

Lecture 2: Exploring data

BIOSTAT 212, Summer 2017



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

Questions from last week?

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

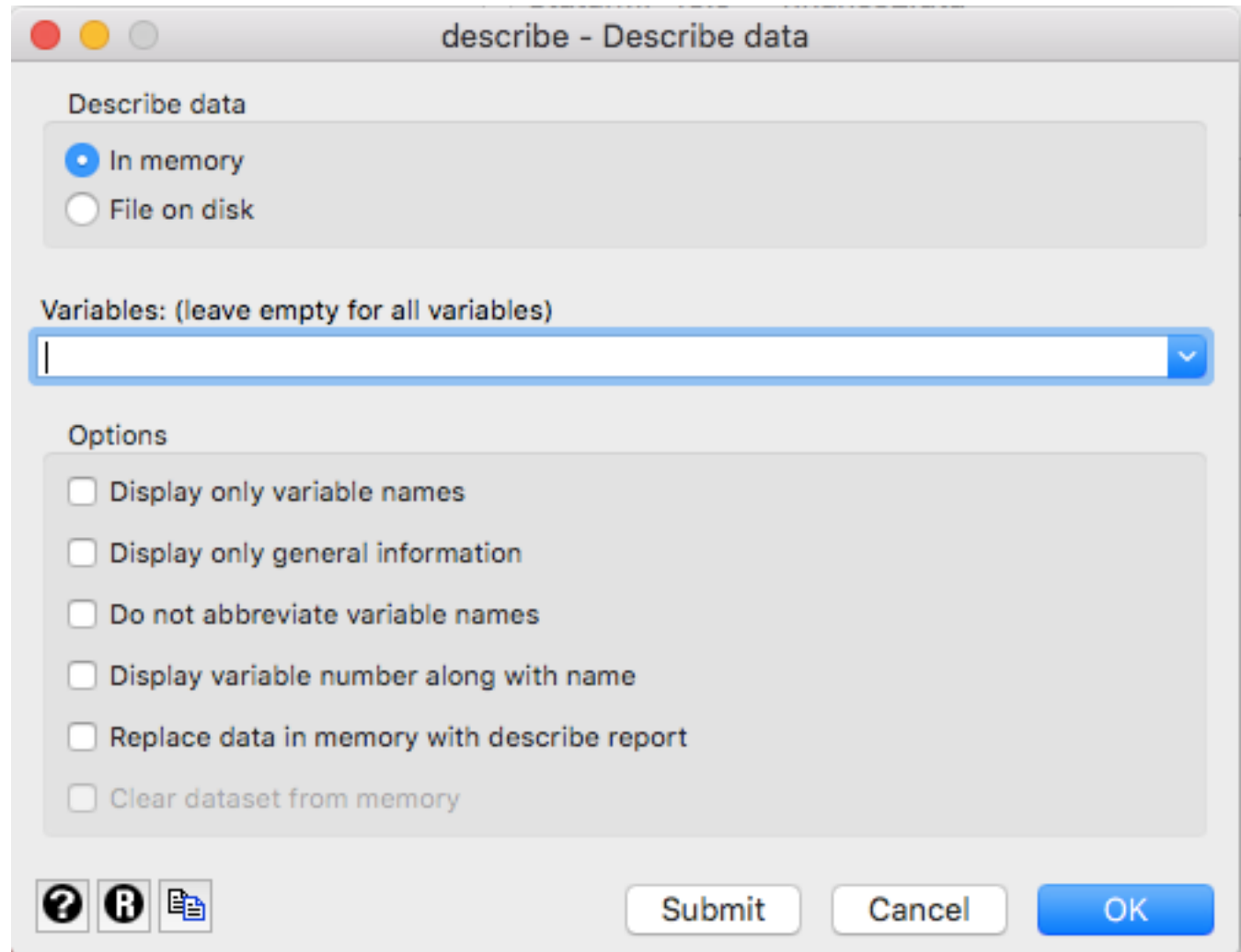
male	smoke		Total
	0	1	
0	2,723 65.00	1,466 35.00	4,189 100.00
1	3,539 57.24	2,644 42.76	6,183 100.00
Total	6,262 60.37	4,110 39.63	10,372 100.00

Review: Why you should avoid “point and click” analyses

- ◇ Easy to forget
- ◇ Not reproducible
- ◇ Slow
- ◇ Mistakes cannot be corrected
- ◇ Cannot be supervised or checked by mentors, reviewers
- ◇ The procedure cannot be shared
- ◇ Cannot be reused

The dialog box

▢ db describe



Notes from last week

- ◇ Download .do files from the course website and open with Stata
- ◇ Labs are for your own practice, you do not need to turn them in

Outline for today

- ◇ **Keyboard shortcuts**
- ◇ Display preferences
- ◇ File types
- ◇ Variable types
- ◇ Immediate commands
- ◇ Logic operators
- ◇ Opening and importing data
- ◇ More data exploring

Useful keyboard shortcuts



Describe data



+



Open a .do file



+



Execute command from
.do file (highlight text)



scroll through previous commands

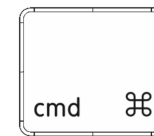


Autocomplete variable name after first few letters

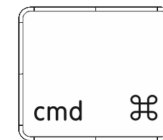
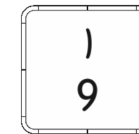
Mac



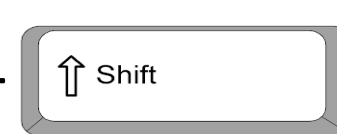
+



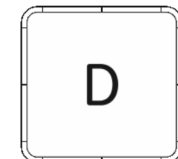
+



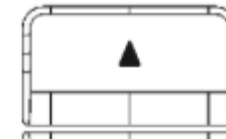
+



+



+



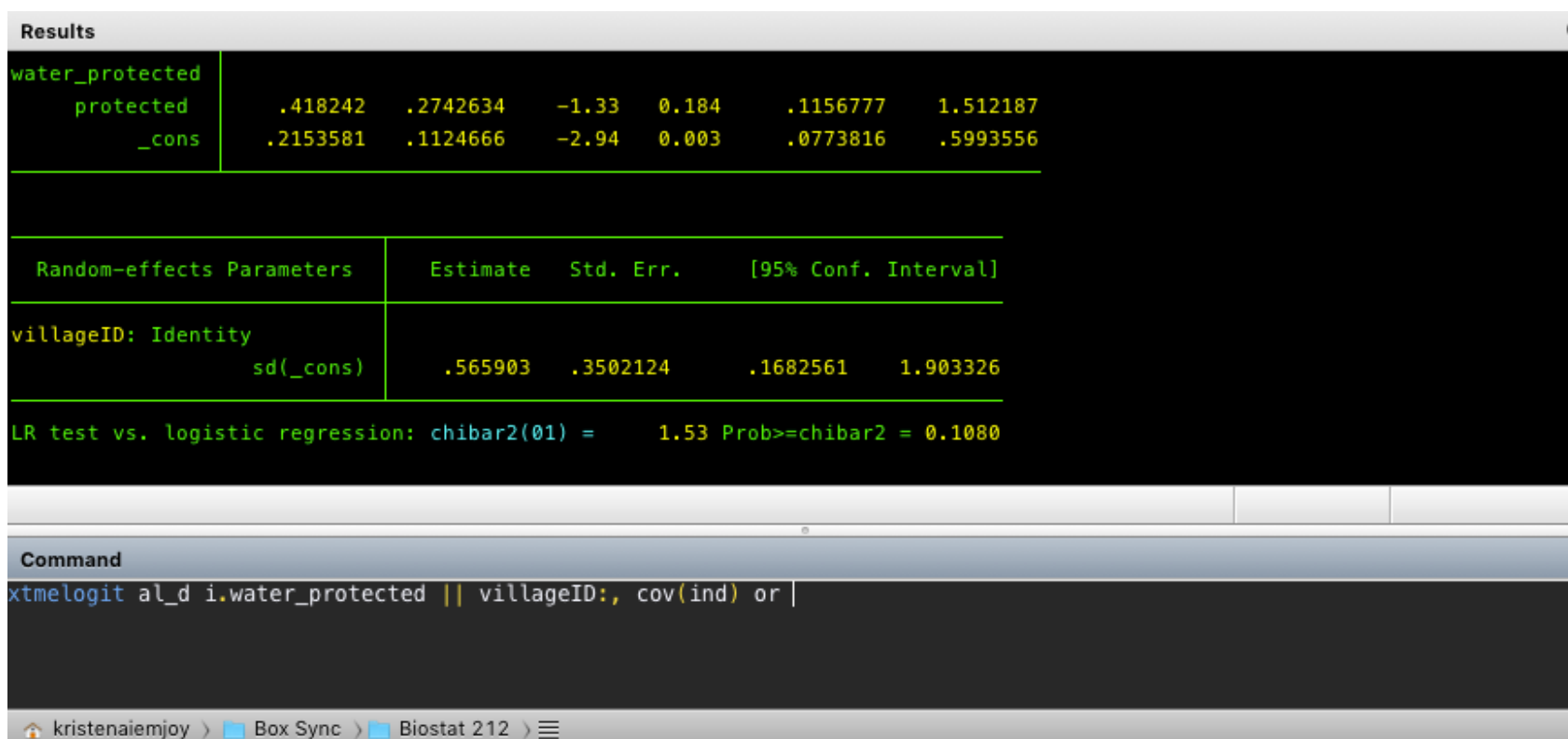
Outline for today

- ◇ Keyboard shortcuts
- ◇ **Display preferences**
- ◇ File types
- ◇ Variable types
- ◇ Immediate commands
- ◇ Logic operators
- ◇ Opening and importing data
- ◇ More data exploring

Display preferences

◇ Drop down menu

- STATA < General preferences < Syntax Highlighting < Color scheme
- STATA < General preferences < Windows < Results < Color scheme



The screenshot shows the STATA Results window with a dark background and yellow text. The output is for a mixed-effects logistic regression model. The first table shows the coefficients for the fixed effects. The second table shows the random-effects parameters for the villageID variable. The LR test result is also displayed.

water_protected	protected	_cons	Estimate	Std. Err.	[95% Conf. Interval]
.418242	.2742634	-1.33	0.184	.1156777	1.512187
.2153581	.1124666	-2.94	0.003	.0773816	.5993556

Random-effects Parameters	Estimate	Std. Err.	[95% Conf. Interval]
villageID: Identity			
sd(_cons)	.565903	.3502124	.1682561 1.903326

LR test vs. logistic regression: $\chi^2(01) = 1.53$ Prob>= $\chi^2 = 0.1080$

Command: `xtmelogit al_d i.water_protected || villageID:, cov(ind) or |`

Navigation bar: kristenaiemjoy > Box Sync > Biostat 212 > ≡

Outline for today

- ◇ Keyboard shortcuts
- ◇ Display preferences
- ◇ **File types**
- ◇ Variable types
- ◇ Immediate commands
- ◇ Logic operators
- ◇ Opening and importing data
- ◇ More data exploring

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 is - rows and columns

Data Editor (Browse) — Coronary Calcium.dta

Filter Variables Properties Snapshots

id[1] 1

	id	age	htn	dm	smoke	chol	cac	dead	obstime	anycac	caccat	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	

Observations



Variables



.DO files: A recipe book of commands you can use again and again

```
activity1
1
2  cd "C:\data\biostat212\"          // create a folder for each assignment/activity
3                                   // and change the directory this folder
4
5  capture log close
6
7  log using "Activity1_yourname.smcl", append    //name the file whatever you choose
8
9  set more off                          // prevents output from getting interrupted for page breaks
10
11 version 14                            // tells stata you are using version 14
12
13
14 /* Activity 1
15
16 */
17
18 use "Coronary Calcium.dta", clear        //make sure this data set is saved in your folder
19
20
21
22 ***** Analysis (this will change for each lab)
23 tab dm
24 tab chol
25
26
27 log close                              //closes log file
28
29 set more on
30
31
32
```

.LOG files: A recording of commands and results every time you look at your data.

```

name: <unnamed>
log: /Users/kristenaiemjoy/Box Sync/Biostat 212/2016/_Assignments/Activity 1/Activity1_yourname.smcl
log type: smcl
opened on: 28 Jul 2016, 10:49:30

.
. set more off                // prevents output from getting interrupted for page breaks
.
.
. version 14                  // tells stata you are using version 14
.
.
. /*      Activity 1
>
> */
.
. use "Coronary Calcium.dta", clear    //make sure this data set is saved in your folder
.
.
.
. ***** Analysis (this will change for each lab)
. tab dm

      diabetes |      Freq.      Percent      Cum.
-----+-----
           no |         9,468         91.30         91.30
           yes |          902          8.70        100.00
      Total   |        10,370        100.00

. tab chol

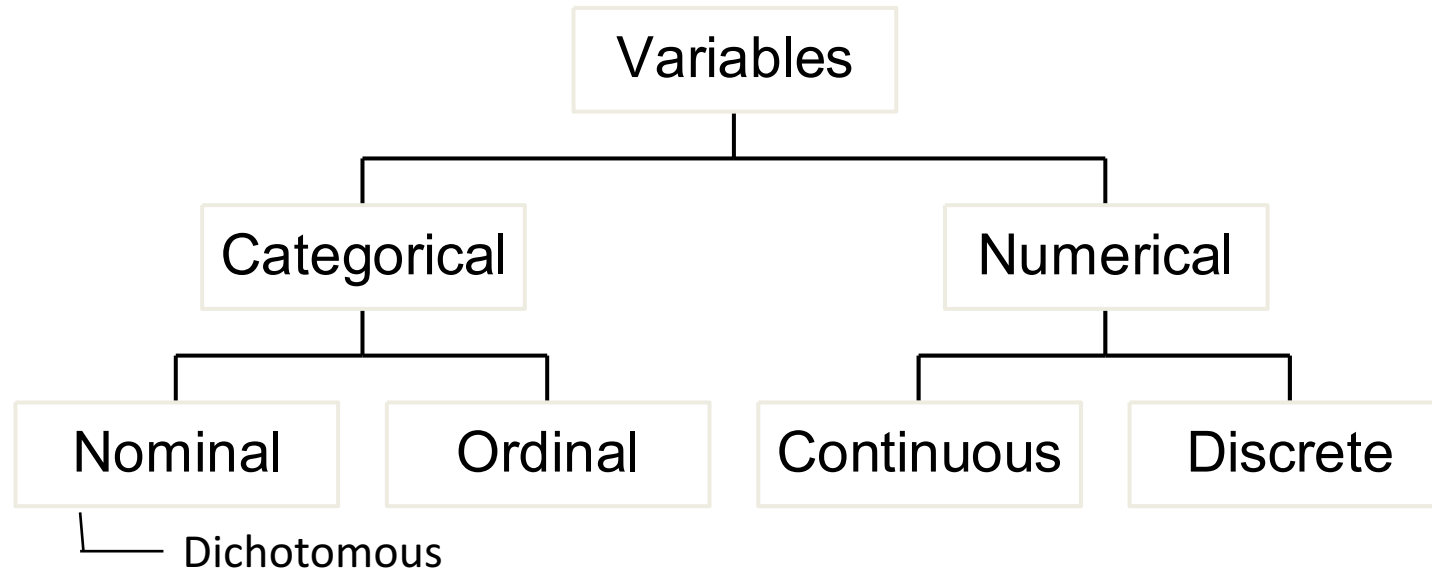
      high
cholesterol |      Freq.      Percent      Cum.
-----+-----
           no |         3,986         38.45         38.45

```

Outline for today

- ◇ Keyboard shortcuts
- ◇ Display preferences
- ◇ File types
- ◇ **Variable types**
- ◇ Immediate commands
- ◇ Logic operators
- ◇ Opening and importing data
- ◇ More data exploring

Types of Variables



Types of Variables

Categorical	
1. Nominal	No order
→ 1a. Dichotomous	No order, only 2 categories
2. Ordinal	Ordered
Numerical	
1. Discrete	Can be counted
2. Continuous	Can take any value within a given range

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	caccat	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

Outline for today

- ◇ Keyboard shortcuts
- ◇ Display preferences
- ◇ File types
- ◇ Variable types
- ◇ **Immediate commands**
- ◇ Logic operators
- ◇ Opening and importing data
- ◇ More data exploring

Immediate commands

- ◇ An immediate command does not require stored data to operate
 1. Calculator (display)
 - Random number generator
 2. Sample size calculator (power)
 3. Statistical/epidemiological immediate commands
 - (end in the letter i)

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

Stata

Add

+

Subtract

-

Multiply

*

Divide

/

To the power of

^

Natural log of

ln (#)

Base 10 log of

log10 (#)

Square Root

sqrt (#)

Immediate commands: random number generator

◇ From a uniform distribution:

- To generate a random number between **a** and **b**

○ `display runiform(a, b)`

- To generate a random integer between **[a, b]**

○ `display int(runiform(a, b))`

QUIZ

◇ Square Root of 457 _____

- `di sqrt(457)`

◇ Natural log of 342 _____

- `di ln(342)`

◇ Random integer between 1 and 100 (uniform distribution)

- `di int(runiform(1,100))`

Outline for today

- ◇ Keyboard shortcuts
- ◇ Display preferences
- ◇ File types
- ◇ Variable types
- ◇ Immediate commands
- ◇ **Logic operators**
- ◇ Opening and importing data
- ◇ More data exploring

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

Outline for today

- ◇ Keyboard shortcuts
- ◇ Display preferences
- ◇ File types
- ◇ Variable types
- ◇ Immediate commands
- ◇ Logic operators
- ◇ **Opening and importing data**
- ◇ More data exploring

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

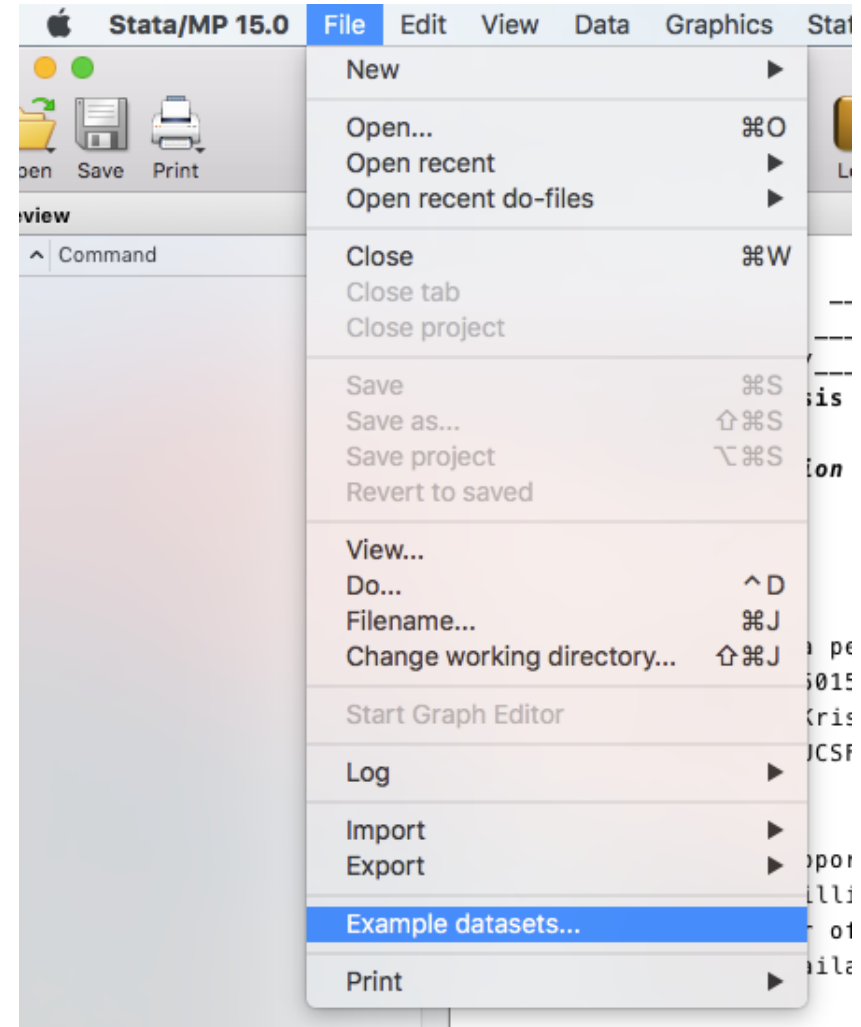
1. Command line

- 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

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.dtatsetURL`

◇ `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

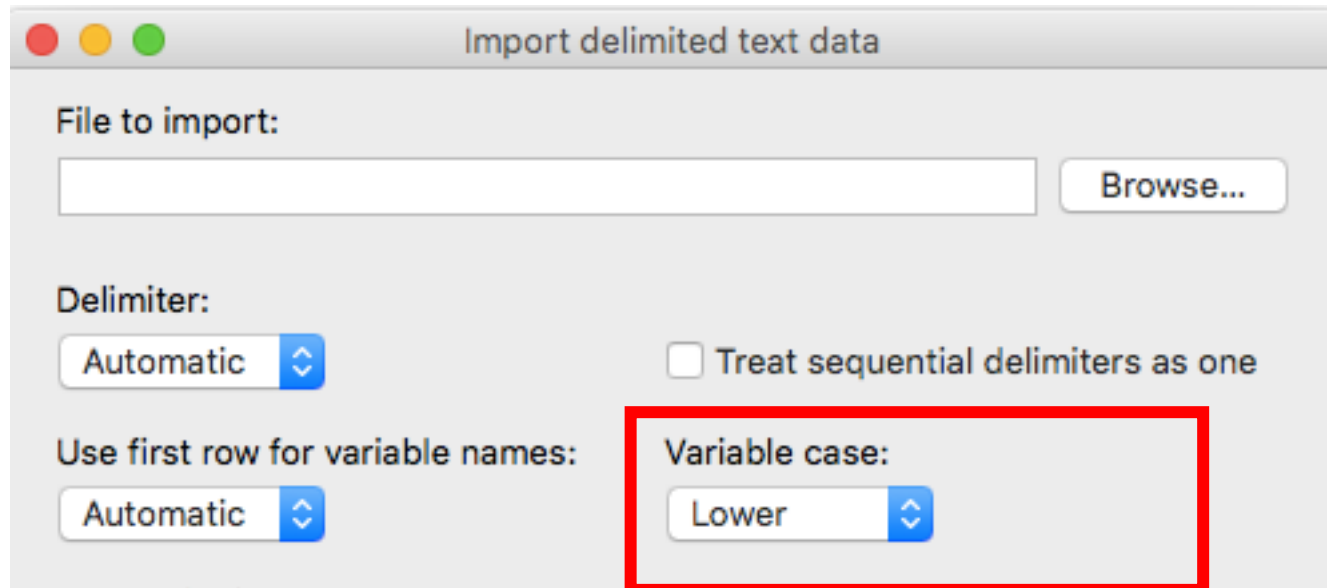
- `import delimited "yourfile.csv"`

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 odcb" for more information

Importing data

- ◇ Check that it worked after import

	id	age	htn	dm	smoke	chol	cac	dead
1	1	25	0	0	0	0	0	0
2	2	25	0	0	0	1	0	0
3	3	25	1	0	1	0	0	0
4	4	25	1	0	1	0	0	0
5	5	25	0	0	0	0	0	0
6	6	25	0	0	1	0	0	0
7	7	25	0	0	0	1	3	0
8	8	26	0	0	0	1	0	0
9	9	26	1	0	1	0	0	0

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 today

Acupuncture dataset

1. Download from course website
2. Check directory
3. `use accupuncture_baseline.dta`

If you have Stata version 13 or lower:

1. Download `accupuncture_baseline_V13.dta`
2. `use accupuncture_baseline_V13.dta`

Outline for today

- ◇ Keyboard shortcuts
- ◇ Display preferences
- ◇ File types
- ◇ Variable types
- ◇ Immediate commands
- ◇ Logic operators
- ◇ Opening and importing data
- ◇ **More data exploring**

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

Explore data: describe

`describe`

- Displays variable names, types, labels for all variables

`describe varlist`

- Describe 1 or more variables

Options

`, short`
`, fullnames`

- Describe only observations, size no variables
- 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
```

```
.
```

QUIZ

◊ How many observations (participants) in this study?

`describe, short -or-
count`

◊ How many participants have a migraine diagnosis at baseline?

`count if migraine==1`

`Bysort migraine: count`

Explore data: `isid`

`isid varname` – check if variable uniquely identifies the data

example `isid id`

*Note: no response given if variable uniquely identified the data but you will get this error message if variable does not uniquely identify data:

variable age does not uniquely identify the observations

Explore data: duplicates

`duplicates report`

`duplicates report varname`

`duplicates drop`

`duplicates drop varname`

`duplicates list`

`duplicates list varname`

- search for duplicate observations
- search for duplicate values of a variable
- drop duplicate observations
- drop duplicate values of a variable
- list duplicate observations
- list duplicate values of a variable

Explore data: duplicates

```
. duplicates report
```

Duplicates in terms of all variables

copies	observations	surplus
1	301	0

No duplicates

Explore data: 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

Explore data: codebook

```
. codebook group
```

group	intervention group
-------	--------------------

```
type: numeric (byte)
label: intervention
```

```
range: [0,1]
unique values: 2
units: 1
missing .: 0/301
```

```
tabulation: Freq.  Numeric  Label
              140      0    control
              161      1  acupuncture
```

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

Quiz

- What is the average baseline headache score [basehs] for participants diagnosed with migraine at baseline [migraine]?

```
sum basehs if migraine==1
bysort migraine: sum(basehs)
tab migraine, sum(basehs) mean
```

- What is the average baseline headache score [basehs] for males [sex] at baseline?

```
sum basehs if sex==0
bysort sex: sum(basehs)
tab sex, sum(basehs) mean
```

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

Number of Observations

					Total	Integers	Nonintegers
#	Negative				-	-	-
#	Zero				-	-	-
#	Positive				301	80	221
#							
#	#	Total			301	80	221
#	#	.	.	.	-		
6.75 94.75					301		
(More than 99 unique values)							

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>		✓	✓

Clearing data and closing STATA

- ◇ `clear` clears data set, **be sure to save first!**
- ◇ `cls` clears the console (where results are displayed)
- ◇ Save `"filename.dta"` save as new filename
- ◇ `save, replace` save and replace last version
- ◇ `exit, clear` closes STATA and clears data, **be sure to save first!**

In-class activity

Table 1: Baseline characteristics. Values are in numbers (percentages) of participants unless otherwise indicated

	Acupuncture (N)	Control (N)
Mean age in years (SD)	Mean (SD)	Mean (SD)
Female	N (%)	N (%)
Migraine diagnosis	N (%)	N (%)
Mean chronicity in years (SD)	Mean (SD)	Mean (SD)