

Lecture 3: Exploring data

BIOSTAT 212, Summer 2016



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

Questions from last week?

.DTA files: Where you data is saved: 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	

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

```
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. *Do not edit*

```

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

```

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 `blue`.
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

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

Acupuncture dataset

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

Baseline data today

Review: opening a data set

Three different ways:

1. Command line

- 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

3. Double click on .dta file

Explore data

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

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`

- Describe only observations, size no variables

`, fullnames`

- Give full variable names

Practice

◊ How many observations (participants) in this study?

`describe` **-or-**

`describe, short`

◊ How many variables in this dataset?

`describe` **-or-**

`describe, short`

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`

by versus bysort

by `varlist`: `stata_cmd`

bysort `varlist`: `stata_cmd`

- ◇ Most Stata commands allow the `by` prefix, which repeats the command for each group of observations for which the values of the variables in `varlist` are the same.
- ◇ `by` without the sort option requires that the data be sorted by `varlist`

Practice

◊ How many participants have a migraine diagnosis at baseline?

```
count if migraine==1
```

◊ How many male patients have migraine diagnosis at baseline?

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

Explore data: lookfor

lookfor "searchterm" – search variable names and labels

example lookfor "calcium"

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

Practice

◊ Does the variable “id” uniquely identify observations?

```
isid id          -or-  
duplicates list id
```

Does the variable “age” uniquely identify observations?

```
isid age
```

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

Practice

◇ How many unique values of age are there?

`codebook age` **-or-**

`codebook age, compact`

◇ What is the range (min and max) for chronicity (years with headache)?

`codebook chronicity` **-or-**

`codebook chronicity, compact`

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

Practice

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

```
sum basehs if migraine==1
```

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

```
sum basehs if migraine==0
```

- What is the median age?

```
sum age, detail
```

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

Practice

- What is the range (min and max) of baseline headache scores for participants who are older than or equal to age 50?

```
inspect basehs if age>=50
```

- What is the range (min and max) of baseline headache scores for participants who are younger than age 50?

```
inspect basehs if age<50
```

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

Practice

- ◇ Sort participants by id in ascending order

```
gsort +id
```

- ◇ What are the first and last id numbers?

```
browse id
```

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

Practice

- ◊ What is the chronicity (number of years with headache) of the 7th participant?

```
list chronicity in 7
```

- ◊ What is the mean baseline headache score in observations 1/10?

```
list basehs in 1/10, mean N sum clean
```

Make sure your data is sorted by id `[gsort +id]`

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

Practice

- What percent of participants with migraines are male?

`tab migraine sex, row` -or-

`tab sex if migraine==1` -or-

`bysort migraine: tab sex`

- What percent of participants were in the acupuncture intervention group?

`tab group`

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