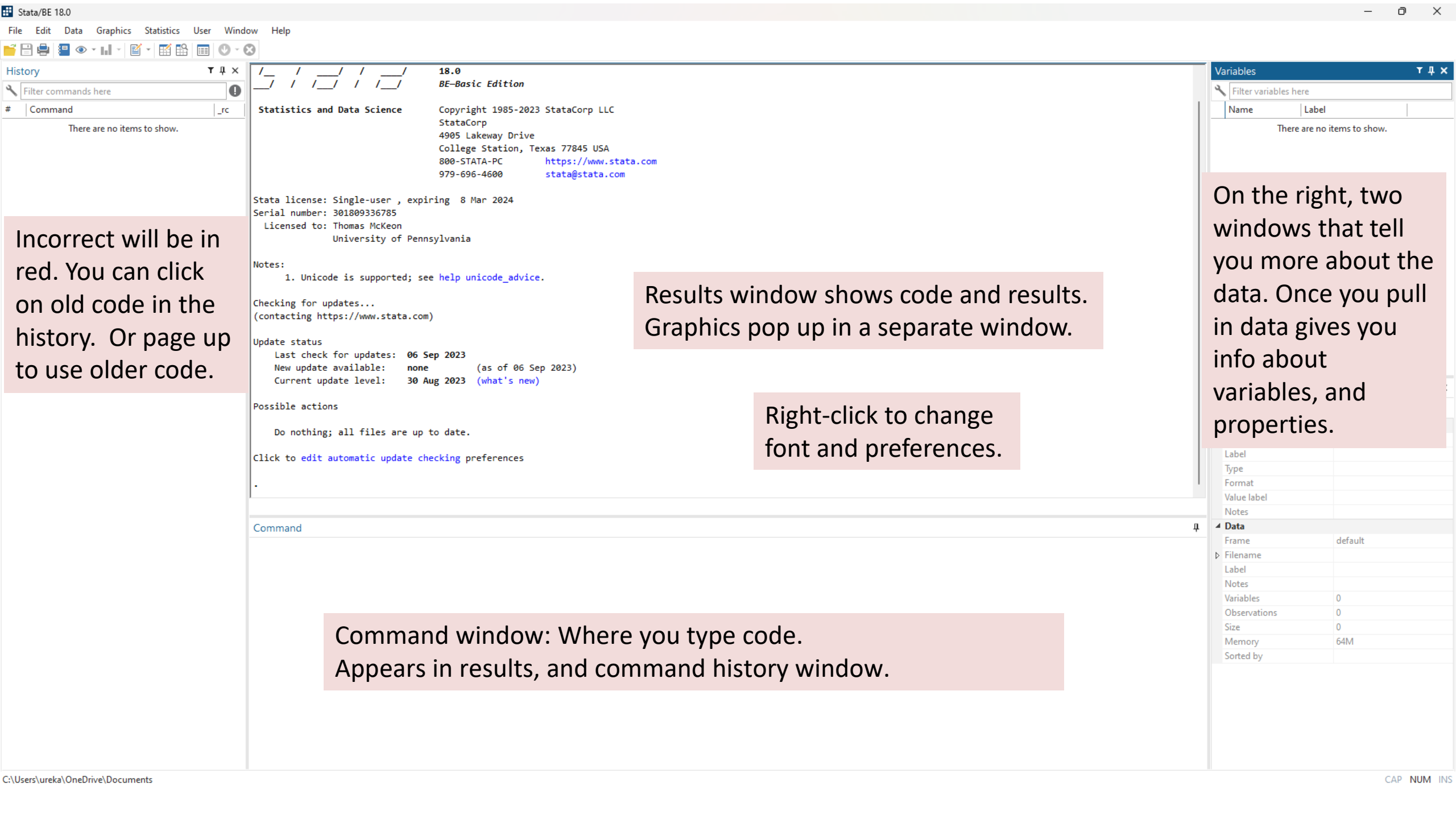
PUBH 501 Biostatistics

STATA INTRODUCTION, TABLES, AND GRAPHS

Overview

- Feel free to have Stata open and play along
- Tour of Stata
- Importing data from excel
- Common commands
- Creating tables and graphs
- Summary statistics


Stata/BE 18.0
File Edit Data Graphics Statistics User Window Help
History
Filter commands here
Command _rc
There are no items to show.
/__ / ____/ / ____/ 18.0
___/ / /___/ / /___/ BE—Basic Edition
Statistics and Data Science Copyright 1985-2023 StataCorp LLC
StataCorp
4905 Lakeway Drive
College Station, Texas 77845 USA
800-STATA-PC https://www.stata.com
979-696-4600 stata@stata.com
Stata license: Single-user , expiring 8 Mar 2024
Serial number: 301809336785
Licensed to: Thomas McKeon
University of Pennsylvania
Notes:
1. Unicode is supported; see help unicode_advice.
Checking for updates...
(contacting https://www.stata.com)
Update status
Last check for updates: 06 Sep 2023
New update available: none (as of 06 Sep 2023)
Current update level: 30 Aug 2023 (what's new)
Possible actions
Do nothing; all files are up to date.
Click to edit automatic update checking preferences
.
Command
Variables
Filter variables here
Name Label
There are no items to show.
Label
Type
Format
Value label
Notes
Data
Frame default
Filename
Label
Notes
Variables 0
Observations 0
Size 0
Memory 64M
Sorted by
Incorrect will be in red. You can click on old code in the history. Or page up to use older code.
Results window shows code and results. Graphics pop up in a separate window.
Right-click to change font and preferences.
On the right, two windows that tell you more about the data. Once you pull in data gives you info about variables, and properties.
Command window: Where you type code.
Appears in results, and command history window.
C:\Users\ureka\OneDrive\Documents
CAP NUM INS

Importing data

- Know your data first
  - Check variable names and formats
- Can import many types of data, including Excel, CSV, text files, SAS, SPSS
- Start with -import excel—
  - File→ import→ excel spreadsheet
  - Import first row as variable names

Datasets

- Practice datasets come pre-installed in Stata
- Low birth weight data we will use in class
- Gun ownership by state we will use in class

Do-files

- Don’t use Stata without them
- A place to store all of your commands,
- Store comments and notes about your project
  - # your comment
  - Code // your comment
  - /* your comment */
- Bookmarks help you keep your place

Log files

- A record of your Stata session
- Allow you to save all your results in a text or Stata file
- Can paste from your log file to Microsoft Word. Helps if you copy tables using the “copy table” option, then you can format them as tables and not plain text
- You can create a new log file, or add to an existing one.

Variable rules

- Names can be 32 characters. Cannot be system-defined name
  - For example, cannot be “tab” which is a command
- Can’t start with a number
- Variables are either string or numeric
- Variables can store different amount of information

Storage type	Minimum	Maximum	Closest to 0 without being 0	Bytes
`byte`	-127	100	± 1	1
`int`	$-32{,}767$	32,740	± 1	2
`long`	$-2{,}147{,}483{,}647$	2,147,483,620	± 1	4
`float`	$-1.70141173319 \times 10^{38}$	$1.70141173319 \times 10^{38}$	$\pm 10^{-38}$	4
`double`	$-8.9884656743 \times 10^{307}$	$8.9884656743 \times 10^{307}$	$\pm 10^{-323}$	8

Precision for `float` is 3.795×10^{-8}.

Precision for `double` is 1.414×10^{-16}.

Variable rules

- String variables can hold 80 characters of any type
- Missing observations
  - Numeric variables missing is “.”
  - String variable missing is “ “

Labeling variables

- -label var- *variable* “variable label”
- -label define- *formatname* 0 “no” 1 “yes”
- -label val- *variable formatname*

```
label var eating “Respondent was eating during the interview”
label def YNFMT 0 “no” 1 “yes”
label val eating YNFMT
```

Need to know about commands

- Everything is case sensitive
- My most used commands
  - -count-
  - -codebook-
  - -describe-
  - -tab-
  - -display-
  - -help-

Descriptive stats

Frequencies

- Command: -tabulate-
  - Can type any letters from “ta” on, including -tab-, -tabul-, etc.
- Available options
  - -after the command, enter a “,” to add options. Common options
  - -missing-, can be shortened to -m-
  - -nolab-, shows you the underlying values instead of the value labels
  - -summarize-, allows you to look at summary statistics stratified by another variable
- -tab1- command shows the frequencies of a list of variables
- -proportion- displays only the percentages

Summary stats

- Command: -summ-
- Available options
  - -detail- shows more info about the variable, including range, IQR, median
- Another command: -tabstat-
  - Option -stats- allows you to call specific stats like mean, sd, and median
  - `tabstat variable, stats(n, mean, sd, p25, p50, p75)`
  - Does not support mode, why? Because there could be more than one. Need to use -tab-

Graphs

Some useful graphs

- Histograms
- Bar charts
- Box plots
- Scatter plots
- *graphs are the only thing that doesn’t show up in the results window. So you must save them or copy into Word. You can also paste them into your log file.
- *you can edit all graphs through the Stata graph editor. Also through the do-file

Histogram

- Visual display of frequency distribution for continuous variables
- Command: -histogram- *variable*, -frequency-
- Options
  - Can stratify by another variable with -by(*variable*)- option
  - Can change the width of the bars or the number of bars (called “bins”)
  - -frequency- option shows the frequency, or number, or observations in the value range. -percent- shows the percentage of all observations in that range. -percent- will sum to 100
  - Add a title with -title- “title” option

- `histogram age, freq bin(10)`

Box and whisker / box plot

- Different way to look at continuous data
  - 25th and 75th percentiles, median, outliers
- Command: -graph box- *variable*
- Options
  - Can stratify by another variable with -over(*variable*)

- `graph box age, over(sex)`

Violin Plot

- Modified version of a box-plot which adds the estimated kernel density to each plot.
- To draw this plot in Stata you will need to install a user-written command called vioplot.

- Installation: `ssc install vioplot, replace`

- Command: `vioplot momweight`

Bar graph

- Display of frequency of categorical variables
  - Nominal and ordinal variables
- Command: -graph bar- (percent), over (*variable*)
- Command: -graph bar- (count), over (*variable*)
- Options
  - Can stratify by another variable with an additional-over(*variable*)- option
  - Can sort by percentage with -sort-
  - Add a title with -title- “title” option
  - Can use another summary statistic instead of count or percent, including median, sum, min, *

- `graph bar (percent), over(smokepreg, sort(1) descending)`
- `graph bar (count), over(docvisits, sort(1) desc) over(smokepreg) title(“Doctor visits by moking status during pregnancy”)`

Dot plot

- Plot single variable , grouped vertically
- Command: --dotplot- *variable*
- Options
  - Can stratify by another variable with an additional-over(*variable*)- option
- `dotplot bwt`
- `dotplot bwt, over(smokepreg)`

Scatter plot

- Plot two continuous variables
- Command: --scatter- *variable1 variable2*
- Options
  - Can stratify by another variable with an additional-over(*variable*)- option

- `scatter momweight bwt`
- `scatter momweight bwt, by(smokepreg)`

Activities

Activity 1, Heights

- Download Heights data from canvas
  - Import data and label variables/values
  - Find mean, median, dispersion for heights
  - Creating a histogram with labels

Activity 2, Low Birth Weight, Part 1

- Identify level of measurement (LOM) for each variable in the data file
  - Without running any analyses, do you think the variables will be normally distributed?
  Why/why not?
  - Find mean, median, for the variables in the file; does the distribution of these variables appear to be normal given these values?

Activity 2, Low Birth Weight, Part 2 (Graphing)

- Create histograms for the momweight and age.
  - Does the distribution of these variables appear to be normal based on the histogram?
  - What do you think would happen to measures of central tendency and dispersion if we had fewer observations? What if we had more observations?
  - What would happen to the range if we had more and/or and fewer people in the class?
  - Challenge each other to create different bin sizes/widths and create these different histograms.
- - Discuss with your group which provides the best representation of the data.