Data Analysis using Stata

Kristin Bott
kbott @ reed.edu

<< Log in to a computer while you’re waiting >>

blogs.reed.edu/datablog
Intro / K. Bott + Data @ Reed

Associate Director, Instructional Technology Services
kbott@reed.edu / ETC 225

Geospatial / Statistical / Computational software
Data analysis + presentation

Data @ Reed : workshops + data support
data@reed.edu
reed.edu/data-at-reed
a very skeletal research process

Question

Data

Results

Conclusions
data @ reed / research + data support

Data analysis

Data discovery

Data citation

Data visualization

Data wrangling

Data management
Where we’re headed (<60 min)

• Step zero : how to find and pilot Stata
• Step one : basic data skills
• Step two : external data
• Step three : data analysis
• Help and resources
where to find / how to access
Where you can find Stata

• Around campus:
  – IRC’s (ETC)
  – Library
  – Eliot 110 / PPW (social sciences)
  – Psychology computer lab

• “Grad Plan” / option for buying your own copy
Accessing Stata

On your computers --

> Applications

>> StataSE

open Stata!

... and let’s take a look.
Review

Variables

Command

Results
What did I do?

Current action is here!

• Point + click (GUI) vs command-line

What data am I working with?

Data details
Stata datasets

• Pre-loaded, useful for training / learning
• Type:

  . set more off
  . sysuse dir

  . sysuse auto
  (1978 Automobile data)
stata: basic data skills
Data is loaded: now what?

What sort of information do you want to know about your data?

How can we get to this information?
Some things you may want to know

**Range** of data / outliers

**Missing** values [How many? How coded?]

**Types** of variables

**Variation** of data

(step one: look at your data – data browser)

锥 browse

K. Bott / Instructional Technology Services
Reed College / Portland, OR
first-glance tools

`.summarize (whole dataset)`
`.summarize rep78
   Observation / mean / StdDev / Min / Max`

`.describe (whole dataset)`
`.describe [var]
   var name / storage type / disp format / value label / var label`

`.codebook (whole dataset)`
`.codebook [var]
   type / range / units / unique / missing / mean / std dev / %tiles`
first-glance tools

.summarize (whole dataset)
.summarize [var]
   Observation / mean / StdDev / Min / Max

.describe (whole dataset)
.describe rep78
   var name / storage type / disp format / value label / var label

codebook (whole dataset)
.codebook [var]
type / range / units / unique / missing / mean / std dev / %tiles
first-glance tools

.summarize (whole dataset)
.summarize [var]
  Observation / mean / StdDev / Min / Max

.describe (whole dataset)
.describe [var]
  var name / storage type, disp format / value label / var label

.codebook (whole dataset)
.codebook [var]
  type / range / units / unique / missing / mean / std dev / %tiles
Variable **storage types**

- **Describe or variable window shows “storage type”**

- **Numbers**
  - byte, int(eger), long, float, double
  - vary in precision + memory they use (bytes)

- **Letters:**
  - String
  - str1, str2, ... str244

- **Question :: Why does this matter?**
Variable **storage types**

- Describe or *variable window* shows “storage type”

- Numbers
  - byte, int(eger), long, float, double
  - vary in precision + memory they use (bytes)

- Letters:
  - String
  - str1, str2, ..., str244

- **Question :: Why does this matter?**

*You can’t find the mean of words...*
first-glance tools

.summarize (whole dataset)
.summarize [var]
  Observation / mean / StdDev / Min / Max

.describe (whole dataset)
.describe [var]
  var name / storage type / disp format / value label / var label

codebook (whole dataset)
.codebook rep78
type / range / units / unique / missing / mean / std dev / %tiles
first-glance tools

`.summarize (whole dataset)`

`.summarize [var]
  Observation / mean / StdDev / Min / Max`

`.describe (whole dataset)`

`.describe [var]
  var name / storage type / disp format / value label / var label`

`.codebook (whole dataset)`

`.codebook [var]
  type / range / units / unique / missing / mean / std dev / %tiles`
some first-glance tools

.tabulate foreign
variable / frequency / percent / cumulative %

.tabulate [var1] [var2]
what does this do?

.tabulate [var2] [var1]
what does this do?
some first-glance tools

```
.tabulate  (whole dataset)
.tabulate [var]
  variable / frequency / percent / cumulative %

.tabulate foreign rep78
  what does this do?

.tabulate [var2] [var1]
  what does this do?
```
some first-glance tools

.tabulate (whole dataset)
.tabulate [var]
variable / frequency / percent / cumulative %

.tabulate make rep78

what does this do?

Hmm ... do I have observations for the same make + repair record?

.tabulate [var2] [var1]

what does this do?
kbott’s first-glance toolbox

For *dataset* or [var] or [var1] [var2]

`.summarize`
`.codebook`
`.describe`
`.tabulate`
`.inspect`  
(be lazy! – use only a few (unique) letters for commands)
`.browse`
`.list`
Basics: \textit{[browse]} subsets of data

Can also use \texttt{view} to see results in main window

\begin{itemize}
  \item \texttt{browse if foreign == 1} \hspace{1cm} (equals)
  \item \texttt{browse if foreign ~= 1} \hspace{1cm} (not equal)
  \item \texttt{browse if foreign != 1} \hspace{1cm} (not equal)

  \item \texttt{browse if mpg > 5 \& mpg < 20} \hspace{1cm} (& joins multiple)

  \item \texttt{browse make mpg in 1/10} \hspace{1cm} (range of values)
\end{itemize}
Basics: alter data

.sort var (sorts from low to high)
.drop var (drop variable, keep rest)
.keep var (keep variable, drop rest)
.replace var (replace existing variable)

.generate var (generate new variable)
.eegen var (extended generate)

.clear [dataset]
(clears from memory, does not erase data)
Basics (stock dataset)

- Histograms
  - `sysuse auto`
  - `hist price, freq`
  - `hist price, freq bin(5)`
  - `hist price, freq bin(15)`
  - `hist price if foreign==1, freq bin(15)`
### Title

[R] **histogram** — Histograms for continuous and categorical variables

### Syntax

```
histogram varname [if] [in] [weight] [, [continuous_opts | discrete_opts] options]
```

#### continuous_opts

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>bin(#)</code></td>
<td>set number of bins to #</td>
</tr>
<tr>
<td><code>width(#)</code></td>
<td>set width of bins to #</td>
</tr>
<tr>
<td><code>start(#)</code></td>
<td>set lower limit of first bin to #</td>
</tr>
</tbody>
</table>

#### discrete_opts

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>discrete</code></td>
<td>specify that the data are discrete</td>
</tr>
<tr>
<td><code>width(#)</code></td>
<td>set width of bins to #</td>
</tr>
<tr>
<td><code>start(#)</code></td>
<td>set theoretical minimum value to #</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>options</th>
<th>Description</th>
</tr>
</thead>
</table>
Basics (stock dataset)

• Scatterplots
  . `sysuse auto`
  . `scatter mpg weight`
  . `scatter mpg weight || lfit mpg weight`
Analyze: two variables

- correlate mpg weight
- regress mpg weight
Analyze: two variables

. correlate mpg weight
. regress mpg weight

Examine relationship by foreign/domestic vehicles
. by foreign: regress mpg weight

Examine relationship for only foreign vehicles
. regress mpg weight if foreign==1
Homework #3

additional + specific tools
Homework #3

• Measures of central tendency + dispersion
  . summarize
  . tabulate

• Visualization
  . scatter

• Analysis
  . regress
  . predict
Homework #3

• Measures of central tendency + dispersion
  . summarize
  . tabulate

• Visualization
  . scatter

• Analysis
  . regress
  . predict
  . help predict
Help + additional resources
Help! + additional tools

• Stata Help Menu
  – contents
  – search
  – command

• At the command line
  – help
  – search
  – findit

• External (to Stata) resources

• Track work via do files
Key to collaboration: Do files

Save you time for repetitious tasks
Minimizes errors
Store your data analysis process

.doedit
or via GUI
Kbott Office Hours
by appointment
help documentation + more
D@R site: reed.edu/data-at-reed

kbott@reed.edu | ETC 225