

Lecture 5: Data Visualization

BIOSTAT 212



Amanda Irish | amanda.irish@ucsf.edu
Department of Epidemiology and Biostatistics
University of California San Francisco

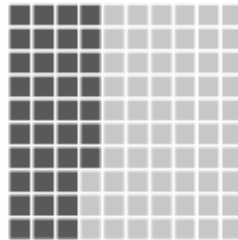
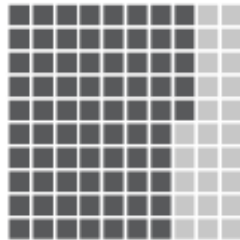
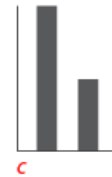
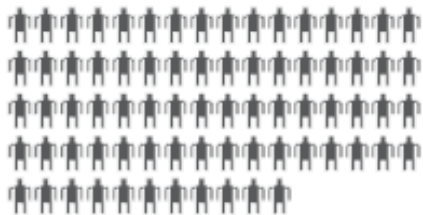
Warm up

Take a minute to sketch out on your own:

How many ways can you represent....

75 and 37

75 and 37



<https://visual.ly/blog/45-ways-to-communicate-two-quantities/>

Good visualization allows you to:

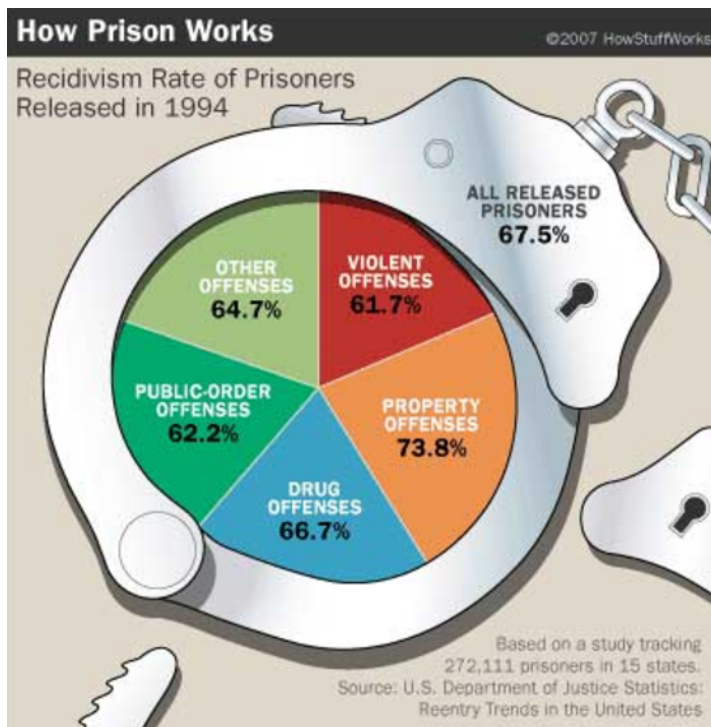
Explore  Analyze  Communicate

- ◇ Summarize information
- ◇ See new relationships
- ◇ Make comparisons
- ◇ Easily communicate your findings

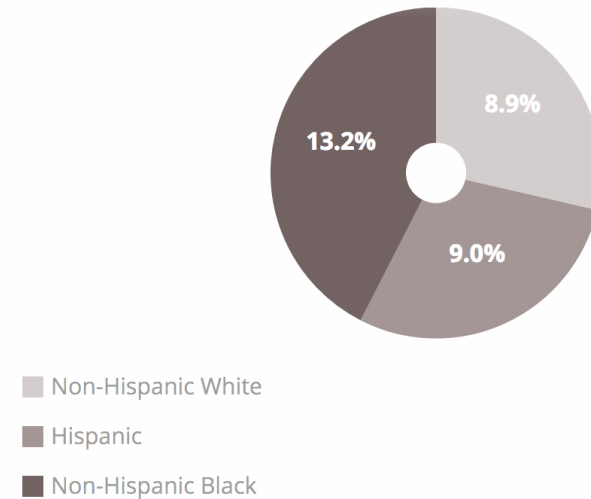
Visualizing data: rules of thumb

- ◇ Think of your audience
- ◇ What is the message you want to convey?
- ◇ Keep it simple
- ◇ Color blind friendly
- ◇ Limited space in publications → usually pie charts and bar graphs take up less room when written as a sentence
 - Also, don't use pie charts

Avoid: percentages that don't add up to 100



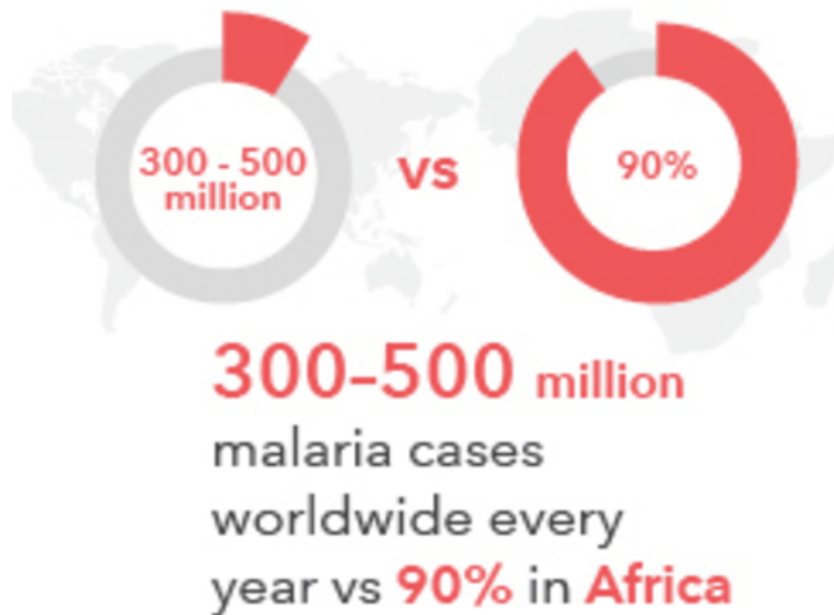
PRETERM BIRTH BY RACE & ETHNICITY



Preterm defined as less than 37 weeks gestation.

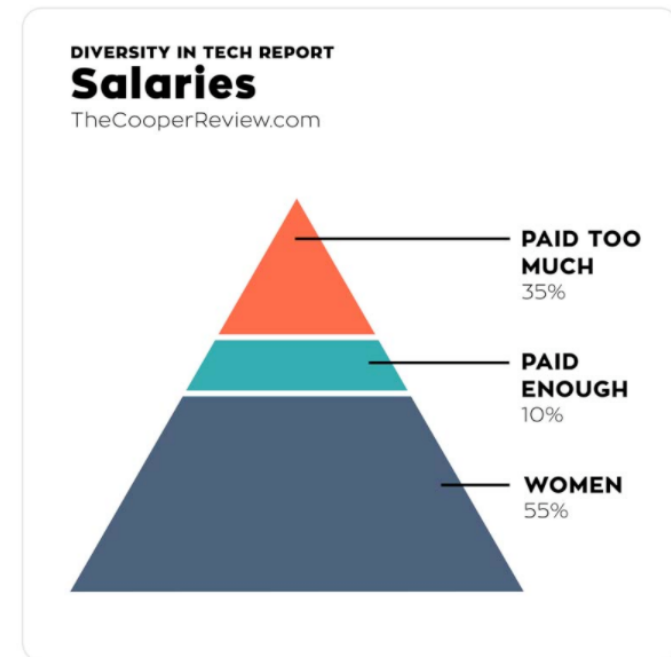
Source: CDC, 2015

Avoid: comparisons that don't make sense



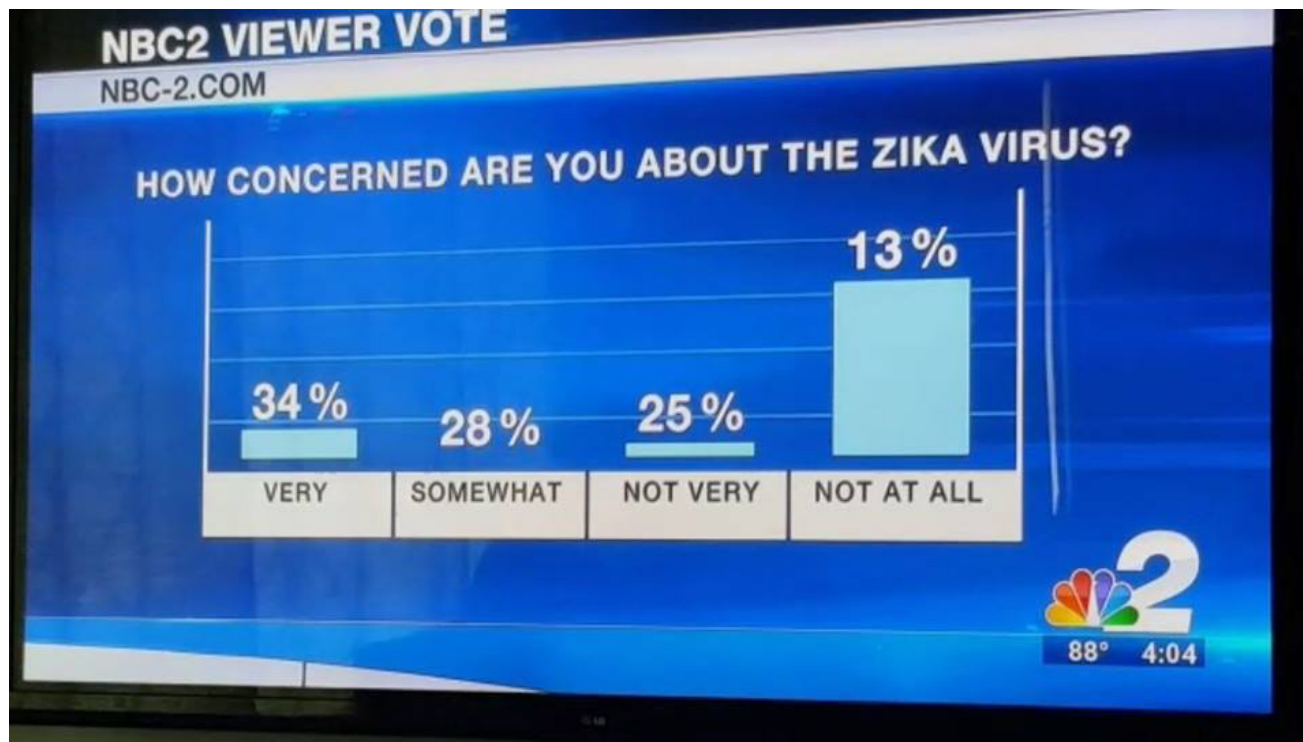
7. Salaries

Salaries are also diverse.

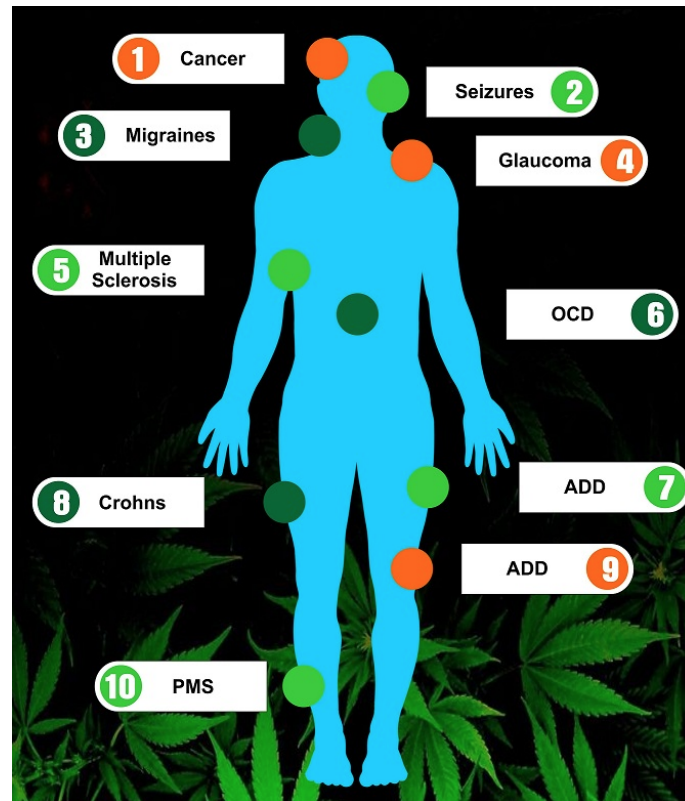


<http://viz.wtf>

Avoid: sizes that don't match percentages



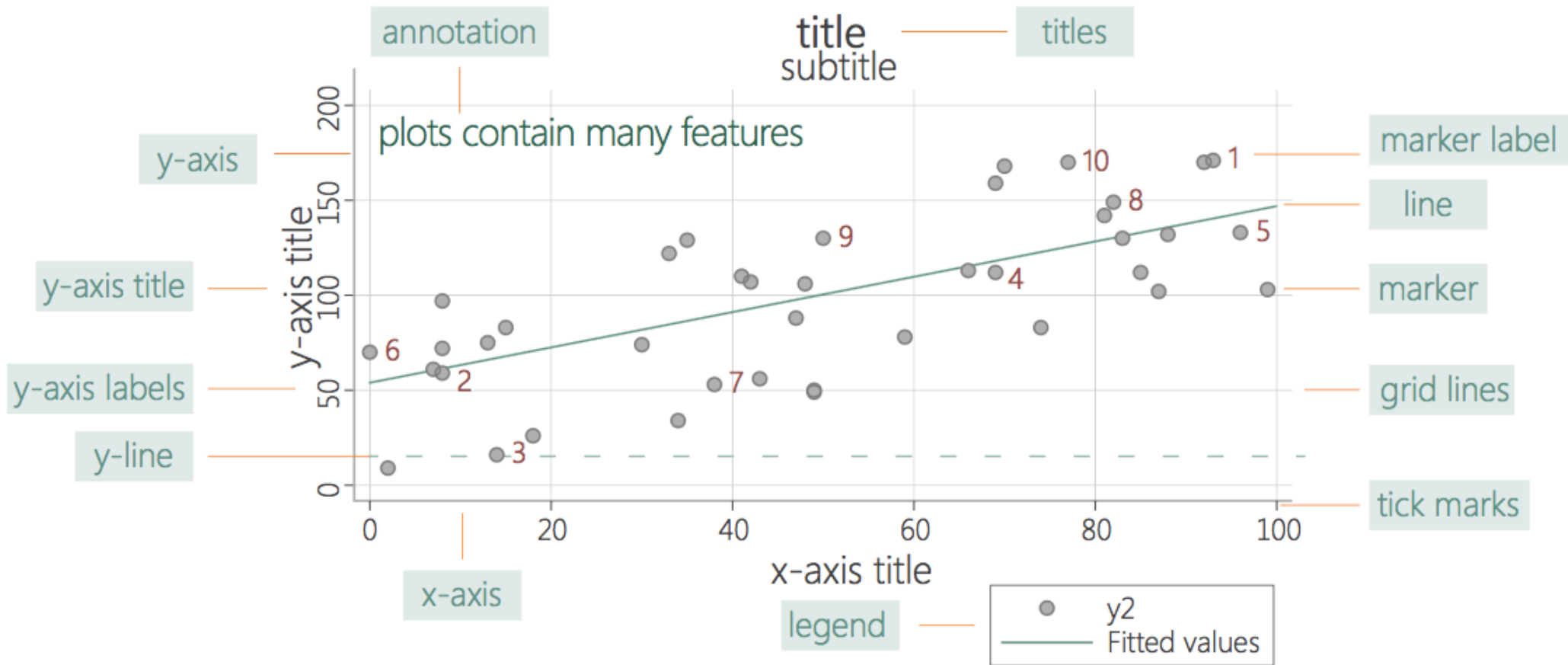
Glaucoma of the shoulder? PMS of the shin?



Graphs in Stata

- ◇ Histogram
 - ◇ Bar chart
 - ◇ Box plot
 - ◇ Dot plot
 - ◇ Scatter plot
 - Linear fit
 - ◇ Layering graphs
 - ◇ Combining graphs
 - ◇ Saving
 - ◇ Exporting
 - ◇ Using putdocx for Word
-
- ◇ Figure options
 - Titles, captions, notes
 - Axis titles and legends
 - Colors
 - Region, margins
 - Line patterns and width
 - Marker symbols and size
 - Text size
 - Schemes

ANATOMY OF A PLOT



Stata plot help files

help histogram

help graph bar

help box plot

help dot plot

help scatter

Let us show you some bar charts:

```
. sysuse citytemp

. graph bar (mean) tempjuly tempjan, over(region)
    bargap(-30)
    legend( label(1 "July") label(2 "January") )
    ytitle("Degrees Fahrenheit")
    title("Average July and January temperatures")
    subtitle("by regions of the United States")
    note("Source: U.S. Census Bureau, U.S. Dept. of Commerce")
    (click to run)

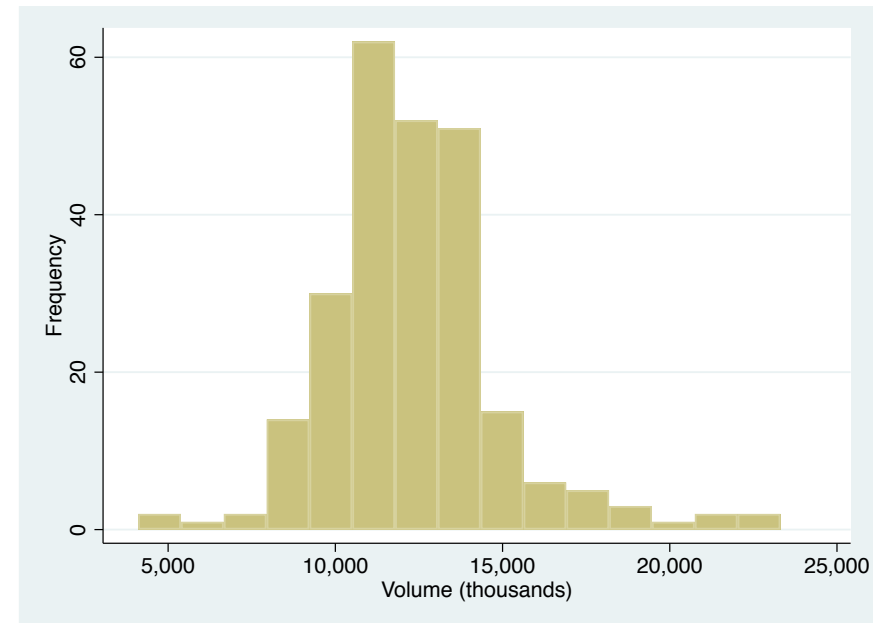
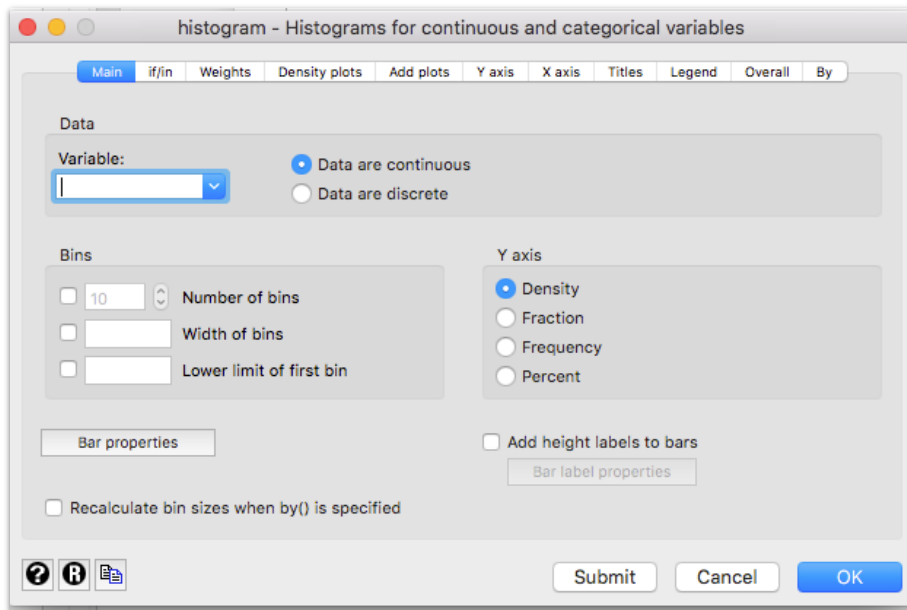
. sysuse citytemp, clear

. graph hbar (mean) tempjan, over(division) over(region) nofill
    ytitle("Degrees Fahrenheit")
    title("Average January temperature")
    subtitle("by region and division of the United States")
    note("Source: U.S. Census Bureau, U.S. Dept. of Commerce")
    (click to run)

. sysuse nlsw88, clear

. graph bar (mean) wage, over(smsa) over(married) over(collgrad)
    title("Average Hourly Wage, 1988, Women Aged 34-46")
    subtitle("by College Graduation, Marital Status,
        and SMSA residence")
    note("Source: 1988 data from NLS, U.S. Dept. of Labor,
        Bureau of Labor Statistics")
    (click to run)
```

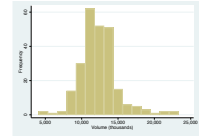
Histogram



Graphics > Histogram

```
sysuse sp500  
histogram volume, frequency
```

Histogram syntax



histogram `varname`

Options

- , normal
 - overlay normal distribution
- , kdensity
 - overlay smoothed estimate of distribution
- , discrete
 - specify variable is dichotomous
- , bin(`#`)
 - set number of bins
- , frequency
 - draw using frequencies (default is density)
- , by(`varname`)
 - histogram for every category of `varname`

Bar chart

graph bar - Bar charts

Main Categories if/in Weights Options Bars Y axis Titles Legend Overall By

Type of data

- ☒ Graph of summary statistics
- ☐ Graph of frequencies within categories
- ☐ Graph of percent of frequencies within categories
- ☐ Graph actual data (asis)

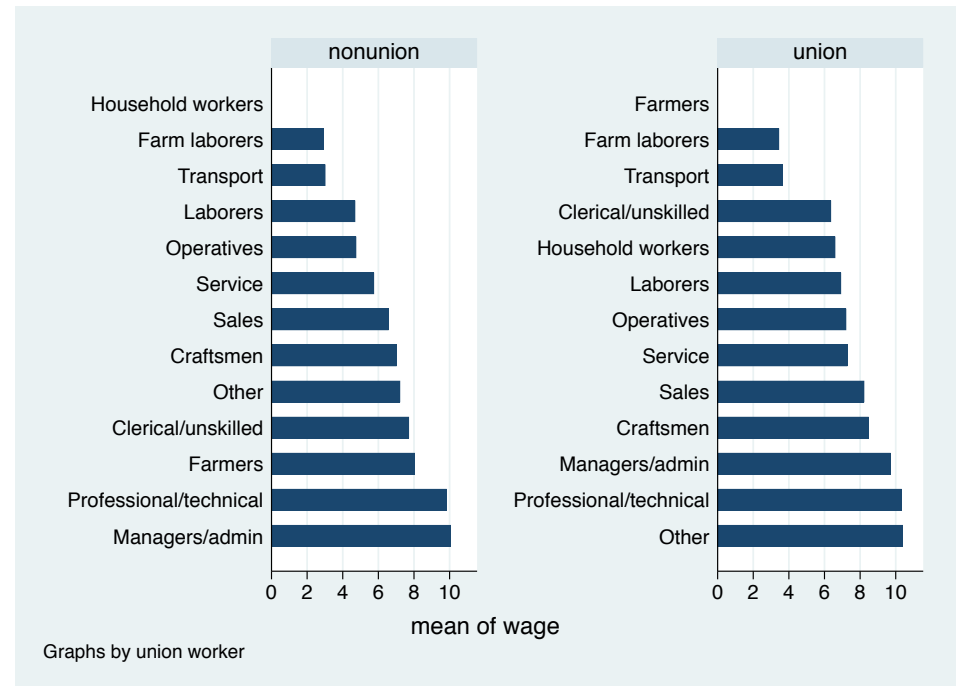
Orientation

- ☒ Vertical
- ☐ Horizontal

Statistics to plot

	Statistic	Variables
☒ 1:	Mean	
☐ 2:	Mean	
☐ 3:	Mean	
☐ 4:	Mean	
☐ 5:	Mean	
☐ 6:	Mean	
☐ 7:	Mean	
☐ 8:	Mean	

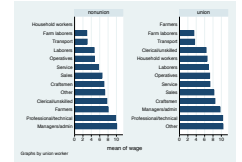
Submit Cancel OK



Graphics > Bar chart

```
sysuse nls88, clear
graph hbar wage, over(occ, sort(1)) by(union)
```

Bar chart syntax



```
graph bar (stat) numeric_var[s], [over(cat_var)]
```

```
graph hbar (stat) numeric_var[s], [over(cat_var)]
```

Options

(stat)

(mean median p1 p2 ... p99

sum count percent min max)

, over(varname)

- bar plot for each category of varname

, missing

- include missing as category of varname

, stack

- stack bars

Load
acupuncture
data set

```
use "acupuncture.dta", clear
```

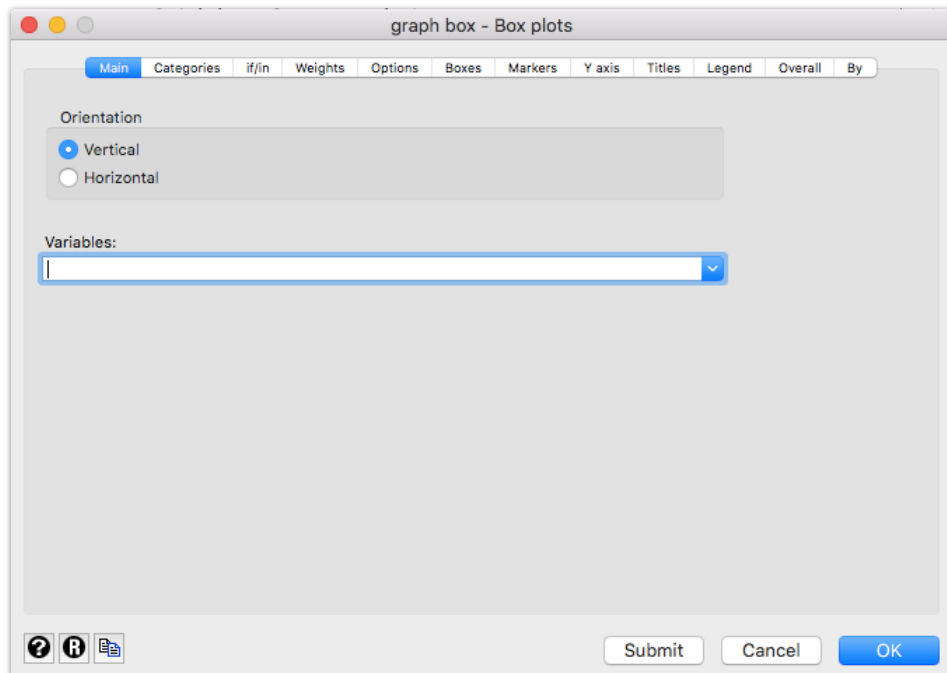
Acupuncture trial dataset

- ◇ **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

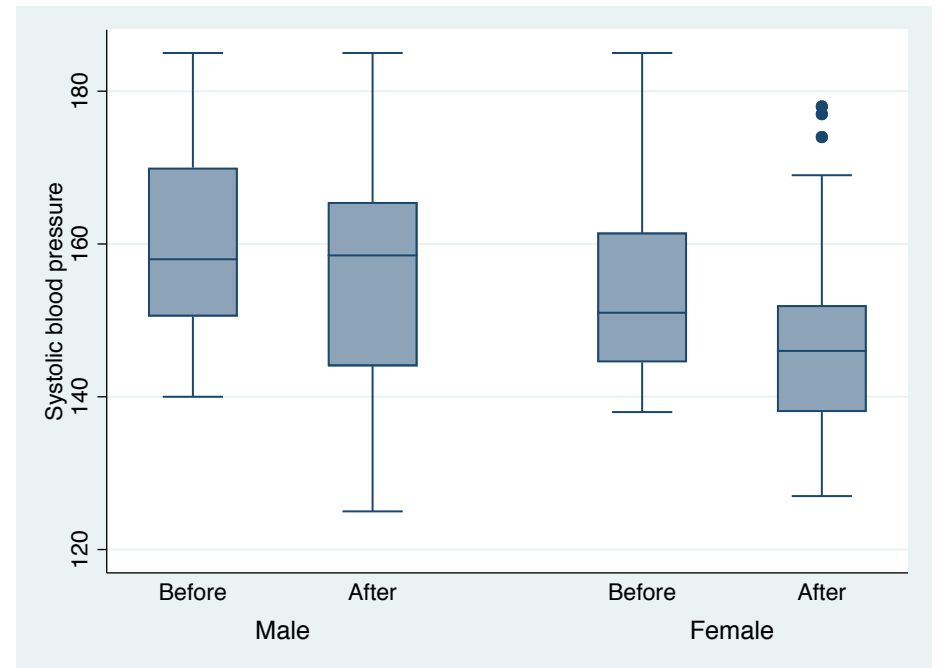
Practice

- ◇ Create a histogram of baseline general health (health_bl)
- ◇ Create a bar chart showing the mean of baseline general health for the treatment vs. control groups

Box plot

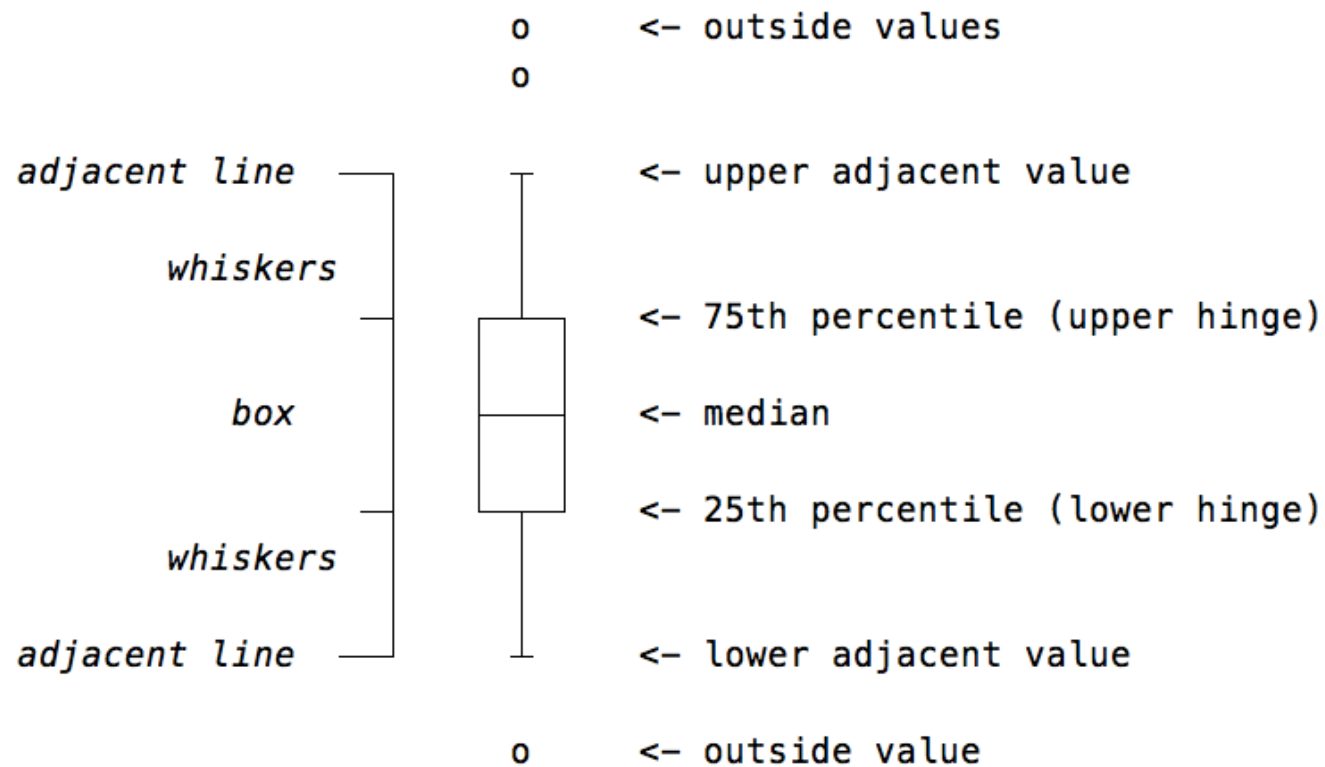
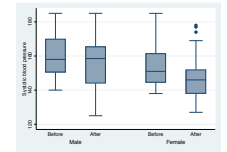


Graphics > Box plot

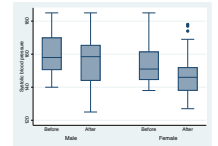


```
sysuse bplong  
graph box bp, over(when) over(sex)
```

Box plot



Box plot syntax



`graph box varname`

- draws vertical box plots. (y axis is numerical, x axis is categorical)

`graph hbox varname`

- draws horizontal box plots

options

`, nooutsides`

- don't display outliers

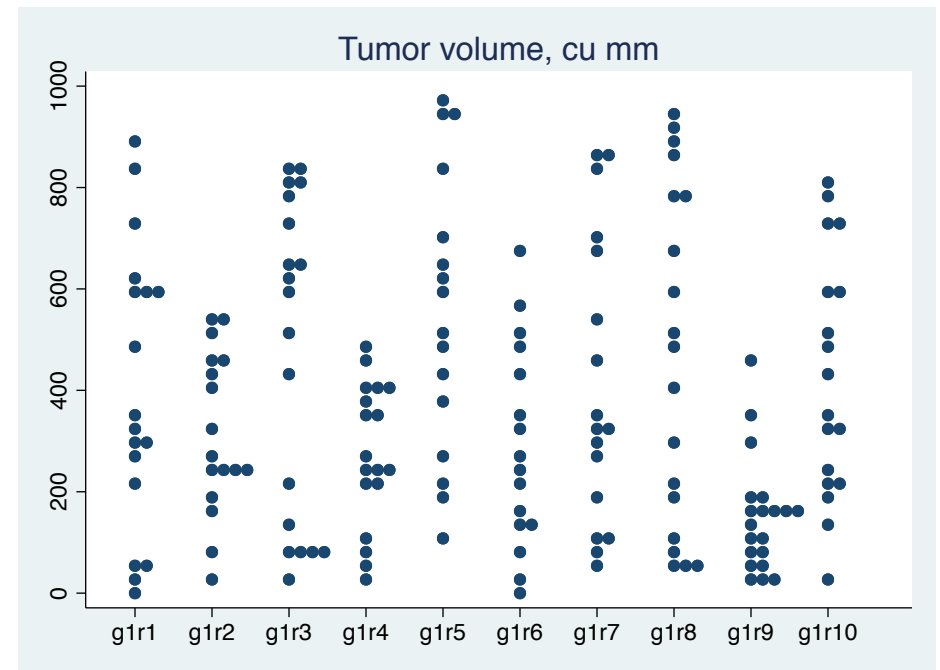
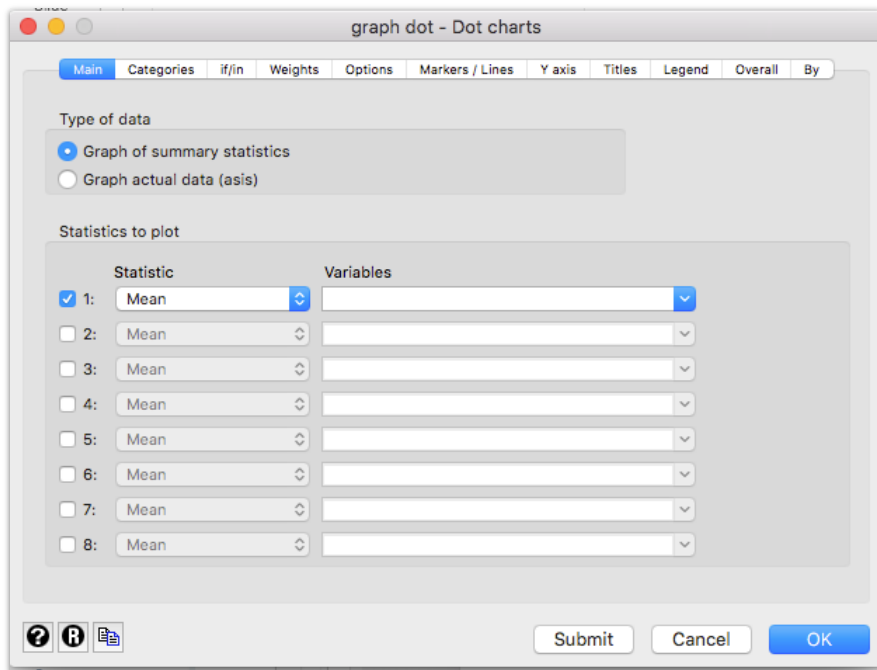
`, over(varname)`

- boxplot for each category of `varname`

`, missing`

- include missing as category of `varname`

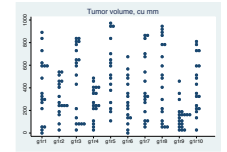
Dot plot



Graphics > Dot chart

```
webuse dotgr  
dotplot g1r1-g1r10, title("Tumor volume, cu mm")
```

Dot plot syntax



`dotplot` `varname`

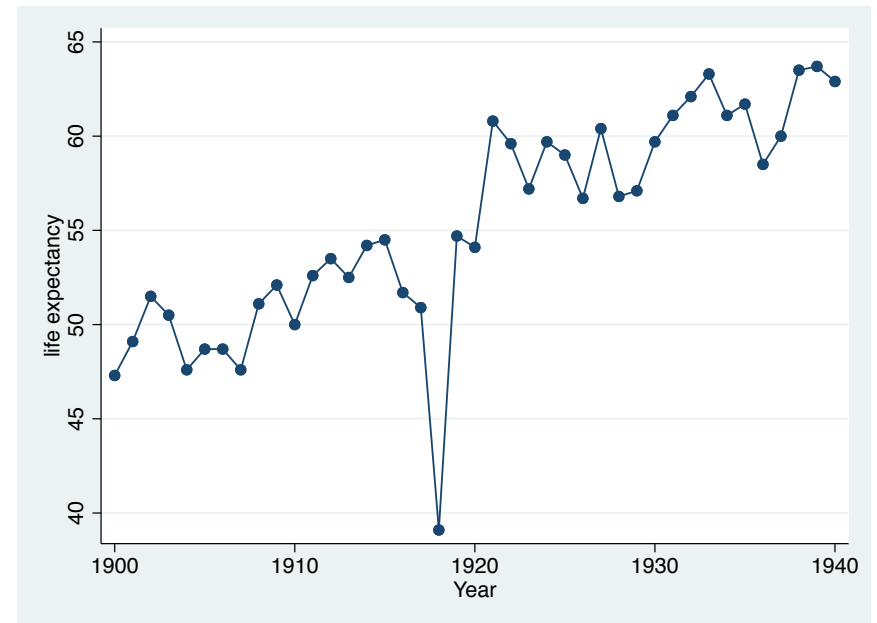
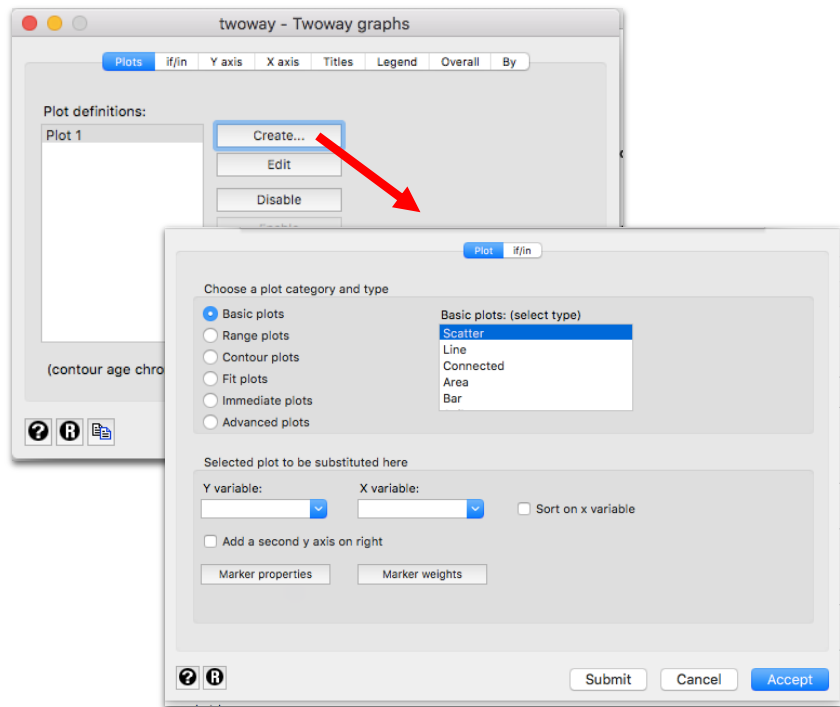
Options

- `, over (varname)` – display dotplot for each value of `groupvar`
- `, center` – center the dot for each column
- `, mean` – place horizontal line at mean
- `, median` – place horizontal line at median

Practice

- ◇ Create a box plot for health_bl comparing acupuncture to control group
- ◇ Create a box plot for health_bl comparing males and females within the acupuncture group and the control group
- ◇ Create a dot plot comparing f1, f2, and f5

Scatter plot



```
sysuse uslifeexp2  
scatter le year, connect(1)
```

Graphics > Twoway

Scatter plot syntax

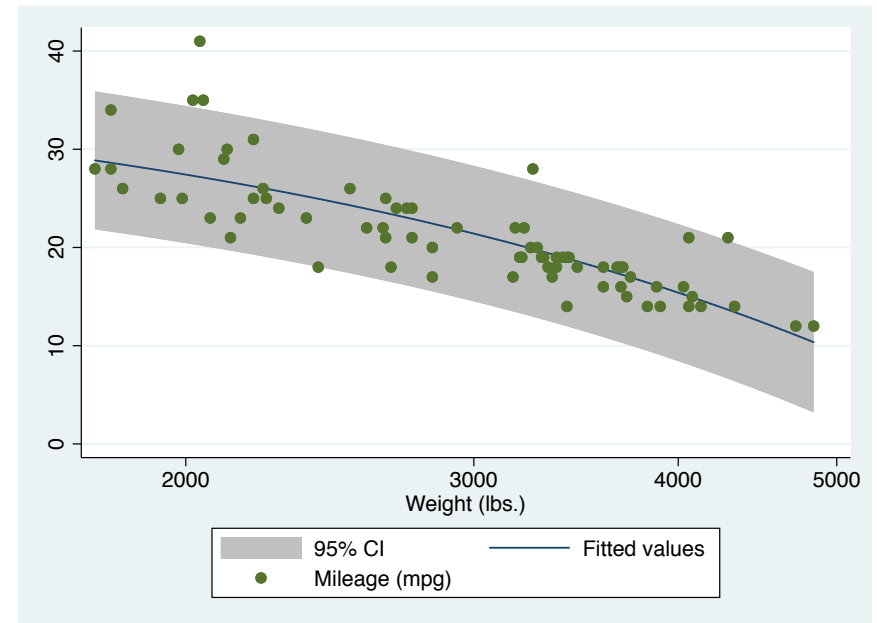
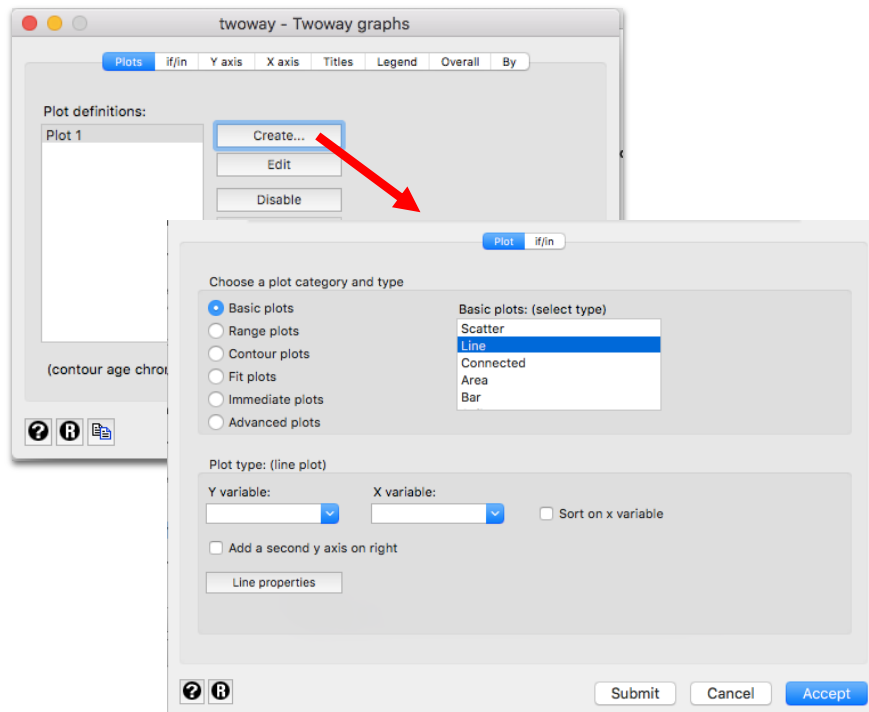
```
scatter yvar xvar
```

Options

```
, connect(style)  
, jitter(#)
```

- connect points
- offset points so they don't overlap

Linear fits



Graphics > Twoway

```
sysuse auto, clear
twoway lfitci mpg weight, stdf ||
scatter mpg weight ||
, xscale(log)
```

Linear fit syntax

twoway lfitci yvar xvar

lowess

lfit

lfitci

- lowess line plot
- linear prediction plot
- linear prediction plot with CIs

Layering graphs

twoway

```
plot1 varlist  
, options  
plot 2 varlist  
, options  
plot 3 varlist  
, options  
, general options
```

```
///  
|| ///  
///  
|| ///  
///  
|| ///
```

	after each plot
///	after each line except last

Practice

- ◇ Create a 2-way scatterplot of `health_bl` as the independent variable (x-axis) and `basehs` as the dependent variable (y-axis) with a lowess line
- ◇ Add a linear fit line to the plot above

Combining graphs: `graph combine`

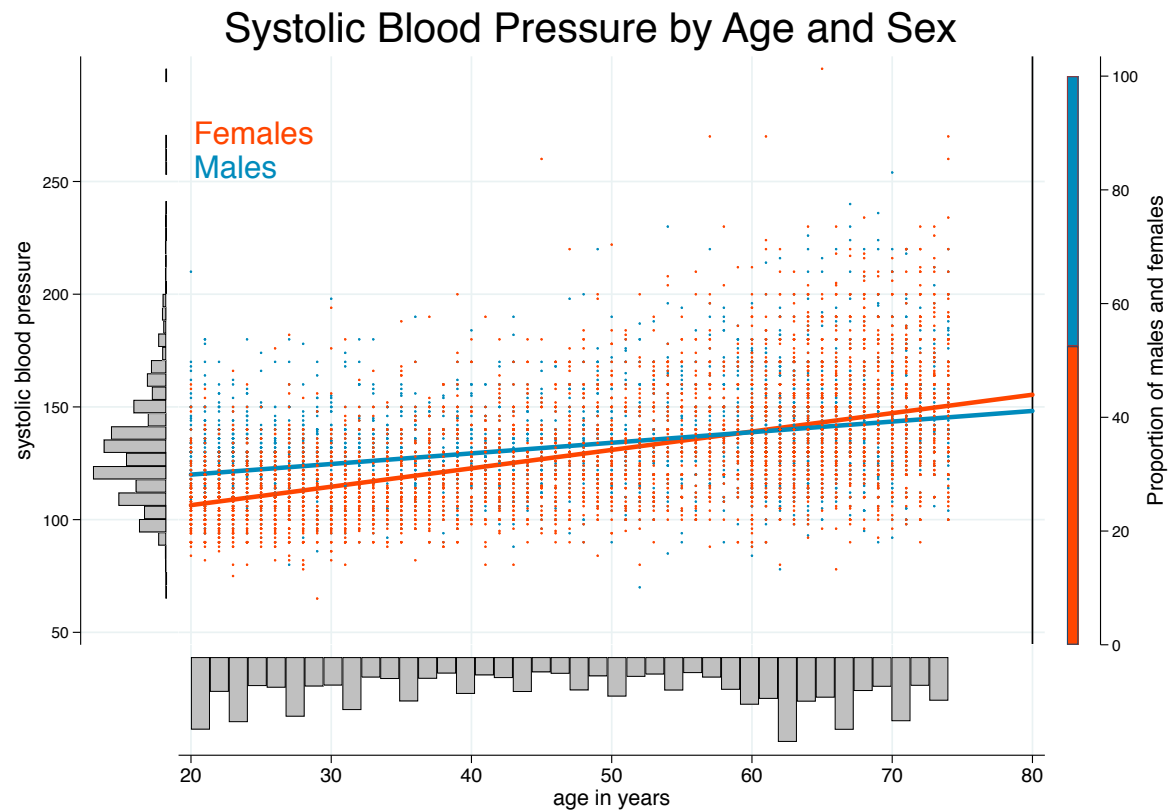
`graph combine graph1 graph2 graph3...`

options

- | | |
|--------------------------|------------------------------------|
| <code>, altshrink</code> | - reduces size of text and markers |
| <code>, cols(#)</code> | - number of columns |
| <code>, rows(#)</code> | - number of rows |

Need to use option `, name(graph1)` with previously created graphs to be able to use `combine`

Graph combine example



Saving, naming & exporting

, saving `graphname`, replace

-Save the graph as a Stata .gph file on disk

, name `graphname`

-graph is saved in memory instead of on disk. Need to clear all, graph drop `graphname`, or quit Stata to clear from memory

graph export `graphname`, as(`extension*`)

-keep graph window open to export!

*extension options: .pdf, .jpg, .png, .tif, and many more
See “help graph export” for all the options

A note on sharing .gph files

- ◇ Stata figures in the native format (.gph) include the data used to create the figure
- ◇ Be careful sharing .gph files if your data set has restrictions/sensitive information



Figures using putdocx

putdocx

- ◇ Allows you to create a Word document in memory
- ◇ Embed figures and tables directly from Stata
- ◇ Can add (and format) text but syntax is clunky

- ◇ More info:
 - <https://www.stata.com/features/overview/create-word-documents/>
 - <https://www.stata.com/manuals/rptputdocxintro.pdf>

Export figures to word with putdocx

1. Create figure as usual
2. `graph export "graphname.png", replace`
3. `putdocx clear`
4. `putdocx begin`
5. `putdocx paragraph, halign(center)`
6. `putdocx image "graphname.png"`
7. `putdocx save "filename" [, replace append]`

Practice

- ✧ Export a graphic as a png file
- ✧ Put the image in a Word document using `putdocx`

Formatting graphics

Title, caption & note options

title options

`title(xyz)`

`subtitle(xyz)`

`caption(xyz)`

`note(xyz)`

- overall title
- subtitle of title
- explanation of plot
- add a note on the plot

Axis and legend options

Axis options

`ytitle(xyz)`

- y axis title

`xtitle(xyz)`

- x axis title

`ttitle(xyz)`

- time axis title

Legend options

`label(xyz)`

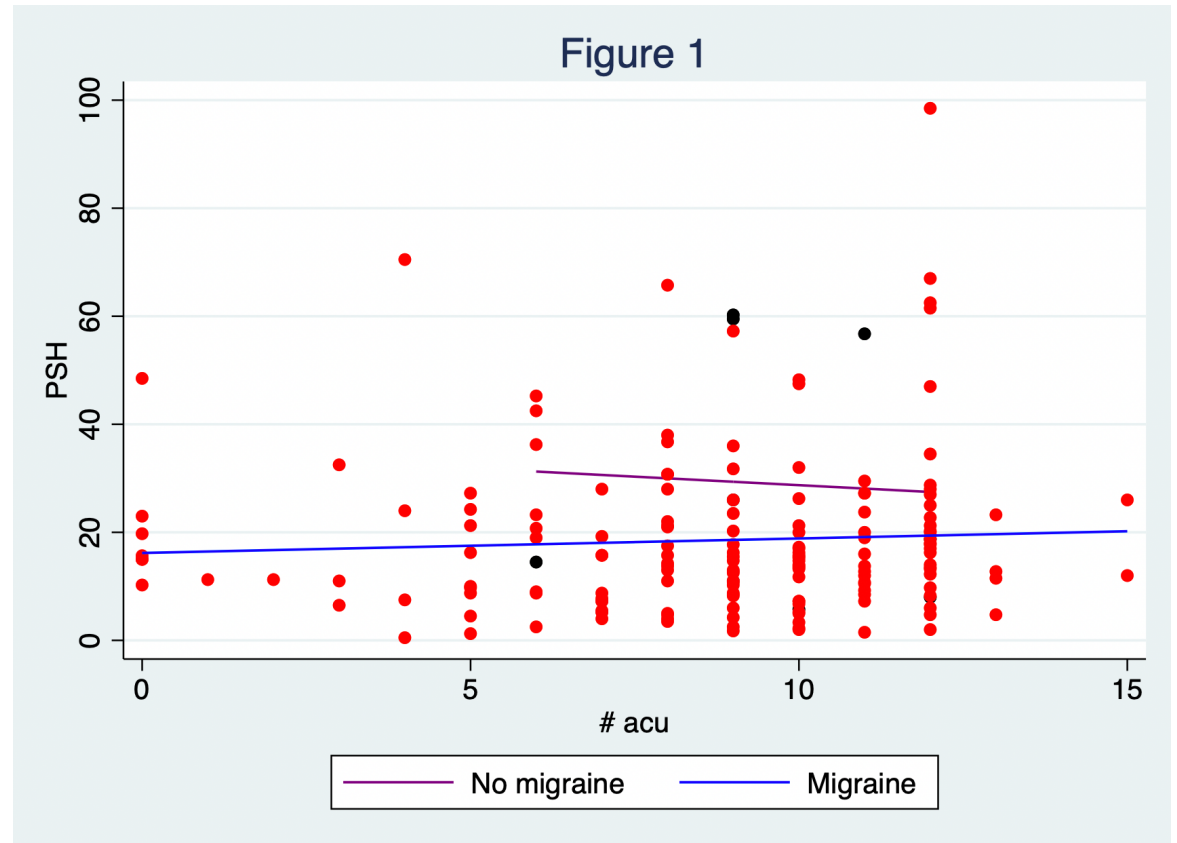
- custom label

`legend(off)`

- turn legend off

Practice: Figure 1

- Run the code for 'Figure 1'



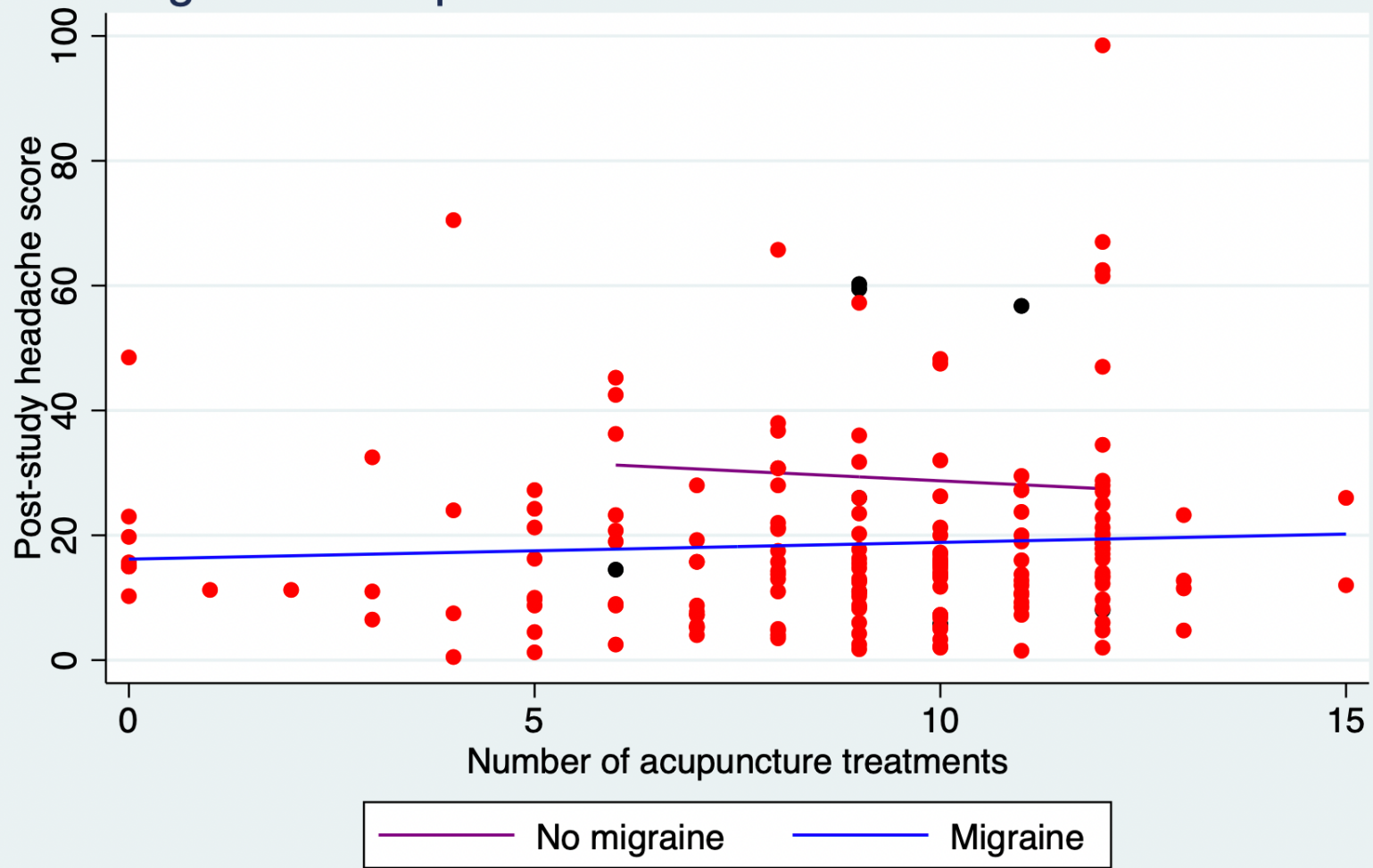
Practice: Figure 1

- ◇ Change the title to “Figure 1: Acupuncture treatments and headache score”
- ◇ Add a note that says “Lecture 5 practice”
- ◇ Change the y-axis to “Post-study headache score”
- ◇ Change the x-axis to “Number of acupuncture treatments”
- ◇ Run your commands for Figure 1 to see if they worked

Practice: Figure 1

- ◇ Change the title to “Figure 1: Acupuncture treatments and headache score”
`title(Figure 1: Acupuncture treatments and headache score)`
- ◇ Add a note that says “Lecture 5 practice”
`note(Lecture 5 practice)`
- ◇ Change the y-axis to “Post-study headache score”
`ytitle(Post-study headache score)`
- ◇ Change the x-axis to “Number of acupuncture treatments”
`xtitle(Number of acupuncture treatments)`

Figure 1 : Acupuncture treatments and headache score



Lecture 5 practice

Text size options

, size(size)

28 pt.	vhuge	10 pt.	medsmall
20 pt.	huge	8 pt.	small
16 pt.	vlarge	6 pt.	vsmall
14 pt.	large	4 pt.	tiny
12 pt.	medlarge	2 pt.	half_tiny
11 pt.	medium	1.3 pt.	third_tiny
		1 pt.	quarter_tiny
		1 pt.	minuscule

Color options

 black	 gs0	 gs1	 gs2	 gs3	 gs4
 gs5	 gs6	 gs7	 gs8	 gs9	 gs10
 gs11	 gs12	 gs13	 gs14	 gs15	 gs16
 white	 blue	 bluishgray	 brown	 cranberry	 cyan
 dimgray	 dkgreen	 dknavy	 dkorange	 eggshell	 emerald
 forest_green	 gold	 gray	 green	 khaki	 lavender
 lime	 ltblue	 ltbluishgray	 ltkhaki	 magenta	 maroon
 midblue	 midgreen	 mint	 navy	 olive	 olive_teal
 orange	 orange_red	 pink	 purple	 red	 sand
 sandb	 sienna	 stone	 teal	 yellow	 ebg
 ebbblue	 edkblue	 eltblue	 eltgreen	 emidblue	 erose

Marker color

`mcolor(blue)`

Marker label color

`mlcolor(blue)`

Line color

`lcolor(blue)`

Text color

`color(blue)`

Gridline color

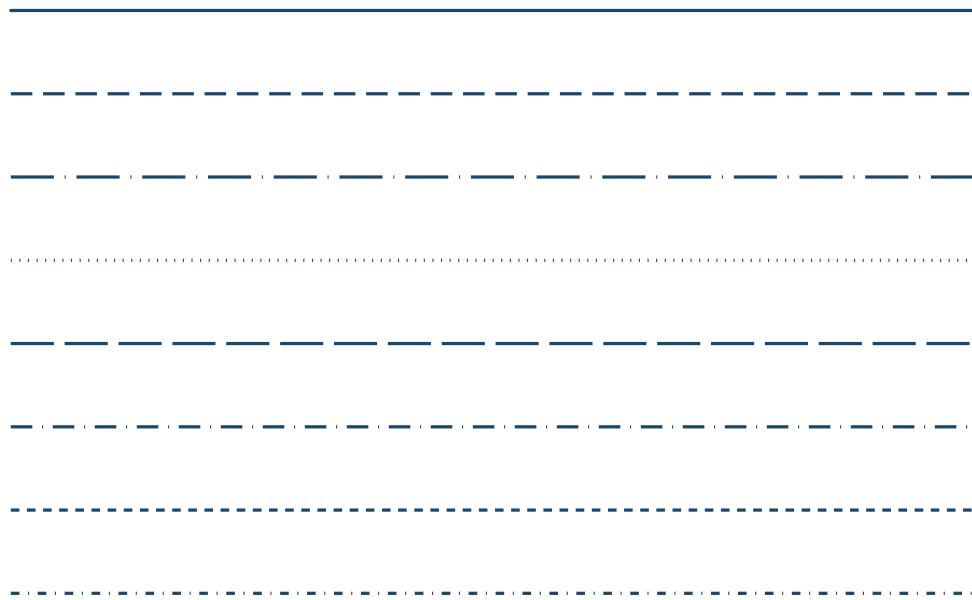
`glcolor(blue)`

Axis label color

`labcolor(blue)`

Line pattern options

, `lpattern(pattern)`



solid

dash

longdash_dot

dot

longdash

dash_dot

shortdash

shortdash_dot

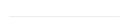
blank

, `glpattern(pattern)`

-grid line pattern

Line width options

, lwidth(width)

	vvvthick		medthin
	vvthick		thin
	vthick		vthin
	thick		vvthin
	medthick		vvvthin
	medium		none

, glwidth(width)

– Grid line width

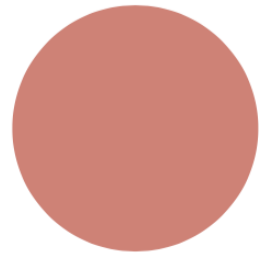
Marker symbol options

, msymbol(symbol)

 O	 D	 T	 S
 o	 d	 t	 s
 Oh	 Dh	 Th	 Sh
 oh	 dh	 th	 sh
 +	 X	 p	none i

Marker size options

, msize(**size**)



ehuge



vhuge



huge



vlarge



large

● medlarge

● medium

● medsmall

● small

● vsmall

● tiny

● vtiny

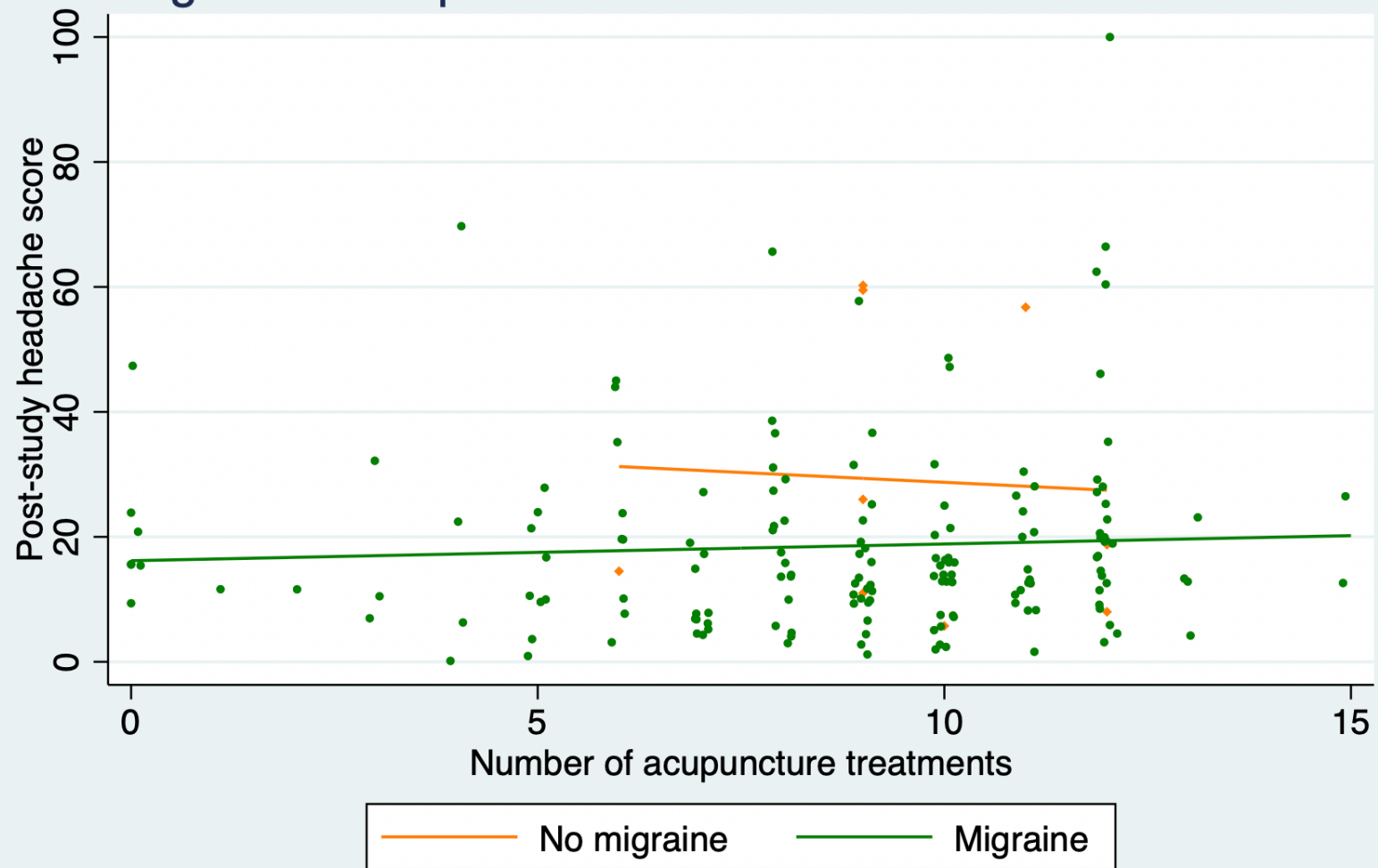
Practice: Figure 1

- ◇ Change the marker & line color to orange for the no migraine group
- ◇ Change the marker & line color to green for the migraine group
- ◇ Change the marker size to tiny (both no migraine and migraine group)
- ◇ Jitter the points in the migraine group so they don't overlap
- ◇ Change the marker symbol for the no migraine group to a diamond
- ◇ Run your commands for Figure 1 to see if they worked

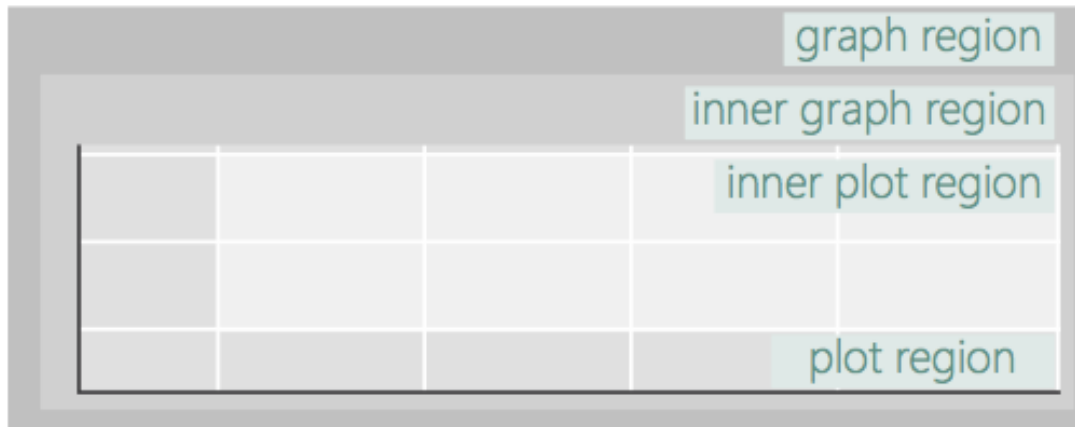
Practice: Figure 1

- ◇ Change the marker & line color to orange for the no migraine group
 - `mcolor(orange)`
 - `lcolor(orange)`
- ◇ Change the marker & line color to green for the migraine group
 - `mcolor(green)`
 - `lcolor(green)`
- ◇ Change the marker size to tiny (both no migraine and migraine group)
 - `msize(tiny)`
- ◇ Jitter the points in the migraine group so they don't overlap
 - `jitter(2)`
- ◇ Change the marker symbol for the no migraine group to a diamond
 - `msymbol(d)`

Figure 1 : Acupuncture treatments and headache score



Region options



Graph region color

`graphregion(color(blue))`

Plot region color

`plotregion(color(blue))`

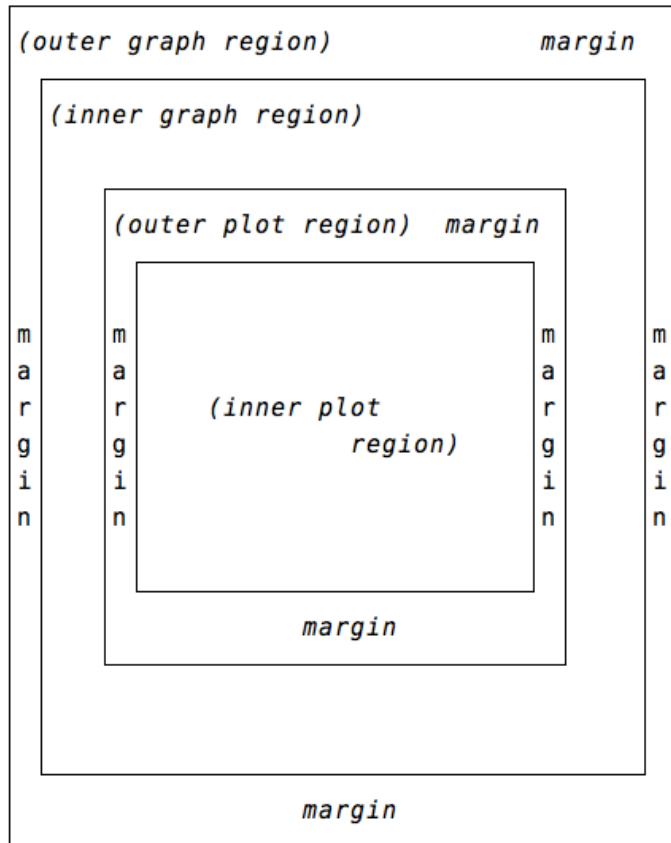
Graph region outline color

`graphregion(lcolor(blue))`

Plot region outline color

`plotregion(lcolor(blue))`

Margin options



Title margin

`title("my title", margin(size))`

Graph region margin

`graphregion(margin(size))`

Size options

zero
tiny
vsmall
small
medsmall
medium
medlarge
large
vlarge

Schemes

, scheme (scheme)

Stata scheme options

s2color

s2mono

s2gcolor

s2manual

s1rcolor

s1color

s1mono

s1manual

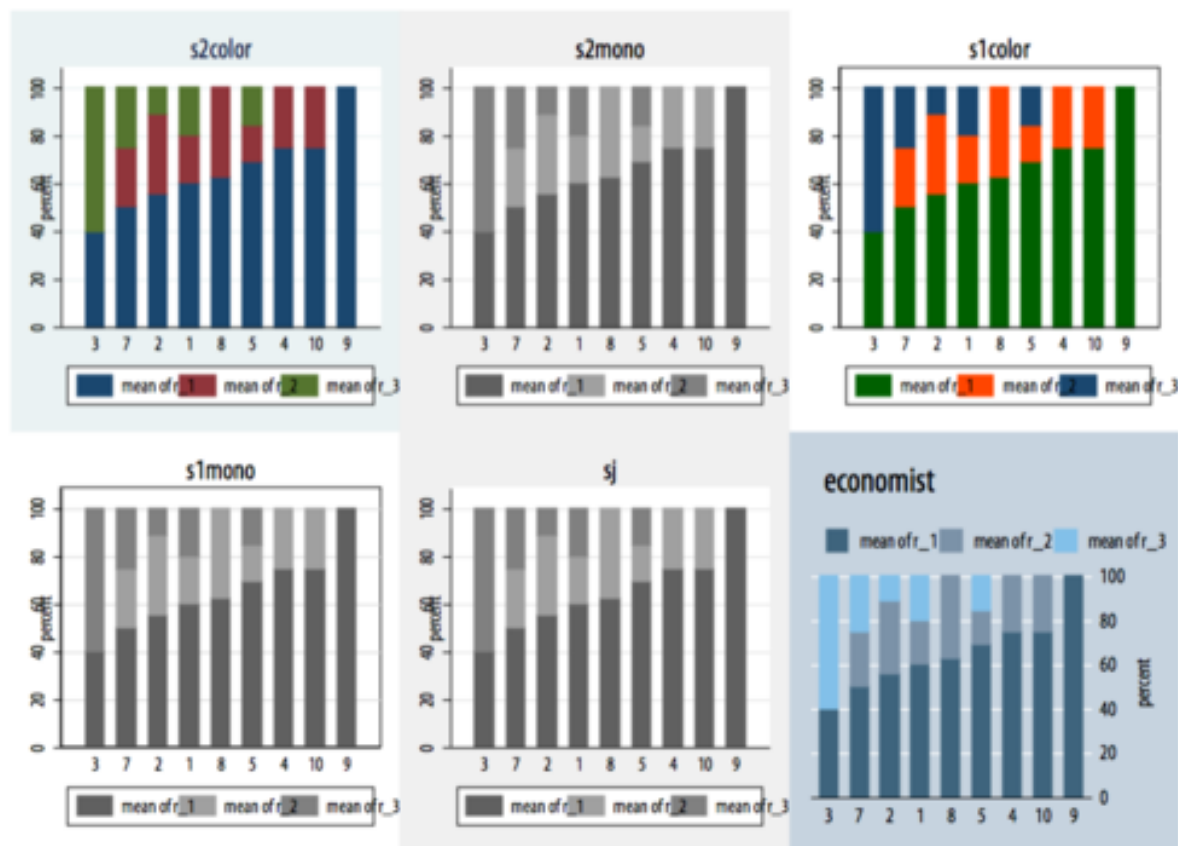
economist

sj

- factory setting
- s2color in monochrome
- color white used in the Stata manuals
- s2gcolor in monochrome
- a plain look on black background
- a plain look
- a plain look in monochrome
- s1mono but smaller
- The Economist magazine
- Stata Journal

Plus many user-contributed options!

Stata schemes



Scheme options

set scheme `schemename`

– set your default scheme (in .do file)

set scheme `schemename`, permanently

– set your default scheme permanently

Or:

[preference > graph preferences]

**you can also select default text here*

Scheme resources

- ◇ Stata journal article with examples of different schemes
 - <https://blog.stata.com/2018/10/02/scheming-your-way-to-your-favorite-graph-style/>
- ◇ Examples of user-generated schemes as well as new user-written packages to simplify creating your own schemes
 - https://www.stata.com/meeting/uk18/slides/uk18_Jann.pdf

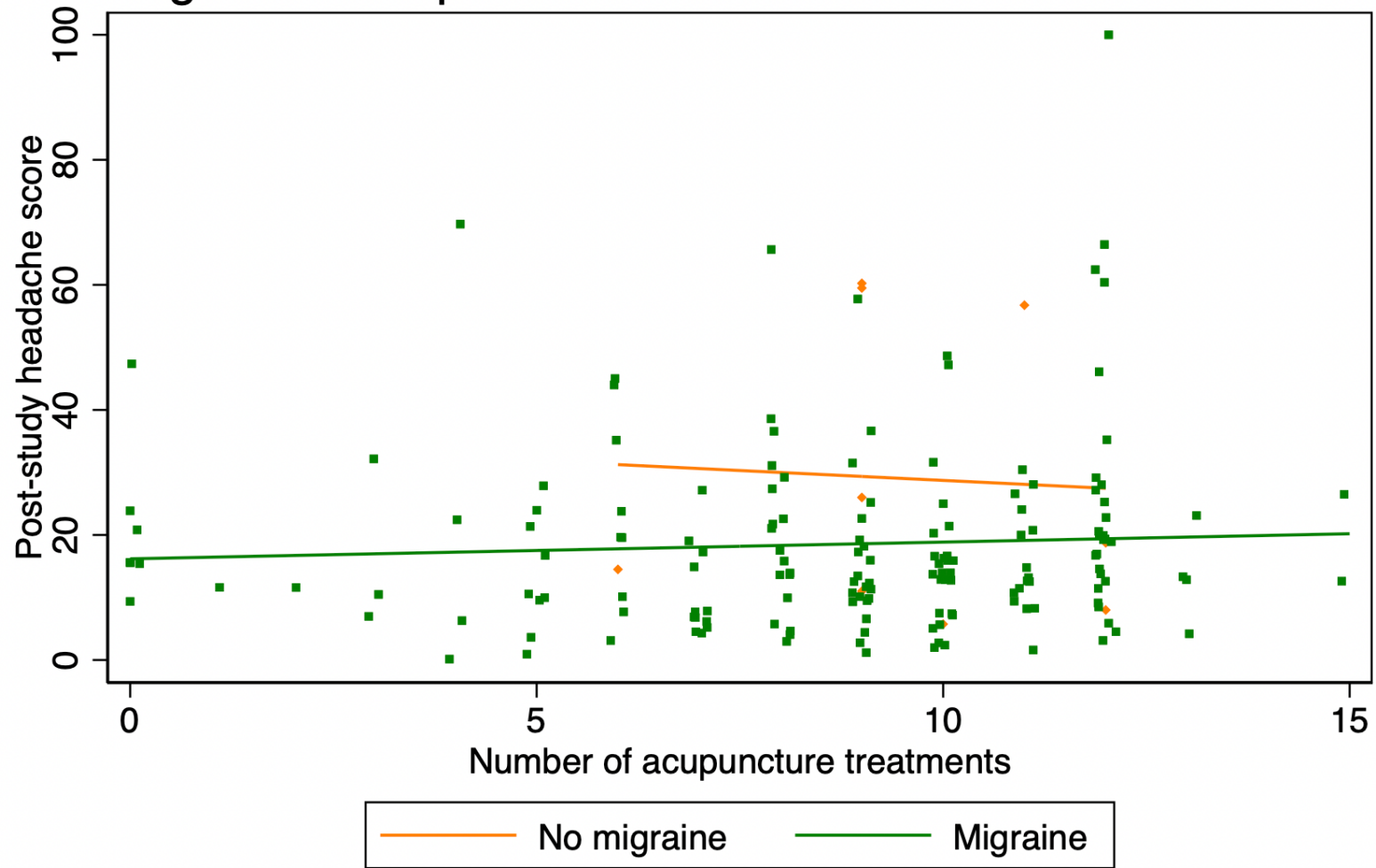
Practice: Figure 1

- ◇ Change the scheme to `s1mono`
- ◇ Run the command for Figure 1 to see if it worked

Practice: Figure 1

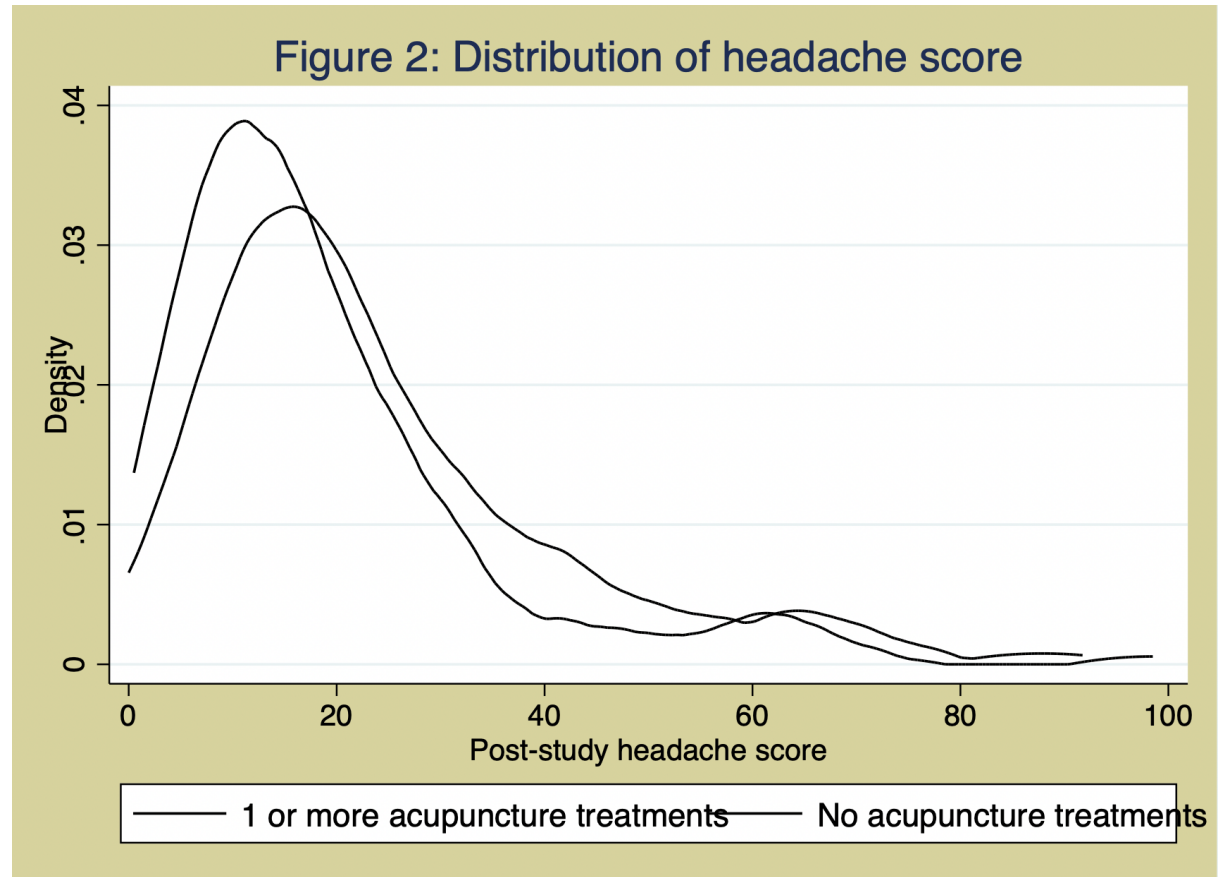
- ◇ Change the scheme to `s1mono`
 - `scheme (s1mono)`
- ◇ Run the command for Figure 1 to see if it worked

Figure 1 : Acupuncture treatments and headache score



Practice: Figure 2

- Run the code for 'Figure 2'



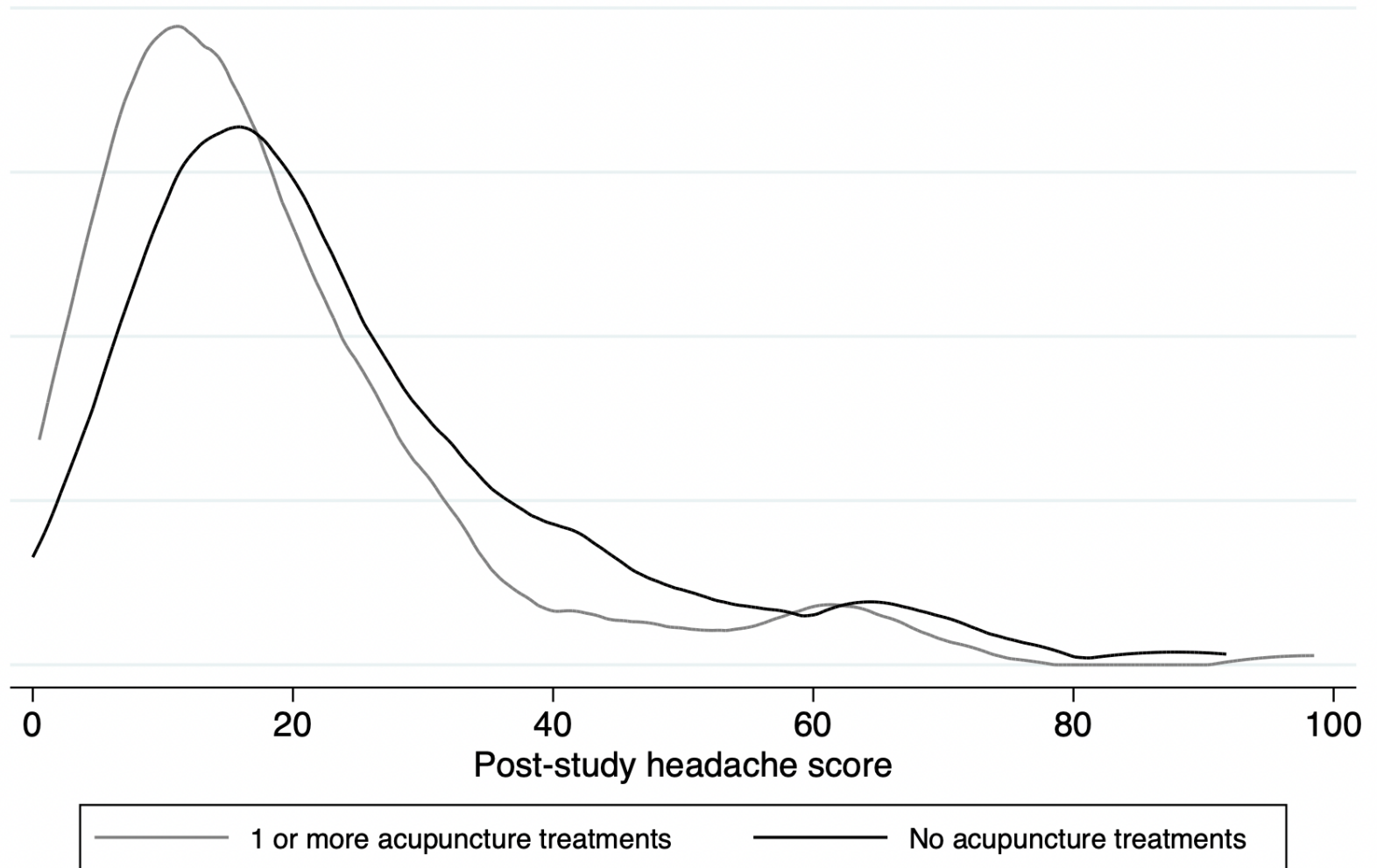
Practice: Figure 2

- ◇ Change the line color for the no acupuncture group to gray
- ◇ Turn off the y axis
- ◇ Change the color of the graph region to white
- ◇ Change the legend font size to small
- ◇ Run your commands for Figure 2 to see if they worked

Practice: Figure 2

- ◇ Change the line color for the no acupuncture group to gray
 - `lcolor(gray)`
- ◇ Turn off the y axis
 - `yscale(off)`
- ◇ Change the color of the graph region to white
 - `graphregion(fcolor(white))`
- ◇ Change the legend font size to small
 - `size(small)`

Figure 2: Distribution of headache score



Resources

◇ Data visualization principles

- <https://rafalab.github.io/dsbook/data-visualization-principles.html>

◇ Stata cheat sheets

- https://geocenter.github.io/StataTraining/portfolio/01_resource/

◇ STATA GRAPHICS REFERENCE MANUAL

- <https://www.stata.com/bookstore/graphics-reference-manual/>

◇ Visual overview for creating graphs

- <https://www.stata.com/support/faqs/graphics/gph/stata-graphs/>

Next Week

Tables