

Lecture 2: Exploring data

BIOSTAT 212, Summer 2020



Amanda Irish | Amanda.Irish@ucsf.edu
Department of Epidemiology and Biostatistics
University of California San Francisco

Announcements

- Problems uploading assignments to CLE – should be fixed starting with the "midterm" and onward
 - In the meantime please email me at amanda.irish@ucsf.edu if you need your assignment opened up to re-submit
- Materials for next week available by end of day Wednesday
 - This will be the default going forward unless otherwise noted

Week 2

Stata Basics

- Keyboard shortcuts
- File types
- Variable types
- Logic operators

Project Organization

- Directories and folders
- Do-file documentation

Basic Descriptive Commands

- Adding on to commands from last week with logic operators
- New commands

Stata Basics

Review: STATA syntax

1. I will use `courier font` for presenting Stata commands, and not elsewhere, so they will stand out.
2. The words that you actually type into Stata will be in `black`
3. Words that are NOT meant to be typed in but instead are descriptive placeholders will be `orange`.
4. Parts of the command that are optional will be placed in `[brackets]`. (do not include these brackets in the command)
5. Shortcuts for commands are indicated by underlining
 - `summarize`

Review: STATA Syntax

tabulate var1 var2, row

tab male smoke, row

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

Useful keyboard shortcuts



Describe data



+



Open a .do file



+



Execute command from
.do file (highlight text)



Scroll through previous commands



Autocomplete variable name after first few letters

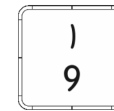
Mac



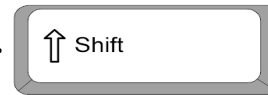
+



+



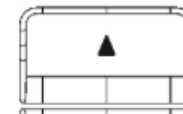
+



+



+



Types of Files

.dta	data files
.do	command files
.ado	programs
.hlp	help files
.gph	graphs
.smcl	Log files (commands and output) in native Stata format
.log	Log files (commands and output) in .txt format

.DTA files: Where your data are - rows and columns

Observations

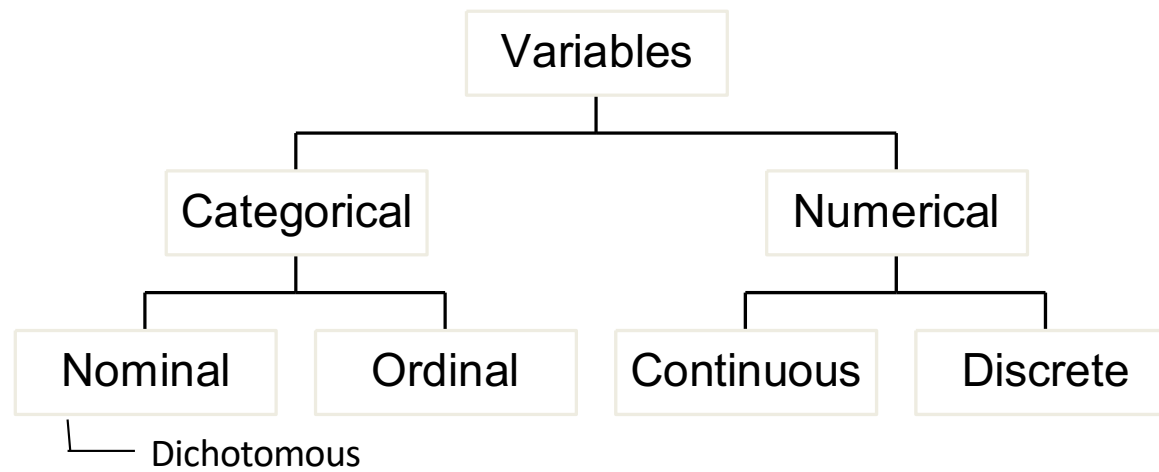


Data Editor (Browse) — Coronary Calcium.dta														
Filter Variables Properties Snapshots														
id[1] 1														
	id	age	htn	dm	smoke	chol	cac	dead	obstime	anycac	cacat	agecat	gender	a
1	1	25	no	no	no	no	0	no	1125	no	0	<30	male	
2	2	25	no	no	no	yes	0	no	1125	no	0	<30	female	
3	3	25	yes	no	yes	no	0	no	1064	no	0	<30	female	
4	4	25	yes	no	yes	no	0	no	1064	no	0	<30	female	
5	5	25	no	no	no	no	0	no	913	no	0	<30	male	
6	6	25	no	no	yes	no	0	no	852	no	0	<30	female	
7	7	25	no	no	no	yes	3	no	699	yes	1-99	<30	male	
8	8	26	no	no	no	yes	0	no	1674	no	0	<30	male	
9	9	26	yes	no	yes	no	0	no	1460	no	0	<30	female	
10	10	26	yes	no	yes	no	0	no	1460	no	0	<30	female	
11	11	26	no	no	no	no	0	no	1368	no	0	<30	male	
12	12	26	no	no	no	no	0	no	1156	no	0	<30	male	
13	13	26	yes	no	yes	yes	0	no	699	no	0	<30	female	
14	14	27	no	no	no	yes	0	no	1613	no	0	<30	male	
15	15	27	no	no	no	yes	0	no	1582	no	0	<30	male	
16	16	27	yes	no	no	yes	0	no	1490	no	0	<30	female	
17	17	27	no	no	no	no	0	no	1460	no	0	<30	female	
18	18	27	no	no	yes	no	0	no	1429	no	0	<30	male	
19	19	27	yes	no	no	no	0	no	1368	no	0	<30	female	
20	20	27	yes	no	no	no	0	no	1278	no	0	<30	male	
21	21	27	yes	no	no	no	0	no	1278	no	0	<30	male	
22	22	27	yes	no	no	no	0	no	1248	no	0	<30	female	
23	23	27	no	no	yes	yes	0	no	1095	no	0	<30	female	
24	24	28	no	no	no	no	0	yes	636	no	0	<30	male	
25	25	28	yes	no	no	yes	0	no	1582	no	0	<30	female	

Variables



Types of Variables



Types of Variables

Categorical		
1. Nominal	No order	e.g. Race (1=Black, 2=White, etc.)
→ 1a. Dichotomous	No order, only 2 categories	e.g. Binary coding of sex (0=male, 1=female)
2. Ordinal	Ordered	e.g. Age (1=10-20, 2=12-30, etc.)
Numerical		
1. Discrete	Can be counted	e.g. How many children in the household?
2. Continuous	Can take any value within a given range	e.g. Weight, age

STATA Variable Storage Types

str1	string 1 character long	words
str#	string # characters long	
strL	string up to 2 billion characters long	
byte	dichotomous	true/false
int	integer values between -32,767 and 32,740	numbers
long	integer values between -2,147,483,647 and 2,147,483,620	
float	real numbers (i.e. numbers with decimal places) with 8 digits of accuracy	
double	real numbers (i.e. numbers with decimal places) with 16 digits of accuracy	
missing	missing data	no data

Variable storage types

obstime	anycac	cacat	agecat	male	age_ge60	sex	haircolor_~e
1125	0	0	<30	1	0	male	blond
1125	0	0	<30	0	0	female	black
1064	0	0	<30	0	0	female	black
1064	0	0	<30	0	0	female	black
913	0	0	<30	1	0	male	black
852	0	0	<30	0	0	female	blond
699	1	1-99	<30	1	0	male	black
1674	0	0	<30	1	0	male	blond
1460	0	0	<30	0	0	female	black
1460	0	0	<30	0	0	female	red
1368	0	0	<30	1	0	male	brown
1156	0	0	<30	1	0	male	blond

Blue

Black

Red

Categorical with label

Numeric or categorical with no label

String

Logic operators in Stata

& and

! not

| or

== tests if something is equal

= assigns a value to a variable

!= not equal

< less than

<= less than or equal to

> greater than

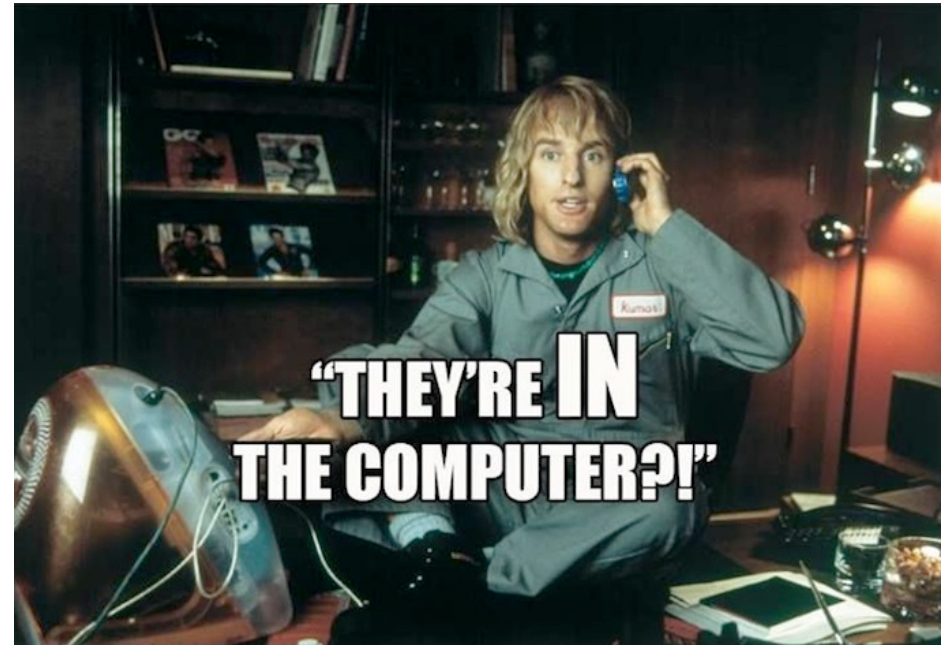
>= greater or equal to

Project Organization

File Paths

Address for where something
lives in your computer

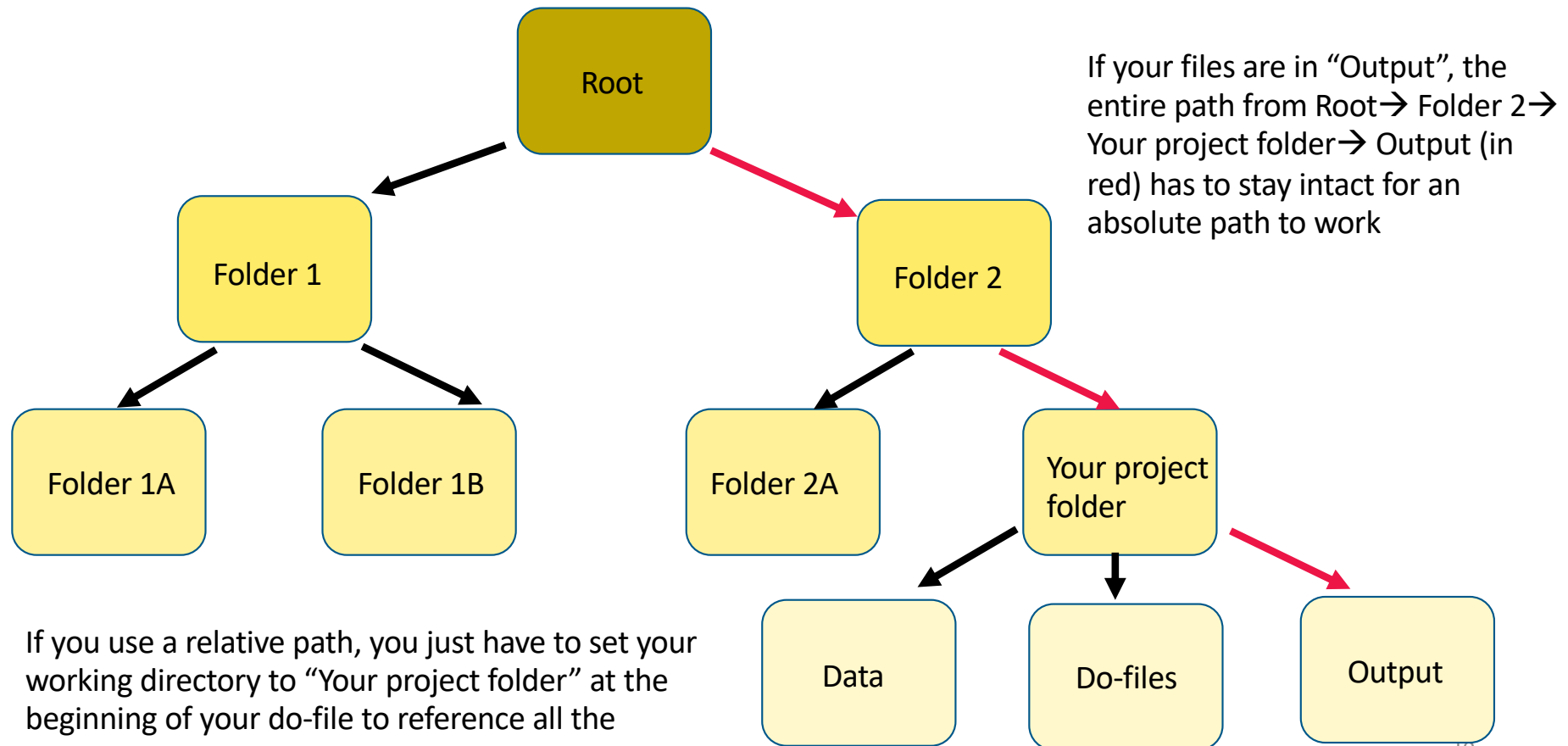
(Walking directions vs GPS
coordinates)



Absolute vs. relative path names

- ◇ Absolute goes all the way back to the root directory for all your computer files (Users/yourusername on a Mac, C:\ on a PC)
 - /Users/amandairish/Desktop/Biostat 212/Week 2
- ◇ Relative means “relative to the parent directory you are in” (set with “cd” at the beginning of your do-file)
 - Just the file name if the file is in the parent directory
 - Subfolder/filename (or subfolder\filename on a PC) for files in subfolders of the parent directory

Relative paths are better for reproducible research



Relative paths are better for reproducible research

- ◇ If you use the full/absolute path for your file, all the folders along that path need to stay in the same place for the command to work
- ◇ Working with others or from another computer
 - Another user (or you on another computer) can easily change the directory to your parent directory at the beginning of your do-file and set up the "downstream" file structure
- ◇ More info: <https://www.kaggle.com/rtatman/reproducibility-tips-absolute-vs-relative-paths>

Working Directories

- Store all files related to a project in one folder
- Organize subfolders for:
 - Data
 - Do files
 - Output

Review from last week

- ◇ To see current working directory:
 - `pwd` displays the path of the current working directory

- ◇ To change working directory:

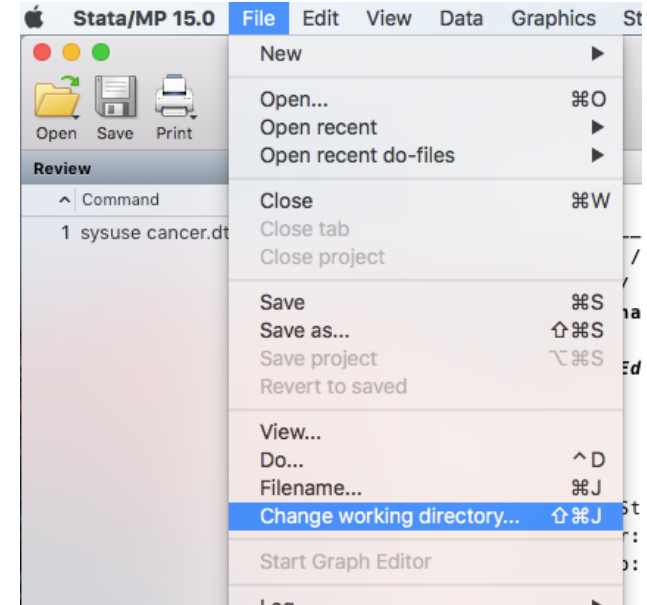
1. Drop-down menu

- File > Change Working Directory

2. Command line

- `cd` change current working directory

- Note that `cd` *alone gets you to home directory*
- Use `dir` to view contents of 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 *

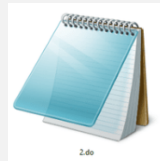
STATA Project Manager

- Groups files related to the same project together
- Makes sharing projects easier
- Allows you to use relative path names

Divide a project
into subfolders



Data



Do-files



Output

Comments

◇ Three ways to make a comment in Stata:

- * at the beginning of a line
- /* */ delimiters for mutli-line comments
 - Also works at the end of a line
- // at the end of a line

◇ Examples

```
1  * This is a comment
2  *or you don't have to have a space
3
4  /* These work well for a longer comment. Remember to
5  put the closing delimiter or Stata will think everything
6  that follows the opening delimiter is a comment */
7
8
9  tab varname // this is a comment for this line of code
10
11 tab varname * this doesn't work - note the color of the text
12
```

Setting up your do-file

- ◇ Best practice to include a “commented-out” header to your do-file with the following information:
 - The who-what-when-why: who created it, what it includes, when it was created and when it was last updated, and its purpose (not necessarily in that order)
 - Consider what you would want to know if you were opening a do-file for the first time for a project you were totally unfamiliar with
- ◇ Also include the Stata version number you used to write the code
 - NOT commented out

```

1  * =====*
2  *
3  * Last updated: 1/2/19 by Amanda Irish
4  *
5  * Analysis for paid family leave (PFL) difference-in-differences
6  * study using NHIS data from 1997-2016
7  *
8  * Code written by Amanda Irish, edited by Rita Hamad
9  *
10 * =====*
11 *
12 * SSC/user-generated programs used:
13 *   - mdesc (~lines 344, 478-479, 585-588)
14 *   - mvpatterns (~line 347)
15 *   - zscore06 (~line 120)
16 *   - zanthro (~line 127)
17 *   - table1 (~lines 598-625)
18 *
19 * Do-file directory
20 *   - Merge datasets to associate parent variables with their child (~line 45)
21 *   - Generate/modify other exposure/outcome variables (~line 82)
22 *   - Regression models (~line 337)
23 *   - Table 1 data (~line 592)
24 *   - Assumption checking (~line 631)
25 *   - Variable availability by year (~line 377)
26 *   - # observations/year (~line 641)
27 *
28 * =====*
29
30 **** NOTE: THIS CODE MUST BE RUN AFTER THE "nhis_1997_2017labels_26Mar2018"
31 **** DO-FILE BY AKANSHA & ANUSHA
32
33 clear matrix
34 set more off, perm
35
36 version 14.2
37 cd "/projects/data/Paid Family Leave"

```

```

1 clear matrix
2 version 14.2
3 set more off
4 set linesize 120
5
6
7 ****
8 ** GRE study **
9 ** PURPOSE: Assess relationship among URM, GRE score, and odds of acceptance **
10 ** DATA USED: Graduate Division application data 2015-2018 **
11 ** STARTING DATE : February 17, 2020 **
12 ** CREATED BY: Amanda Irish **
13 ****
14
15 cd "/Users/amandairish/Desktop/GRE study"
16 clear
17
18

```

Importing and Exporting Data

Opening a data set (.dta) saved on your computer

1. Command line

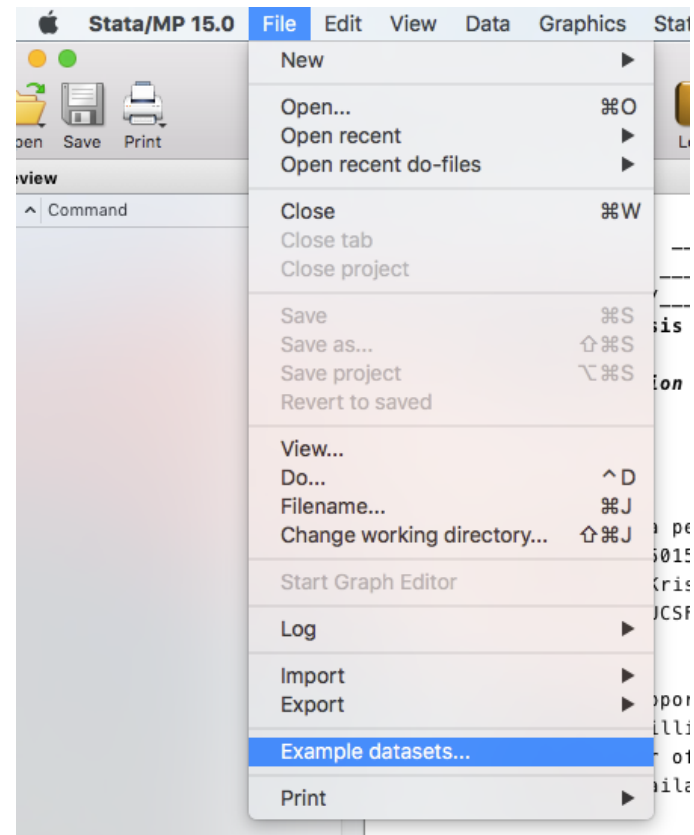
o use `"yourStataDataFile.dta"`, `clear`

- *Note: this will open only from your current working directory*
- Type `dir` to see what files are in your directory

2. File > Open

Review: opening a 'system' data set

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



Opening a web data set

- `webuse nhanes2`
- `webuse www.datasetURL`

◇ `webuse query` //tells you what URL Stata is looking for data files in

A note on example data sets

- ◇ The datasets used to demonstrate Stata are often fictional. If you want to know whether a dataset is real or fictional, and its history, type:
 - `notes // dataset must be installed`
- ◇ A few datasets have no notes. This means that the datasets are believed to be real but that they were created so long ago that information about their original source has been lost.

Importing data: .CSV files

1. Use the command line

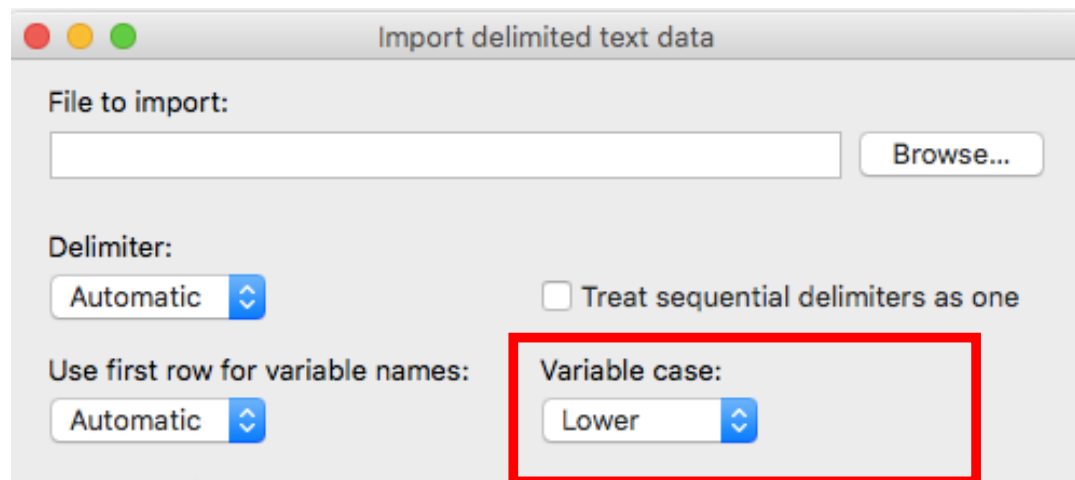
- `import delimited "yourfile.csv"`

2. File > Import > Text Data (delimited, *.csv, ...)

3. Copy paste into data editor – **dangerous!!**

- .CSV files use commas to separate (delimit) data
- Export data from other programs (like excel) as a .csv file
- Note: .CSV files will not preserve variable labels

Importing data: File > import



- ◇ First row = variable names
- ◇ Variable case = lower
- ◇ Use the dropdown menu and then copy command into your .do file

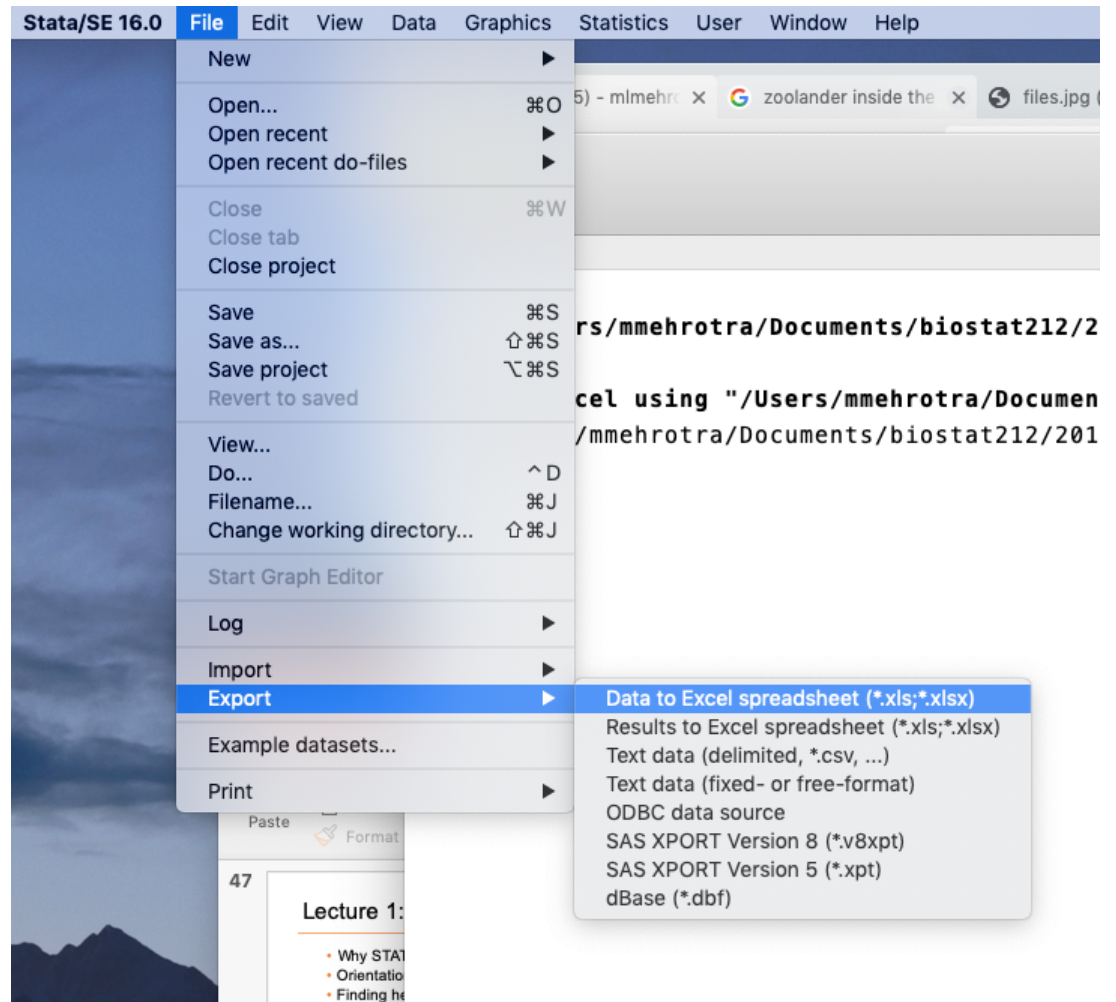
Other import formats

Source program	Export as	Stata command
Excel	.csv	import delimited "filename"
Excel	.xls	import excel "filename"
SPSS	.dta	Use "filename"
SAS	.xpf	import sasexport "filename"
SQL		odbc "help odbc" for more information



- ◇ save "filename.dta", replace save as a **NEW** filename.
- ◇ **NEVER** modify raw data. Always save a new copy with a new file name.

Exporting Data



Exporting Data

```
. export excel using "data/accupuncture_baseline_onlymales.xlsx"  
file data/accupuncture_baseline_onlymales.xlsx saved
```

Clearing data and closing STATA

- ◇ `clear` clears data set, be sure to save first!
- ◇ `cls` clears the console (where results are displayed)
- ◇ `exit, clear` closes STATA and clears data, be sure to save first!

Descriptive commands

Acupuncture trial dataset

- ◇ **Objective** To determine the effects of acupuncture on headache
- ◇ **Design** Randomized trial
- ◇ **Setting** General practices in England and Wales
- ◇ **Participants** patients with chronic headache, predominantly migraine

We will use baseline data for this week

Explore data

- ◇ browse
- ◇ describe
- ◇ count
- ◇ lookfor
- ◇ isid
- ◇ duplicates
- ◇ codebook
- ◇ summarize
- ◇ inspect
- ◇ list
- ◇ gsort
- ◇ tabulate

	Continuous	Categorical	Dichotomous
browse	✓	✓	✓
<u>describe</u>	✓	✓	✓
isid	✓		
duplicates	✓		
codebook	✓	✓	✓
<u>summarize</u>	✓		
<u>inspect</u>	✓	✓	✓
<u>list</u>	✓	✓	✓
gsort	✓	✓	✓
<u>tabulate</u>		✓	✓

Explore data: `browse`

`browse`

- Opens browser window

`browse varlist`

- Opens browser window and displays specific variables

Options

`if` – browse using logic operators

example `browse if sex==1 & age>=36`

Explore data: describe

`describe`

- Displays variable names, types, labels for all variables

`describe varlist`

- Describe 1 or more variables

Options

`, short`

- Describe only observations, size no variables

`, fullnames`

- Give full variable names

Prefix: `bysort`

`bysort varname: stata_cmd`

- ◇ Most Stata commands allow the `bysort` prefix, which repeats the command for each group of observations
- ◇ `by` without the `sort` option requires that the data be sorted by `varname`

Explore data: count

count

– count number of observations

Options

`if` – count using logic operators

example `count if sex==1`

`by/bysort varname:` – count by categories of another variable

example `bysort race: count if sex==1`

Prefix: bysort

```
. bysort sex: count if migraine==1
```

```
-> sex = male
```

```
    43
```

```
-> sex = female
```

```
   241
```

```
.
```

Explore data: summarize

summarize

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

summarize `varlist`

- Summarize 1 or more variables

Options

`if`

- Summarize using logic operators

example `sum if alive==1`

`, detail`

- produces additional statistics, including skewness, kurtosis, the four smallest and four largest values, and various percentiles.

`by/bysort` `varname:`

- sum by categories of another variable

Explore data: `inspect`

`inspect`

- Display histogram of data, number of missing and 0 observations, number of integer observations for all variables

`inspect varlist`

- Inspect for 1 or more variables

Options

`if`

- Inspect using logic operators

example `inspect chol if age<50`

`by/bysort varname:`

- inspect by categories of another variable

Explore data: inspect

```
. inspect basehs
```

basehs: baseline headache score

#				Negative
#				Zero
#				Positive
#				
#	#			Total
#	#	.	.	Missing
6.75		94.75		
(More than 99 unique values)				

Number of Observations

Total	Integers	Nonintegers
-	-	-
-	-	-
301	80	221
301	80	221
-		
301		

Explore data: gsort

gsort +varlist

- arranges observations to be in ascending order of the specified variables

gsort -varlist

- arranges observations to be in descending order of the specified variables

options

, mfirst

- puts missing values first

Explore data: list

- `list` – list values of all variables
- `list varlist` – list values of 1 or more variables

options

- `if` – list using logic operators
example `list sex chol if age<50`
- `in` – list a certain number of observations
example `list sex chol in 1/10`
- `, clean` – clean option gives a table format
- `, sum/mean/N` – a line added to the bottom of the list with the sum, mean and N
example `list sex chol in 1/10, mean N sum clean`
- `by/bysort varname:` – list by categories of another variable

Explore data: tabulate

tabulate `varname`

- One-way table of frequencies

tabulate `varname1 varname1`

- Two-way table of frequencies

Options

`if`

- Use logic operators

example `tab chol smoke if age<50`

`, missing`

- include missing data

`, row`

- gives row percentages

`, col`

- gives column percentages

example `tab chol smoke, col`

`by/bysort varname:`

- tab by categories of another variable

	Continuous	Categorical	Dichotomous
browse	✓	✓	✓
<u>describe</u>	✓	✓	✓
isid	✓		
duplicates	✓		
codebook	✓	✓	✓
<u>summarize</u>	✓		
<u>inspect</u>	✓	✓	✓
<u>list</u>	✓	✓	✓
gsort	✓	✓	✓
<u>tabulate</u>		✓	✓

Lecture 2: Exploring data

BIOSTAT 212, Summer 2020



Amanda Irish | Amanda.Irish@ucsf.edu
Department of Epidemiology and Biostatistics
University of California San Francisco