**

, append

, replace

log close

log off

log on

open a new log file

add to end of existing file

replace existing file

close the log

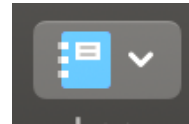
suspends the log

resumes the log

# How to open a log file

---

- File > Log > Begin (or Append)
- Log button (begin or append)
- Command line
- From your do-file **\*\*recommended**



# Immediate commands

---

- Do not require stored data to operate
- Calculator (display)
- Sample size calculator (power)
- Statistical/epidemiological immediate commands
  - End in the letter “i”
  - E.g., `csi` or `cci`

# Immediate commands: calculator

---

- Stata can work as a calculator using the display command

```
display 3+5
```

```
8
```

```
display (54.3-21)*83
```

```
2763.9
```

```
display 23*8
```

```
184
```

```
display 17^3
```

```
4913
```

```
display 3+5
```

```
8
```

```
display sqrt(156)
```

```
12.489996
```

# Mathematical operators in Stata

---

## English

Add

Subtract

Multiply

Divide

To the power of

Natural log of

Base 10 log of

Square Root

## Stata

+

-

\*

/

^

ln ( # )

log10 ( # )

sqrt ( # )

# Stata datasets

---

- Stata comes with some datasets already installed
- Typically small example datasets useful for learning how Stata works

`sysuse data.dta`

loads dataset

`sysuse dir`

lists all the datasets available

`help dta_examples`

more information on the datasets

# Basic dataset exploration

---

browse

describe

summarize

tabulate

codebook

# Browse

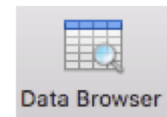
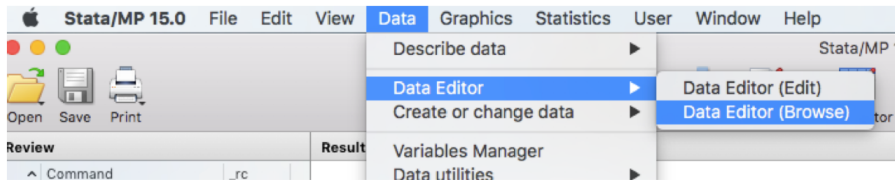
---

browse

- Opens browser window

browse varlist

- Opens browser window and displays specific variables



# Describe

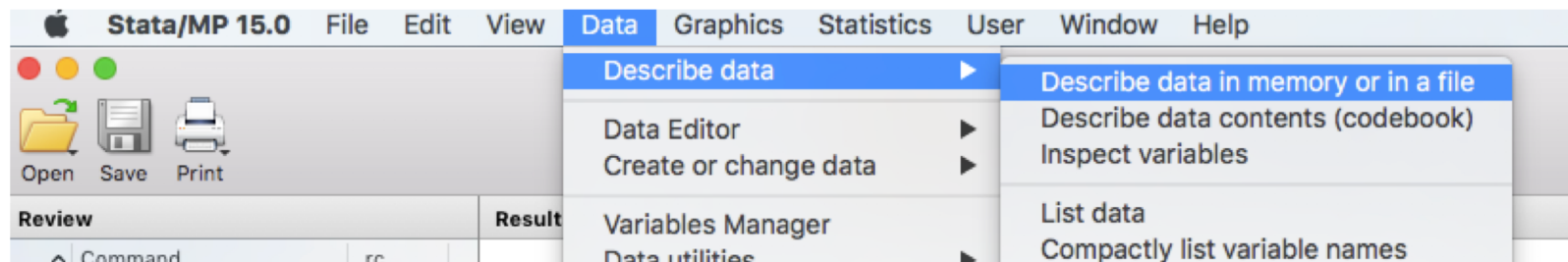
---

`describe`

- Displays variable names, types, labels for all variables

`describe varlist`

- Describe 1 or more variables



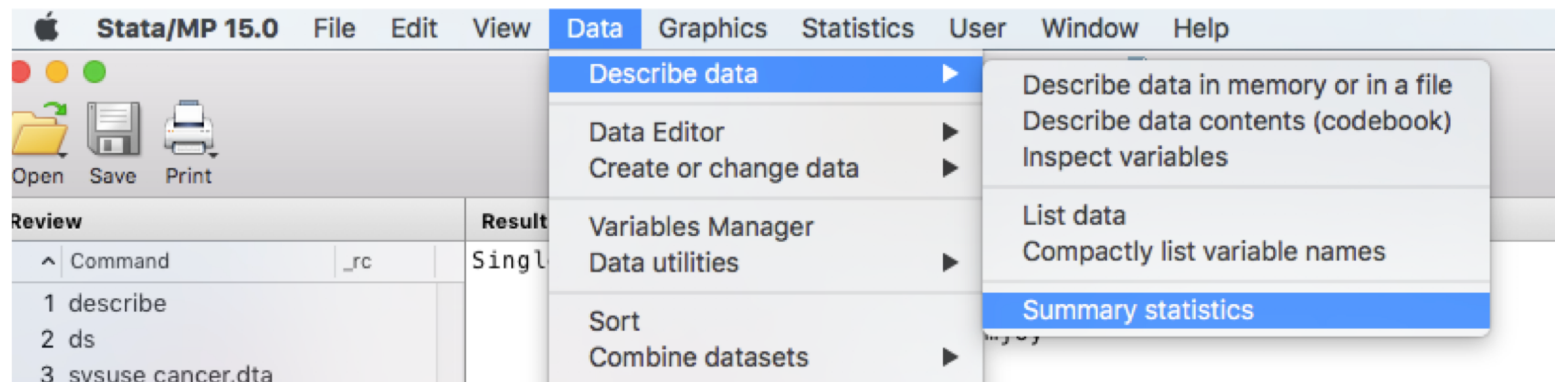
# Summarize

summarize

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

summarize varlist

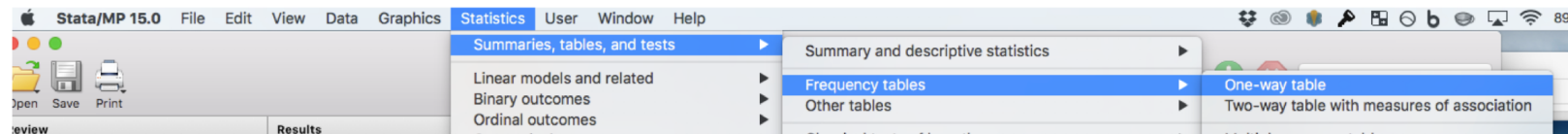
- Summarize 1 or more variables



# Tabulate

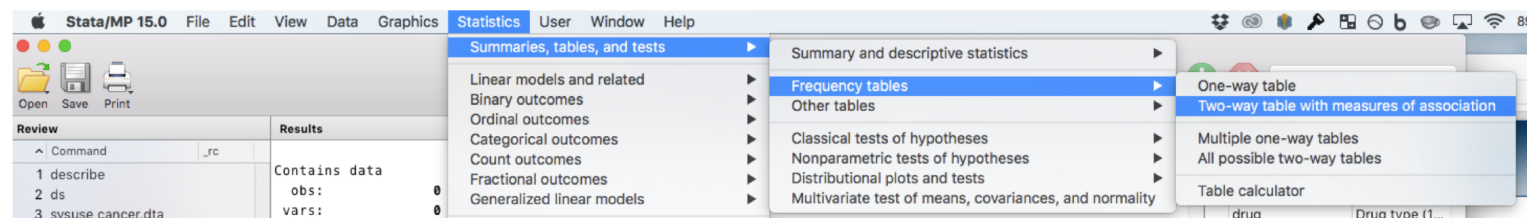
`tabulate varname`

- One-way table of frequencies



`tabulate varname1 varname2`

- Two-way table of frequencies



# Codebook

---

`codebook` – Overview of variable types, stats, number of missing/unique values for all variables

`codebook varlist` – Codebook for 1 or more variables

## options

`, problems` – Report potential problems with dataset

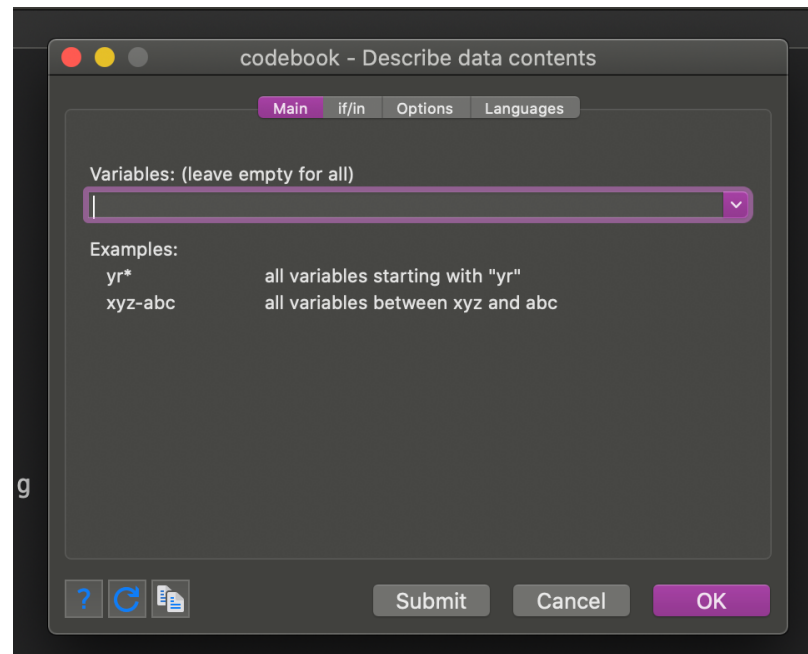
**example** `codebook, problems`

`, compact` – Display compact report for all variables in dataset

# Codebook - menu

---

- Data > Describe data > Describe data contents (codebook)



# Lab 1: CLE quiz and Homework 1

# Lecture 1: Introduction to Statistical Computing

---

BIOSTAT 212, Summer 2020



Amanda Irish, DVM MPH PhD(c)  
Amanda.Irish@ucsf.edu  
Department of Epidemiology and Biostatistics  
University of California San Francisco