Applied Epidemiology I: Data clearance
A review of using Stata

Enoch Yi-Tung Chen

Department of Medical Epidemiology and Biostatistics, Karolinska Insitutet

November 20, 2020

Acknowledgements

This course material in data clearance is based on my learning from Anastasia Lam's teachings in last year's Applied Epidemiology I lab sessions, and readings from *A First Course in Probability and Statistics* by Goldsman and Goldsman [1], *Principles of Biostatistics* by Pagano and Gauvreau [2], and *Biostatistics I* by Gabriel and Frumento [3].

Outline (also learning outcomes)

1. Set up working directory
2. Import and save data
   - Import
   - Save
3. Manage datasets
   - Merge
   - Append
4. Get to know the data
   - Summarize
   - Describe
   - Codebook
   - List
5. Manage variables
   - Numeric and string
   - Drop/Keep
   - Label
   - Rename, recode, generate, replace
   - Sort, by, if, in
   - Operators
6. References

Set up working directory

- Working directory is the folder where all your files are stored, and should be set each time you start.
- Where is it?

```
. cd
/Users/Desktop
. pwd
/Users/Desktop
```

- Change working directory

  -
    ```
    // Change working directory to Download
    . cd "/Users/Download"
    ```
  - Click File - Change Working Directory

Import and save data: Import

- **Excel (.xls or .xlsx)**

  ```
  import excel filename, clear firstrow
  ```

- **Delimited (.csv) or text (.txt)**

  ```
  import delimited filename, clear
  infile filename, clear
  ```

- **Stata (.dta)**

  ```
  use filename, clear
  ```

- **SAS (.xpt)**

  ```
  fdause filename, clear
  ```

Import and save data: Save

- Save your dataset as a Stata file:

  ```
  save "filename", replace
  ```

- The `replace` option lets you overwrite the existing dataset.

  ```
  save "filename", replace
  ```

Manage datasets: Merge

merge adds new variables from a second dataset to your existing dataset. (Make the dataset wider)

```
. sysuse cancer, clear
(Patient Survival in Drug Trial)

. gen id = _n

. keep id

. merge 1:1 id using cancer

    Result                           # of obs.
    -----------------------------------------
    not matched                             0
    matched                                48  (_merge==3)
    -----------------------------------------
```

Manage datasets: Append

`append` adds new observations to existing variables in your current dataset. (Make the dataset longer)

```
. use cancer_drug12, clear
(Patient Survival in Drug Trial)

. append using cancer_drug3.dta // append patients using drug 3
```

Get to know the data: Summarize

`summarize` gives summaries for all your variables, such as number of observations, mean, standard deviation, etc.

```
. sysuse cancer, clear
(Patient Survival in Drug Trial)

. keep if drug ==1 | drug == 2
(14 observations deleted)

. summarize age  // One variable only (age)

    Variable |        Obs        Mean    Std. Dev.       Min
-------------+-----------------------------------------------------
         age |         34    56.41176    6.010686         47
```

Get to know the data: Describe

`describe` gives descriptions for all your variables, such as storage type and labels.

```
. describe  age

              storage   display value
variable name   type    format  label variable label
--------------------------------------------------------------------
age             byte    %8.0g         Patient's age at start of exp
```

Get to know the data: Codebook

`codebook` is a combination of summarize and describe and will give a detailed summary of all your variables, including mean, sd, range, percentiles, missing, frequency, etc.

```
. codebook  age

----------------------------------------------------------------------------------------------------------------
age                                                                       Patient's age at start of exp.
----------------------------------------------------------------------------------------------------------------

                  type:  numeric (byte)

                 range:  [47,67]                      units:  1
         unique values:  15                       missing .:  0/34

                  mean:   56.4118
              std. dev:   6.01069

           percentiles:        10%       25%       50%       75%       90%
                                49        51        56        61        65
```

Get to know the data: List

`list` lists the observations of specified variables.

```
. list      age if age < 50

     +-----+
     | age |
     |-----|
 12. |  49 |
 15. |  49 |
 18. |  49 |
 25. |  49 |
 33. |  47 |
     +-----+
```

Manage variables: Numeric and string

Numeric: byte, integer, long, float, double – all types of numeric variables that just differ based on min and max length
String: character variables with a certain length (*str*#)

Manage variables: Drop/Keep

- drop is used to delete variables or observations.
- keep is used to keep variables or observations.

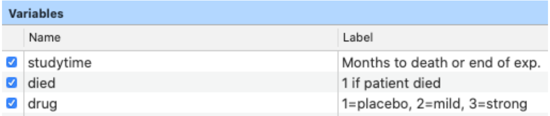
```
. sysuse cancer, clear
(Patient Survival in Drug Trial)
. drop if drug ==1 | drug == 2
(34 observations deleted)

. sysuse cancer, clear
(Patient Survival in Drug Trial)
. keep if drug ==1 | drug == 2 // So drug == 3 will be dropped
(14 observations deleted)
```

Manage variables: Label

- `label` helps you keep track of your dataset and variables, and helps others understand your data.
- `label define` to a variable (usually the one you defined)
- `label values` attaches the labels defined using.

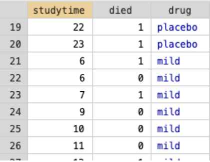
```
. label variable drug "1=placebo, 2=mild, 3=strong"
```



Variables

	Name	Label
☑	studytime	Months to death or end of exp.
☑	died	1 if patient died
☑	drug	1=placebo, 2=mild, 3=strong

```
. label define drug 1 "placebo" 2 "mild" 3 "strong"
. label values drug drug
```



	studytime	died	drug
19	22	1	placebo
20	23	1	placebo
21	6	1	mild
22	6	0	mild
23	7	1	mild
24	9	0	mild
25	10	0	mild
26	11	0	mild

Manage variables: Rename, recode, generate, replace

- `rename` changes the name of a variable.

```
. rename died death
```

- `recode` changes variable values.

```
. recode drug (3=4)
```

- `generate` creates a new variable.

```
. generate placebo = 1 if drug == 1
```

- `replace` replaces existing variables (or variable values).

```
. replace placebo = 0 if drug != 1
```

Manage variables: Sort, by, if, in

- sort orders observations in ascending order.

```
. sort death
```

- by executes a command within a specified variable (e.g. by age group), but data should be sorted first.

```
. by death: summarize
```

- bysort combines the by and sort commands into one.

```
. bysort death: summarize // by death, sort: summarize
```

- if is used to select by a condition.

```
. list age if death == 1
```

- in is used to select by observations.

```
. gen id = _n
. list id 1/10
```

Manage variables: Operators

Operator	Purpose	Example
==	Evaluates if true/false	summarize if sex==1
~= or !=	Indicates 'not equal'	summarize if sex!=0
<, <= >, >=	Less than (equal to) or greater than (equal to)	summarize if age<35
&	Indicates 'and'	summarize outcome if sex==1 & age>=60
|	Indicates 'or'	gen x=1 if a==1 & (b==1 | c==1)

References

1. David Goldsman PG. *A First Course in Probability and Statistics*. Georgia Institute of Technology, 2020.
2. Marcello Pagano KG. *Principles of Biostatistics*. Cengage Learning, Inc, 2000. ISBN 0534229026.
3. Erin Gabriel PF. Epidemiology PhD program, Karolinska Institutet, 2020.