

Lecture 4: Visualizing data

BIOSTAT 212, Summer 2018



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

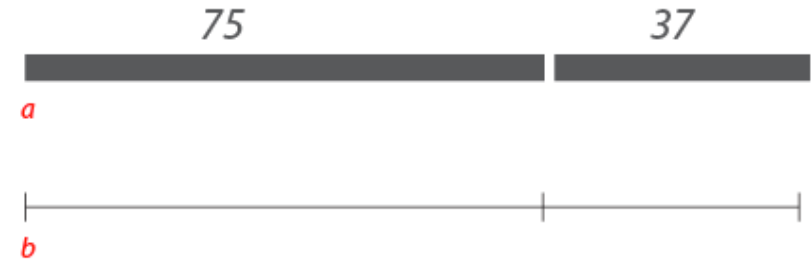
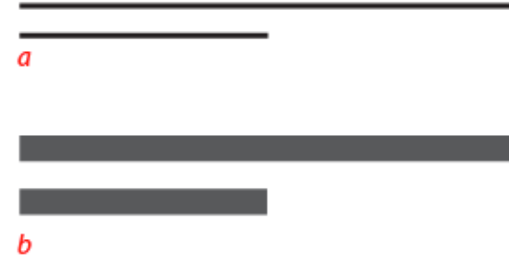
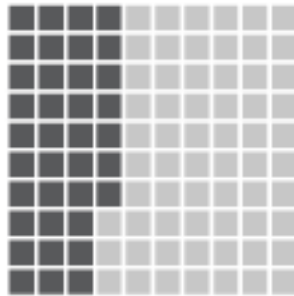
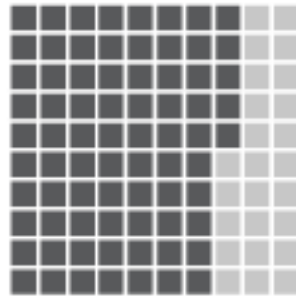
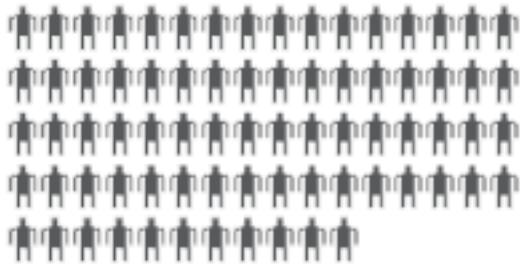
Warm up

Work with your neighbor to sketch out:

How many ways can you represent....

75 and 37

75 and 37



Good visualization allows you to:

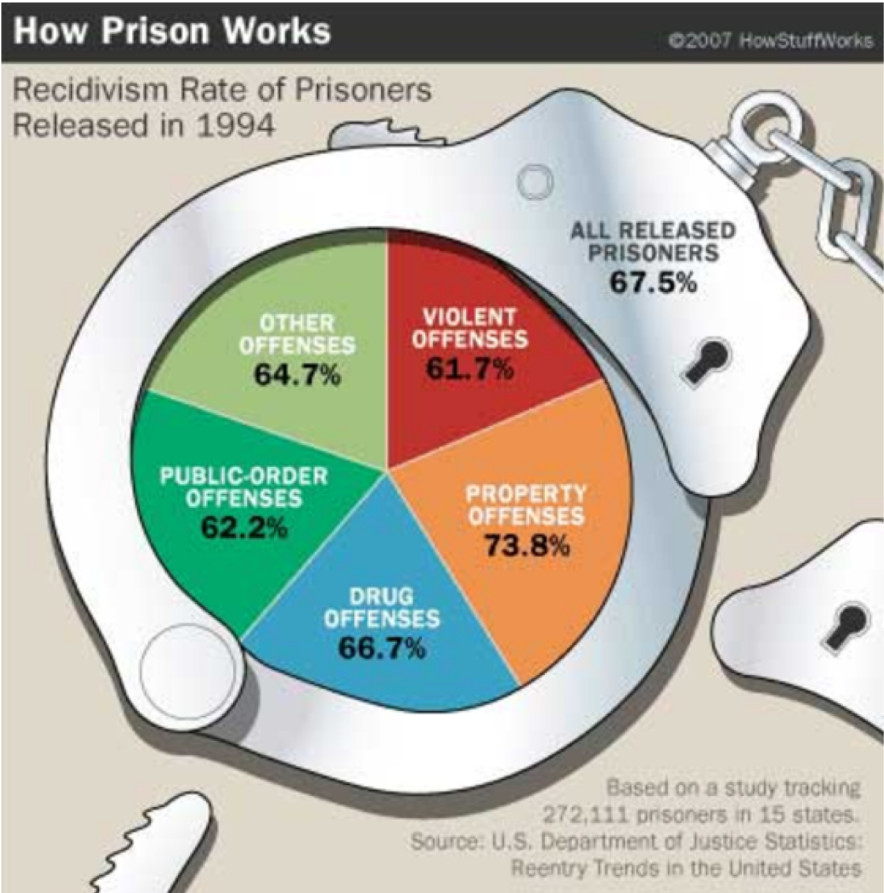
Explore  Analyze  Communicate

- ◇ See information at a glance
- ◇ See new relationships
- ◇ Make comparisons
- ◇ Easily communicate your findings

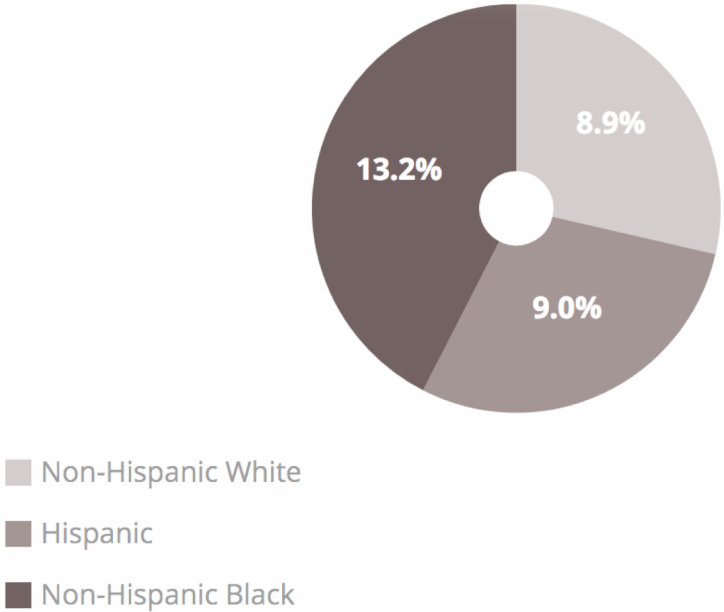
Visualizing data: rules of thumb

- ◇ Think of your audience
- ◇ What is the message you want to convey?
- ◇ Keep it simple
- ◇ Usually your goal is to make some kind of comparison
- ◇ Color blind friendly
- ◇ Limited space in publications → usually pie charts and bar graphs take up less room when written as a sentence

Avoid: percentages that don't add up



PRETERM BIRTH BY RACE & ETHNICITY



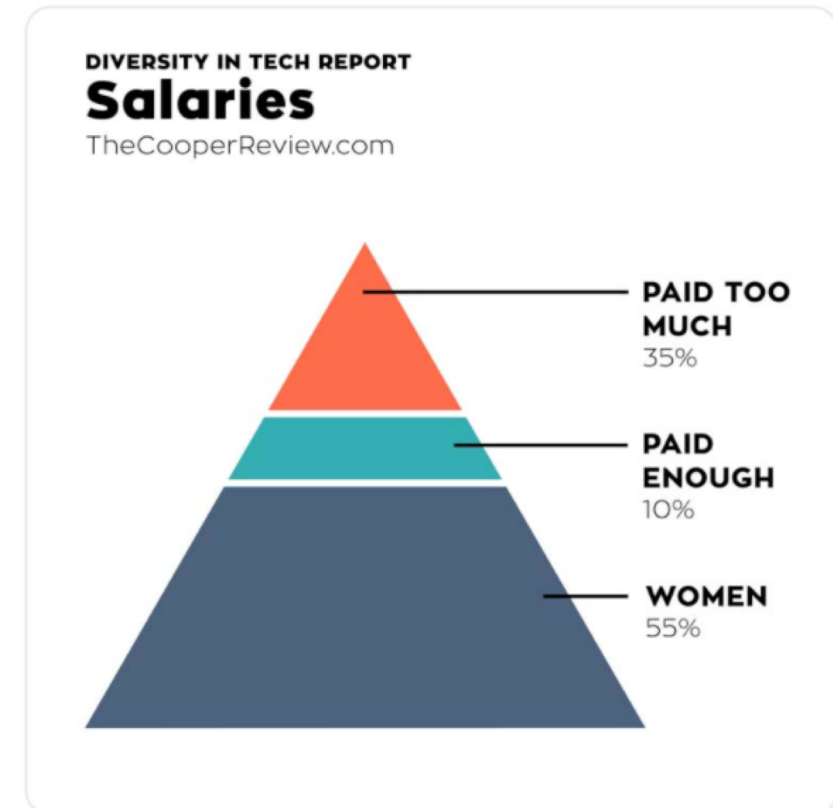
Preterm defined as less than 37 weeks gestation.
Source: CDC, 2015

Avoid: comparisons that don't make sense

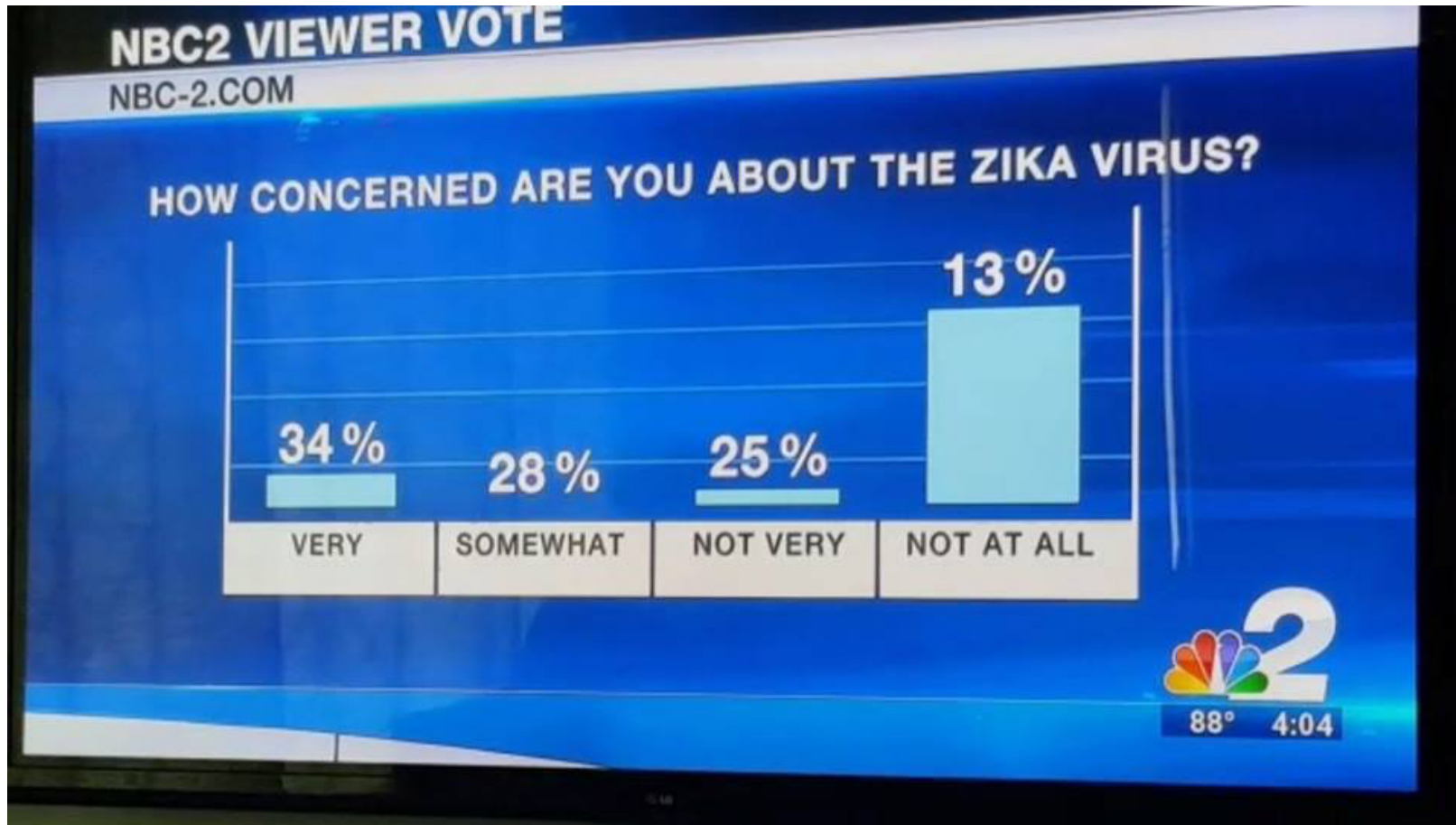


7. Salaries

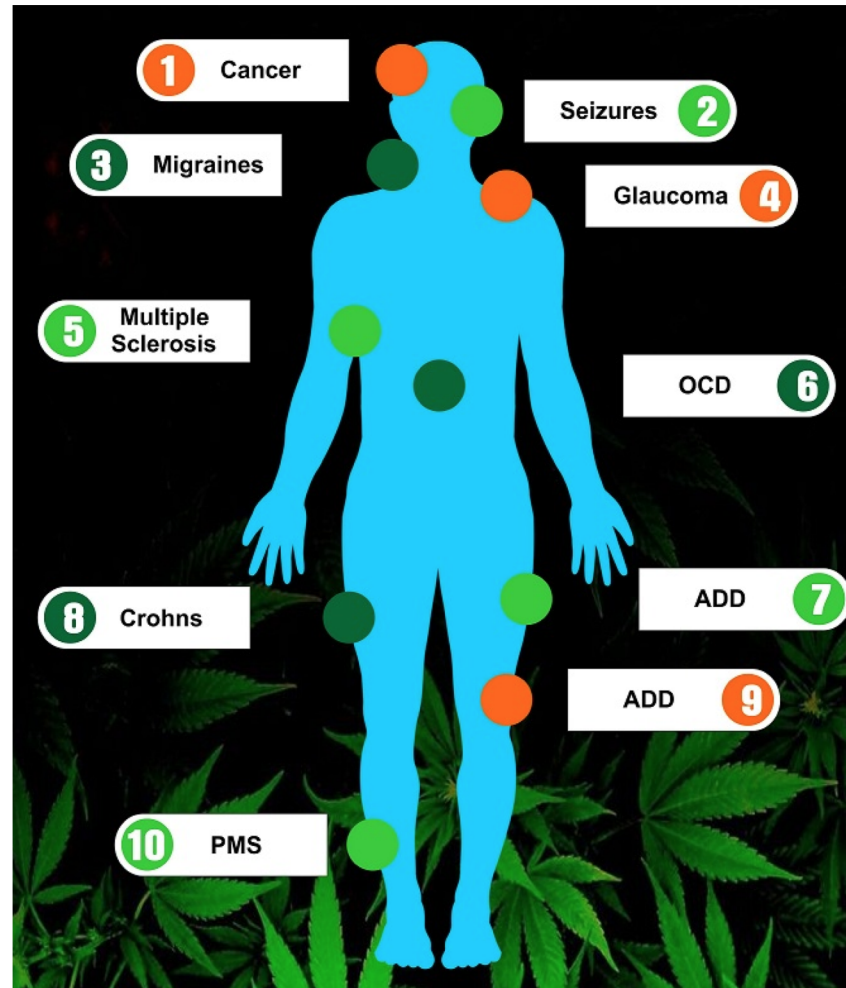
Salaries are also diverse.



Avoid: sizes that don't match percentages



Glaucoma of the shoulder? PMS of the shin?

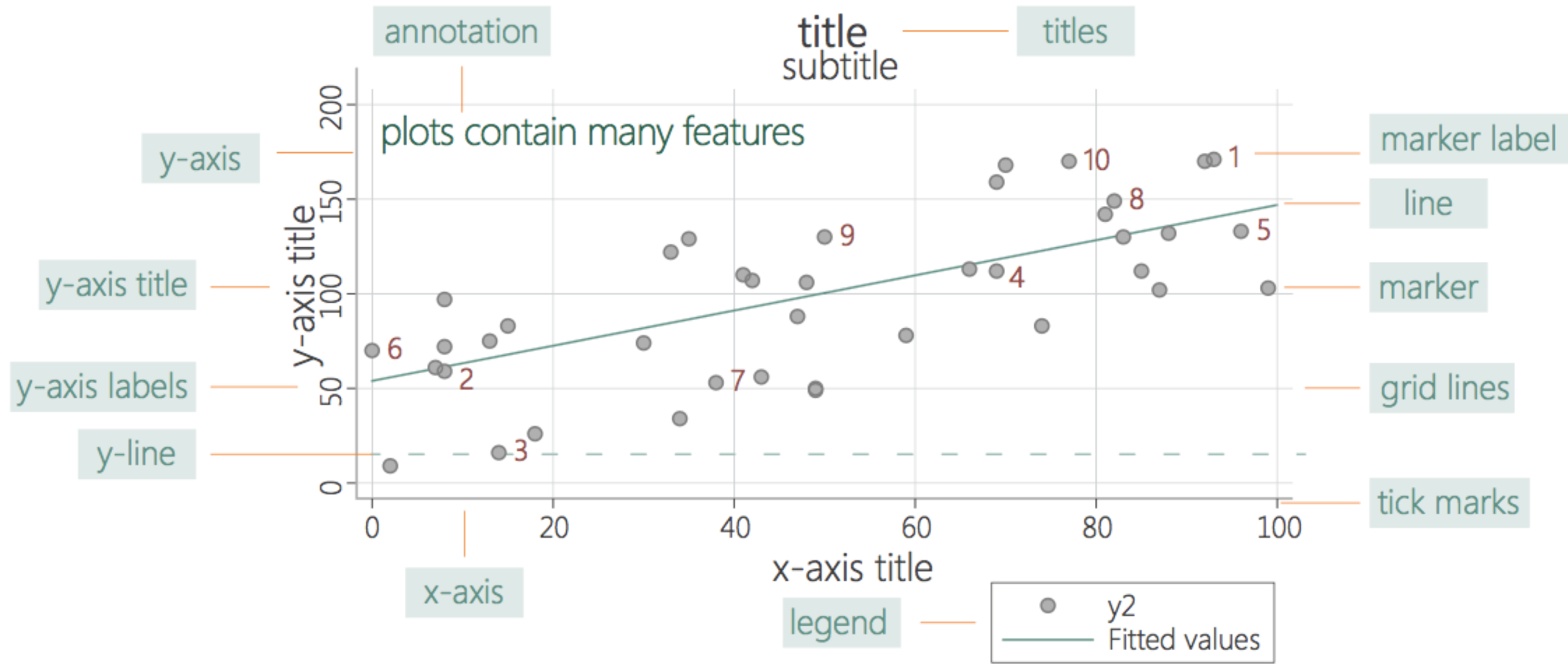


Outline

- ◇ Anatomy of a plot
- ◇ Help files
- ◇ Histogram
- ◇ Bar chart
- ◇ Box plot
- ◇ Dot plot
- ◇ Coefficient plot
- ◇ Scatter
 - Linear fit

- ◇ Layering graphs
- ◇ Combining graphs
- ◇ Saving and naming
- ◇ Figure options
 - *Titles, captions, notes*
 - *Axis titles and legends*
 - *Colors*
 - *Region, margins*
 - *Line patterns*
 - *Marker symbols*
 - *Marker size*
 - *Text size*
 - *Line width*
 - *Schemes*

ANATOMY OF A PLOT



Outline

- ◇ Anatomy of a plot

- ◇ **Help files**

- ◇ Histogram

- ◇ Bar chart

- ◇ Box plot

- ◇ Dot plot

- ◇ Coefficient plot

- ◇ Scatter

- Linear fit

- ◇ Layering graphs

- ◇ Combining graphs

- ◇ Saving and naming

- ◇ Figure options

- *Titles, captions, notes*
 - *Axis titles and legends*
 - *Colors*
 - *Region, margins*
 - *Line patterns*
 - *Marker symbols*
 - *Marker size*
 - *Text size*
 - *Line width*
 - *Schemes*

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)
```

Outline

- ◇ Anatomy of a plot
- ◇ Help files

- ◇ **Histogram**

- ◇ Bar chart

- ◇ Box plot

- ◇ Dot plot

- ◇ Scatter

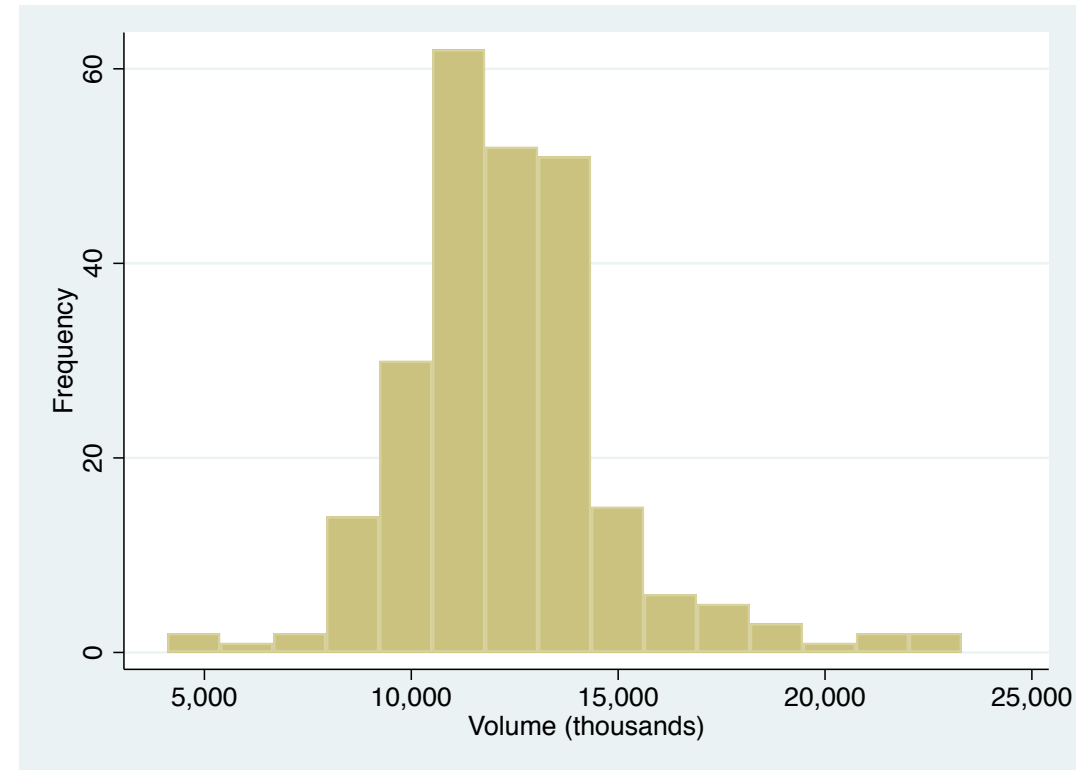
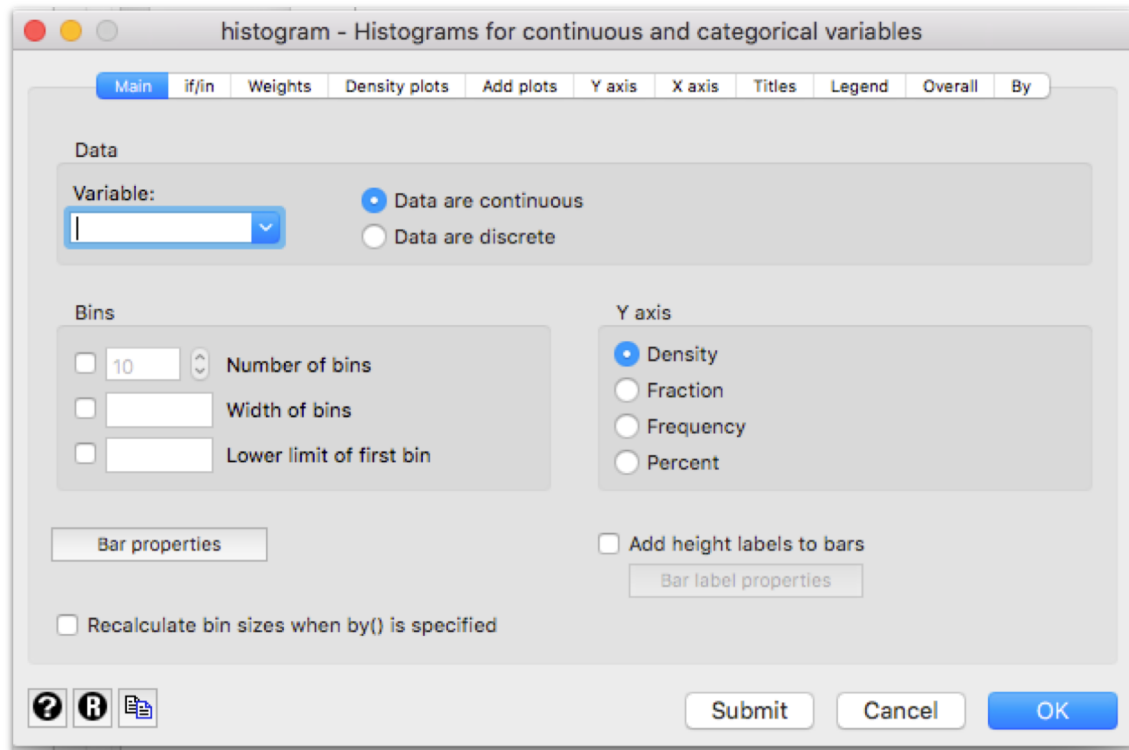
- Linear fit

- ◇ Layering graphs
- ◇ Combining graphs
- ◇ Saving and naming

- ◇ Figure options

- *Titles, captions, notes*
 - *Axis titles and legends*
 - *Colors*
 - *Region, margins*
 - *Line patterns*
 - *Marker symbols*
 - *Marker size*
 - *Text size*
 - *Line width*
 - *Schemes*

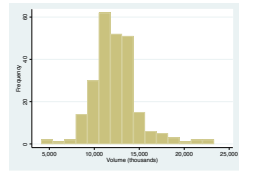
Histogram



Graphics < Histogram

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

Histogram syntax



histogram `varname`

Options

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

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

Outline

- ◇ Anatomy of a plot
- ◇ Help files
- ◇ Histogram
- ◇ **Bar chart**
- ◇ Box plot
- ◇ Dot plot
- ◇ Coefficient plot
- ◇ Scatter
 - Linear fit

- ◇ Layering graphs
- ◇ Combining graphs
- ◇ Saving and naming
- ◇ Figure options
 - *Titles, captions, notes*
 - *Axis titles and legends*
 - *Colors*
 - *Region, margins*
 - *Line patterns*
 - *Marker symbols*
 - *Marker size*
 - *Text size*
 - *Line width*
 - *Schemes*

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)

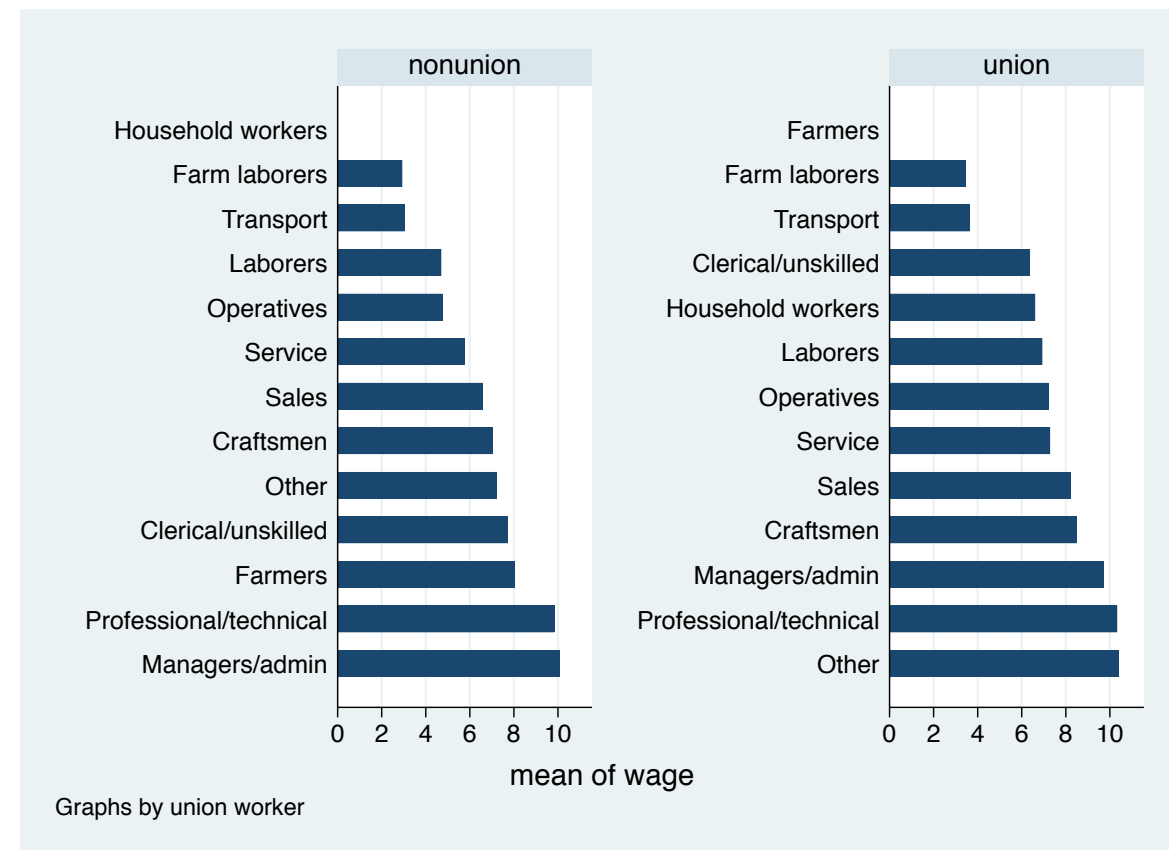
Orientation

- ☒ Vertical
- ☐ Horizontal

Statistics to plot

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

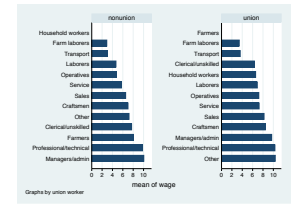
Submit Cancel OK



Graphics < Bar chart

```
sysuse nlsw88, clear  
graph hbar wage, over(occ, sort(1)) by(union)
```

Bar chart syntax



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

```
graph hbar (stat) numeric_var, [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

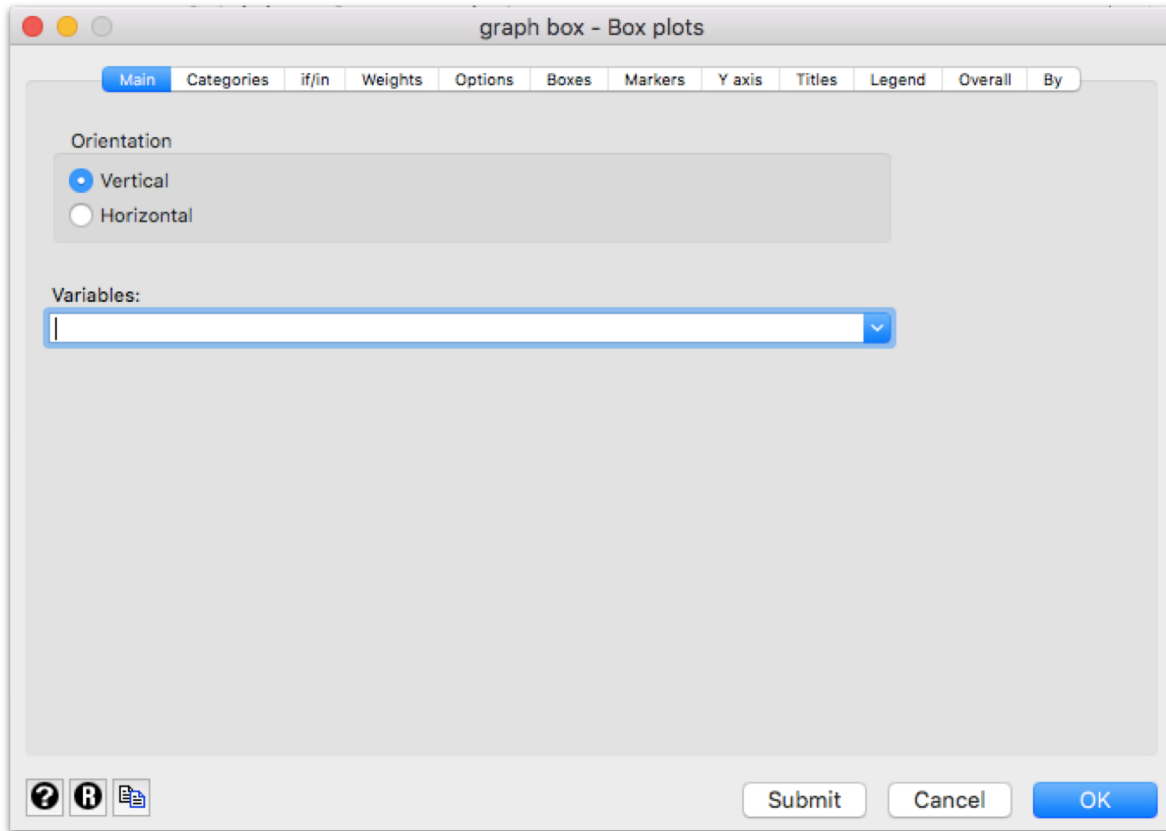
- stack bars

Outline

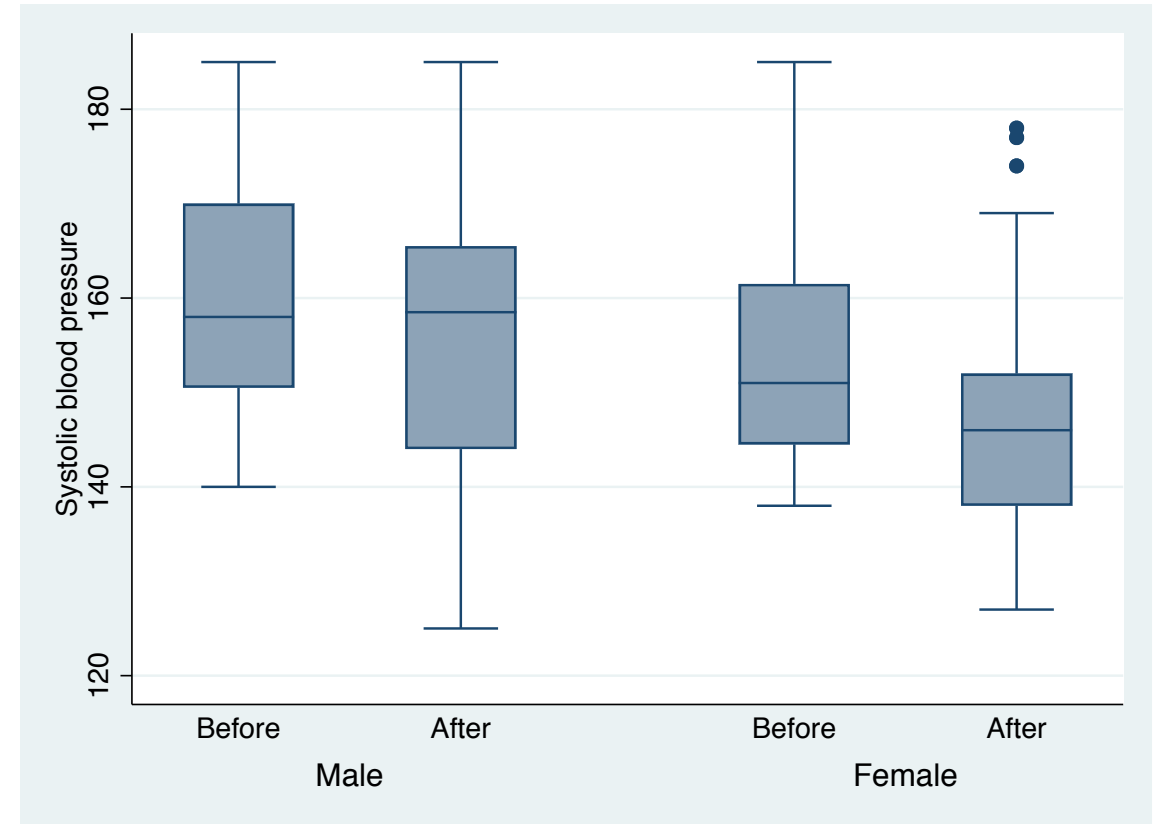
- ◇ Anatomy of a plot
- ◇ Help files
- ◇ Histogram
- ◇ Bar chart
- ◇ **Box plot**
- ◇ Dot plot
- ◇ Coefficient plot
- ◇ Scatter
 - Linear fit

- ◇ Layering graphs
- ◇ Combining graphs
- ◇ Saving and naming
- ◇ Figure options
 - *Titles, captions, notes*
 - *Axis titles and legends*
 - *Colors*
 - *Region, margins*
 - *Line patterns*
 - *Marker symbols*
 - *Marker size*
 - *Text size*
 - *Line width*
 - *Schemes*

Box plot

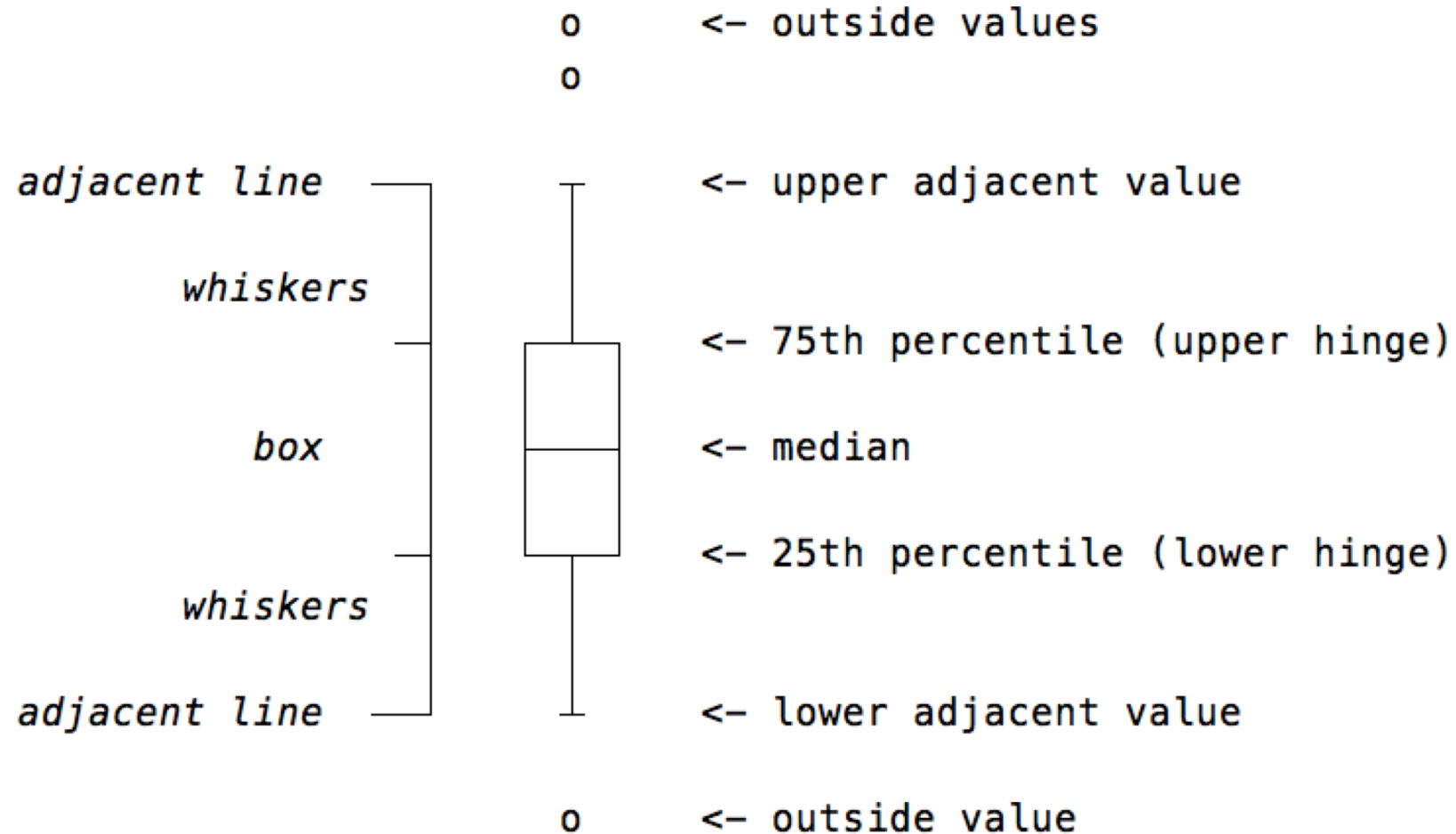
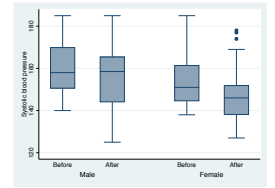


Graphics < Box plot

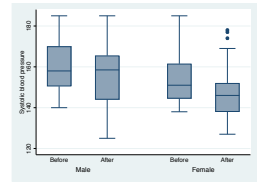


```
sysuse bplong  
graph box bp, over(when) over(sex)
```

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`

`if`

- subset data using logic operators

`in`

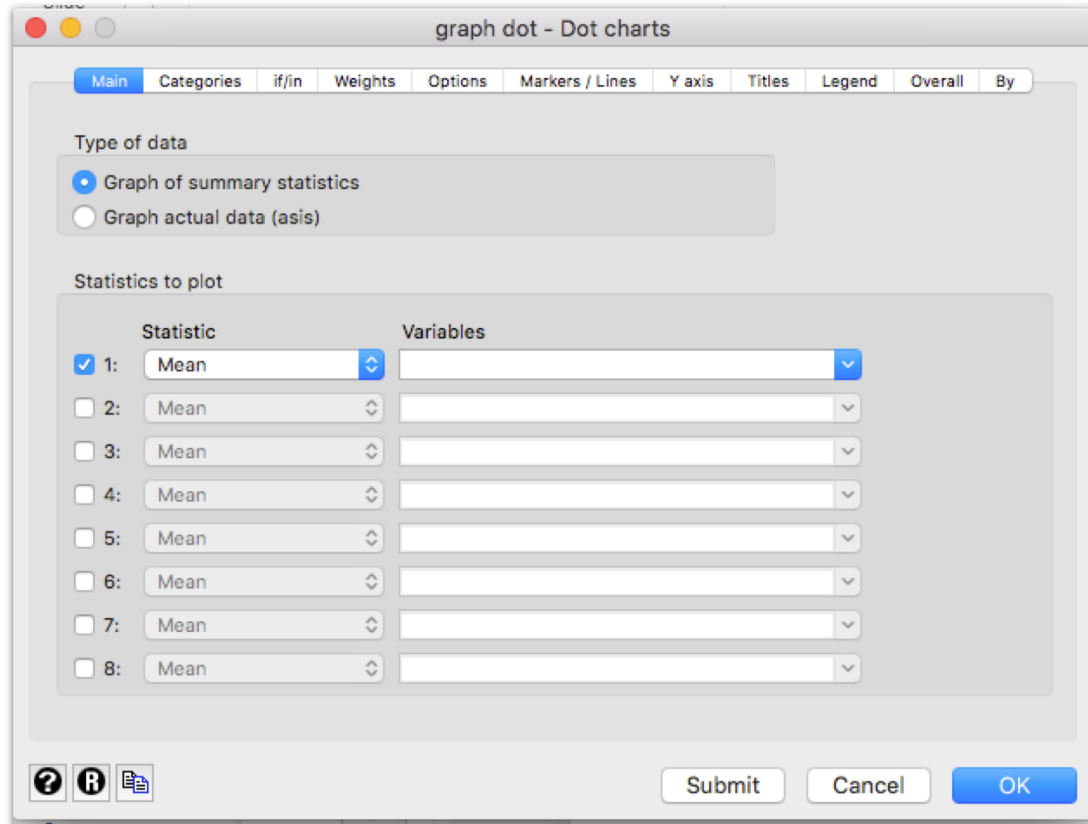
-subset data

Outline

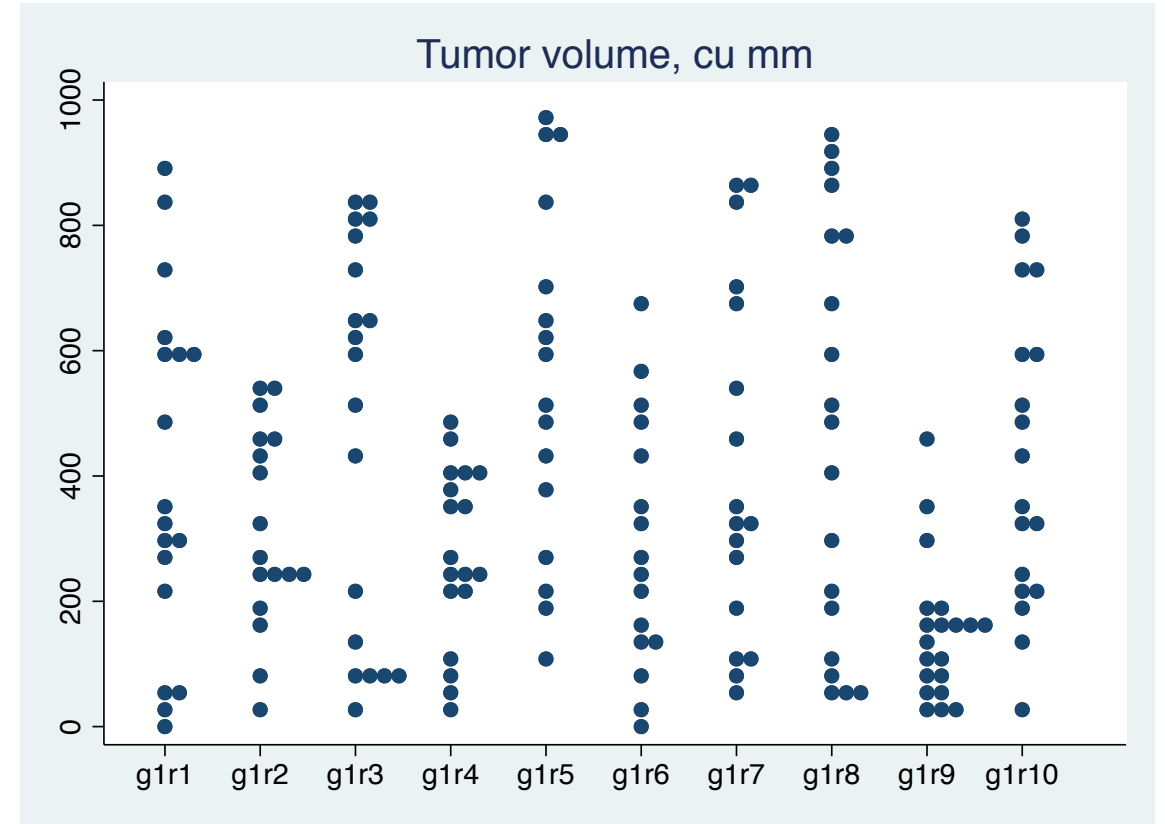
- ◇ Anatomy of a plot
- ◇ Help files
- ◇ Histogram
- ◇ Bar chart
- ◇ Box plot
- ◇ **Dot plot**
- ◇ Coefficient plot
- ◇ Scatter
 - Linear fit

- ◇ Layering graphs
- ◇ Combining graphs
- ◇ Saving and naming
- ◇ Figure options
 - *Titles, captions, notes*
 - *Axis titles and legends*
 - *Colors*
 - *Region, margins*
 - *Line patterns*
 - *Marker symbols*
 - *Marker size*
 - *Text size*
 - *Line width*
 - *Schemes*

Dot plot

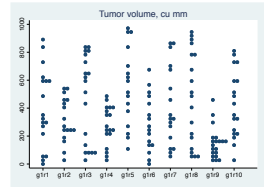


Graphics < Dot chart



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

Dot plot syntax



`dotplot varname`

Options

`, over(varname)`

`center`

`if`

`in`

`mean`

`median`

- display dotplot for each value of `groupvar`
- center the dot for each column
- use logic operators
- subset data
- place horizontal line at mean
- place horizontal line at median

Outline

- ◇ Anatomy of a plot
- ◇ Help files
- ◇ Histogram
- ◇ Bar chart
- ◇ Box plot
- ◇ Dot plot
- ◇ **Coefficient plot**
- ◇ Scatter
 - Linear fit

- ◇ Layering graphs
- ◇ Combining graphs
- ◇ Saving and naming
- ◇ Figure options
 - *Titles, captions, notes*
 - *Axis titles and legends*
 - *Colors*
 - *Region, margins*
 - *Line patterns*
 - *Marker symbols*
 - *Marker size*
 - *Text size*
 - *Line width*
 - *Schemes*

Coefficient plot

Load package: `ssc install coefplot`

`help coefplot`

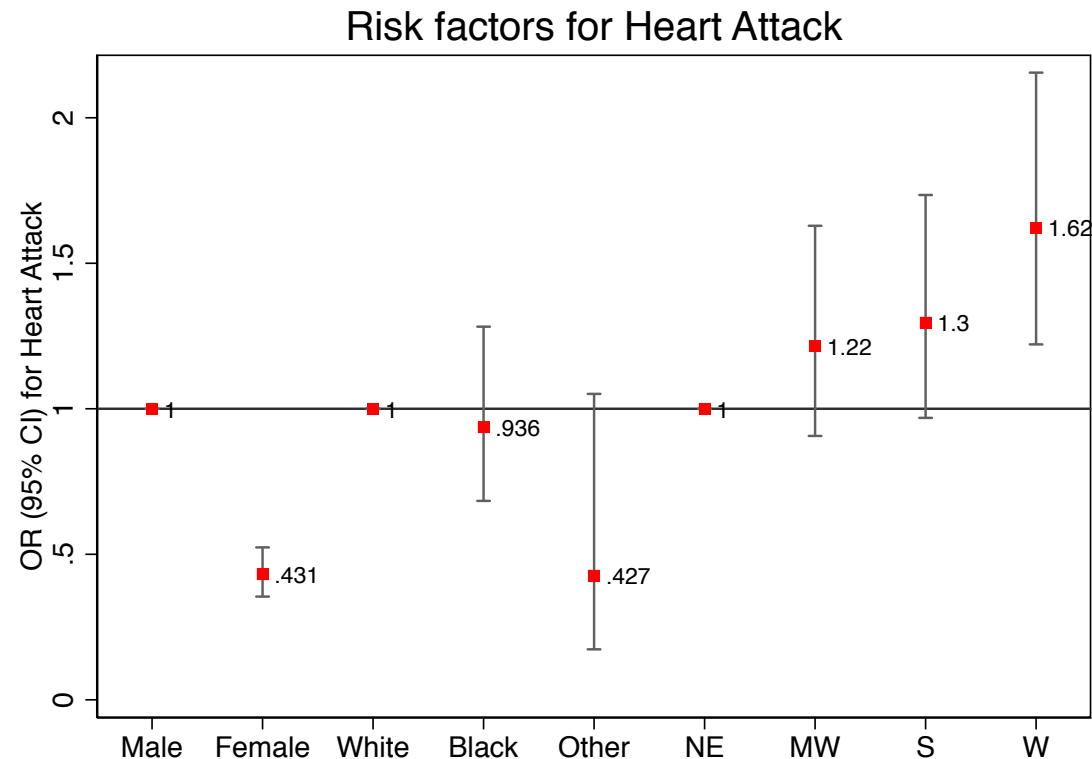
*No drop down menu

Coefficient plot example

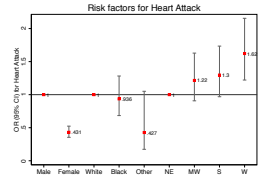
```
webuse nhanes2, clear

logit heartatk i.sex i.race i.region

coefplot, vertical eform drop(_cons)
baselevels yline(1) ///
ciopts(recast(rcap)) ///
scheme(s1mono) ///
mlabel mcolor(red) msymbol(s) format(%9.3g) ///
ytile(OR (95% CI) for Heart Attack) ///
title (Risk factors for Heart Attack)
```



Coefficient plot syntax



First step: Run regression

```
coefplot
```

Options

```
eform
```

- plot exponentiated point estimates and confidence intervals

```
baselevels
```

- include base levels

```
drop(coeflist)
```

- omit coefficients from plot

```
levels
```

- omit confidence intervals

```
mlabels
```

- add custom marker labels

```
veritcal
```

- coefficient values are on y axis

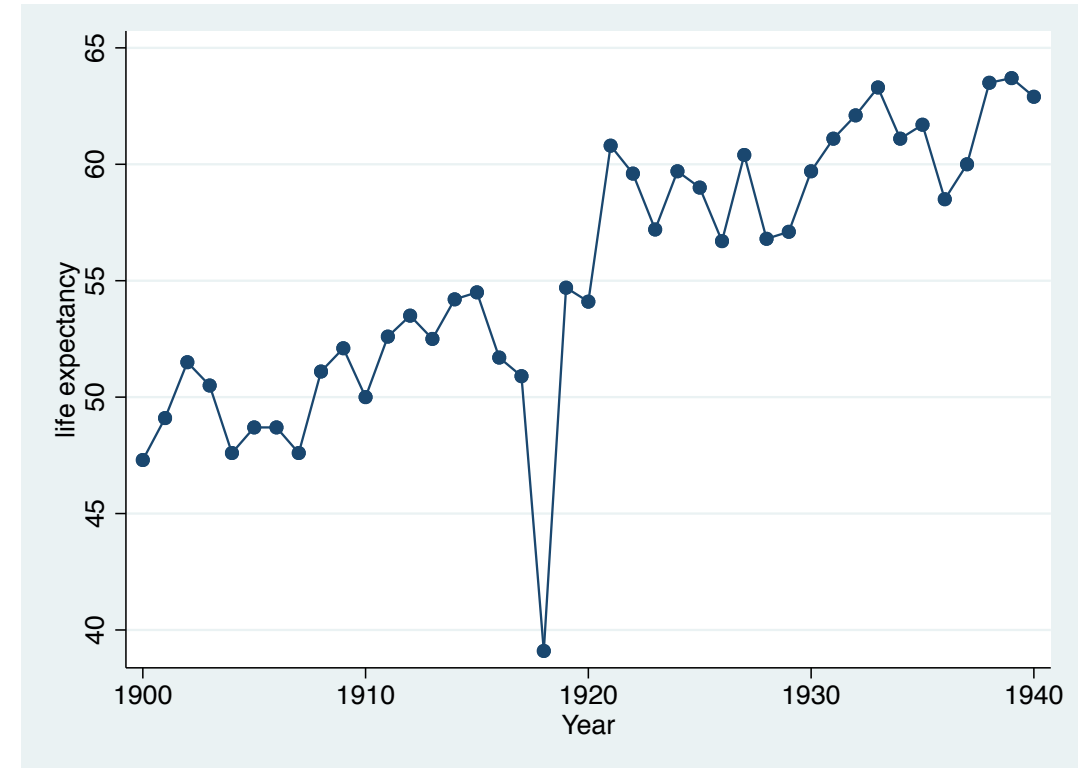
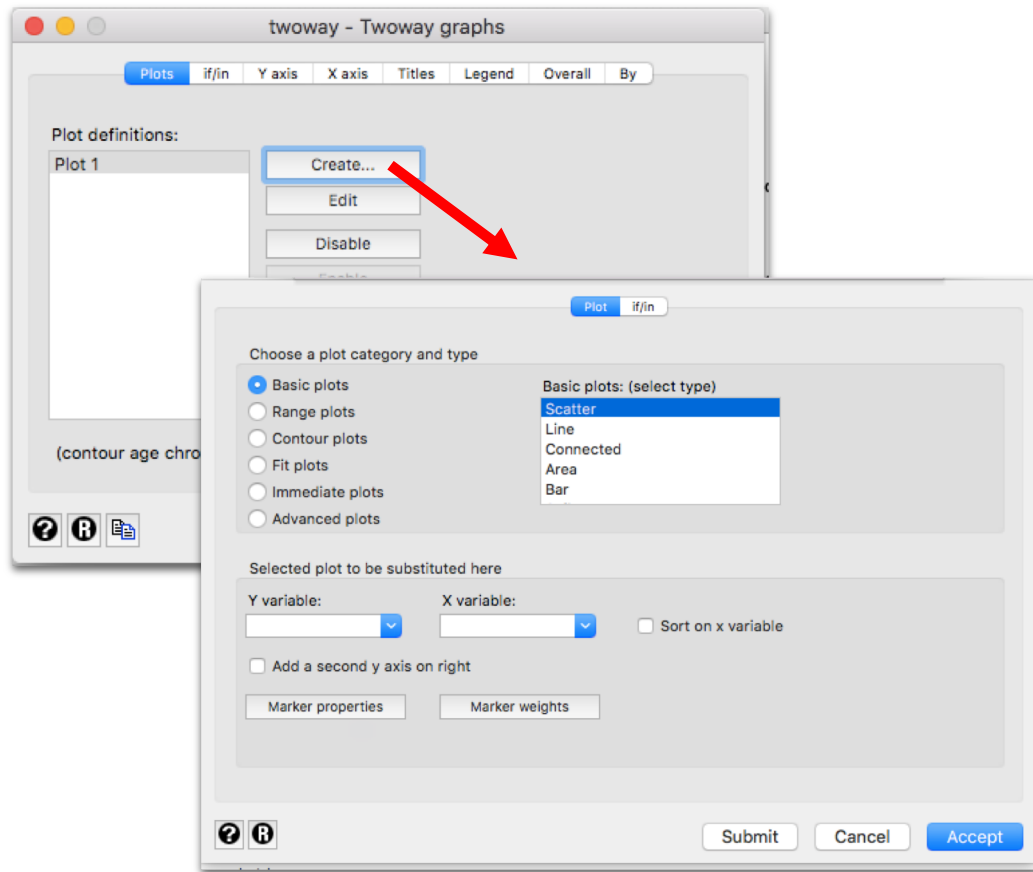
...many more options!

Outline

- ◇ Anatomy of a plot
- ◇ Help files
- ◇ Histogram
- ◇ Bar chart
- ◇ Box plot
- ◇ Dot plot
- ◇ Coefficient plot
- ◇ **Scatter**
 - **Linear fit**

- ◇ Layering graphs
- ◇ Combining graphs
- ◇ Saving and naming
- ◇ Figure options
 - *Titles, captions, notes*
 - *Axis titles and legends*
 - *Colors*
 - *Region, margins*
 - *Line patterns*
 - *Marker symbols*
 - *Marker size*
 - *Text size*
 - *Line width*
 - *Schemes*

Scatter plot



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

Graphics < Twoway

Scatter plot syntax

```
scatter var1 var2
```

Options

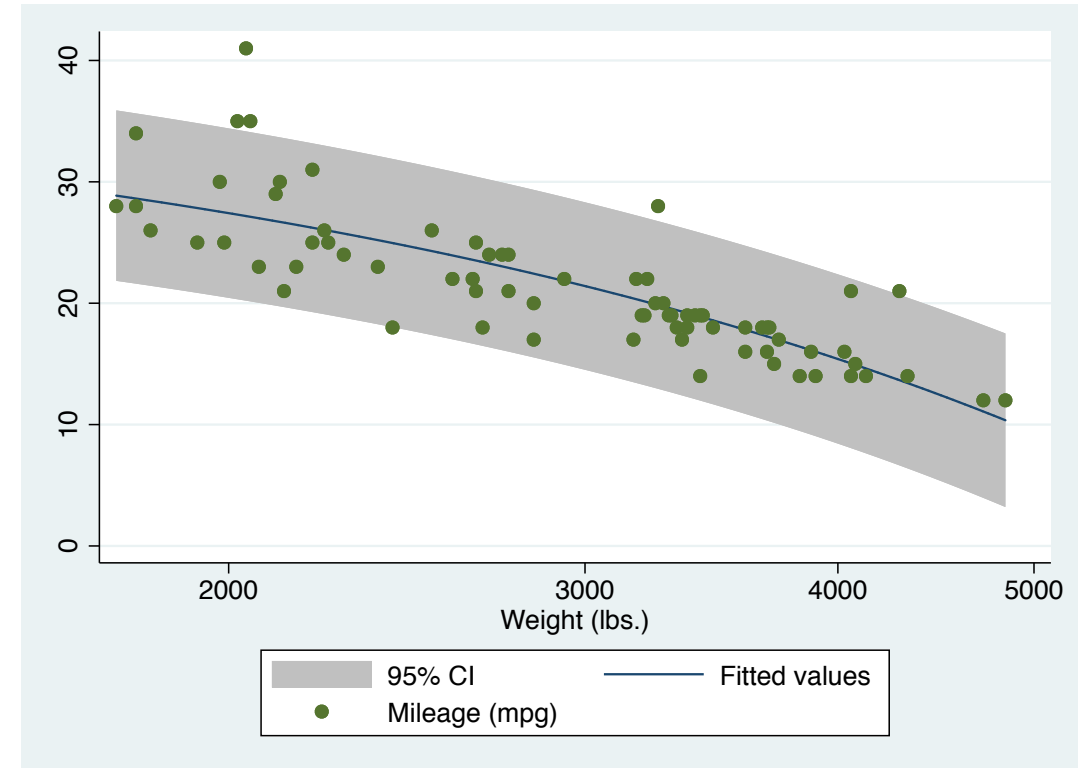
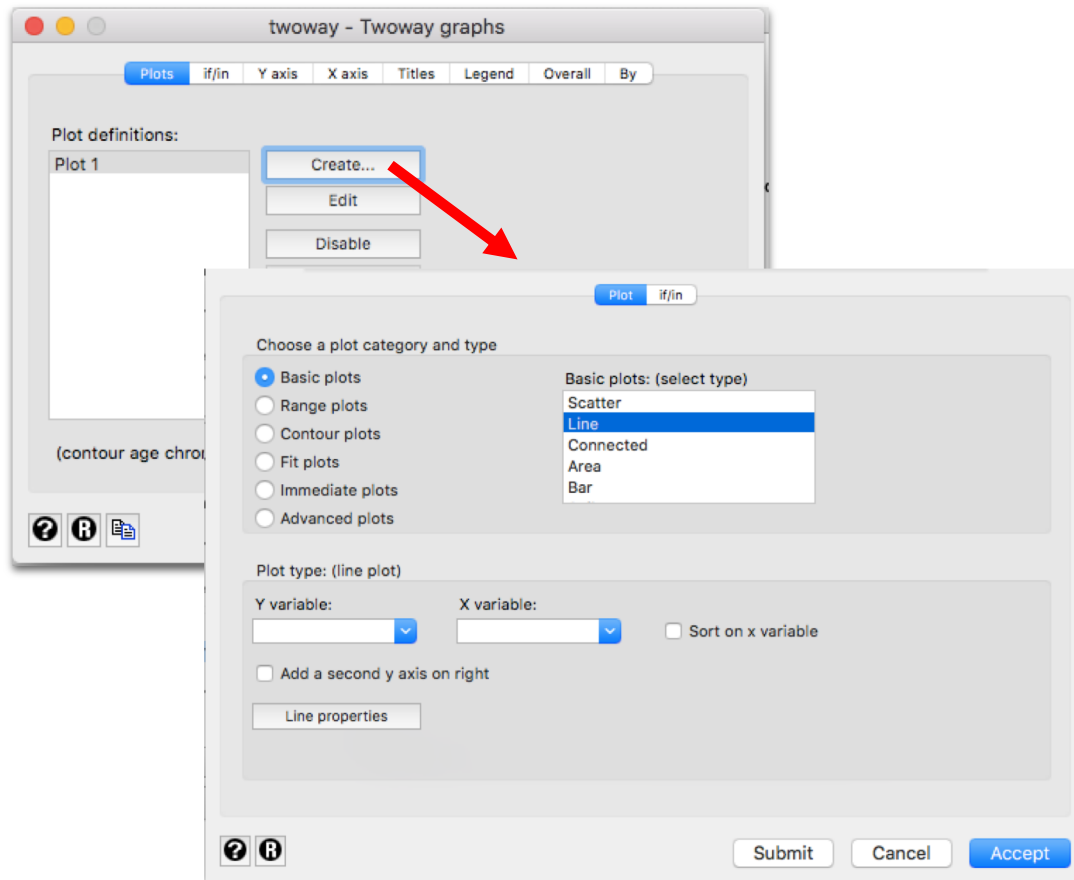
```
connect (style)
```

```
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

Outline

- ◇ Anatomy of a plot
- ◇ Help files
- ◇ Histogram
- ◇ Bar chart
- ◇ Box plot
- ◇ Dot plot
- ◇ Coefficient plot
- ◇ Scatter
 - Linear fit

- ◇ **Layering graphs**
- ◇ Combining graphs
- ◇ Saving and naming
- ◇ Figure options
 - *Titles, captions, notes*
 - *Axis titles and legends*
 - *Colors*
 - *Region, margins*
 - *Line patterns*
 - *Marker symbols*
 - *Marker size*
 - *Text size*
 - *Line width*
 - *Schemes*

Layering graphs

twoway

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

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

	after each plot
///	after each line except last

Outline

- ◇ Anatomy of a plot
- ◇ Help files
- ◇ Histogram
- ◇ Bar chart
- ◇ Box plot
- ◇ Dot plot
- ◇ Coefficient plot
- ◇ Scatter
 - Linear fit

- ◇ Layering graphs
- ◇ **Combining graphs**
- ◇ Saving and naming
- ◇ Figure options
 - *Titles, captions, notes*
 - *Axis titles and legends*
 - *Colors*
 - *Region, margins*
 - *Line patterns*
 - *Marker symbols*
 - *Marker size*
 - *Text size*
 - *Line width*
 - *Schemes*

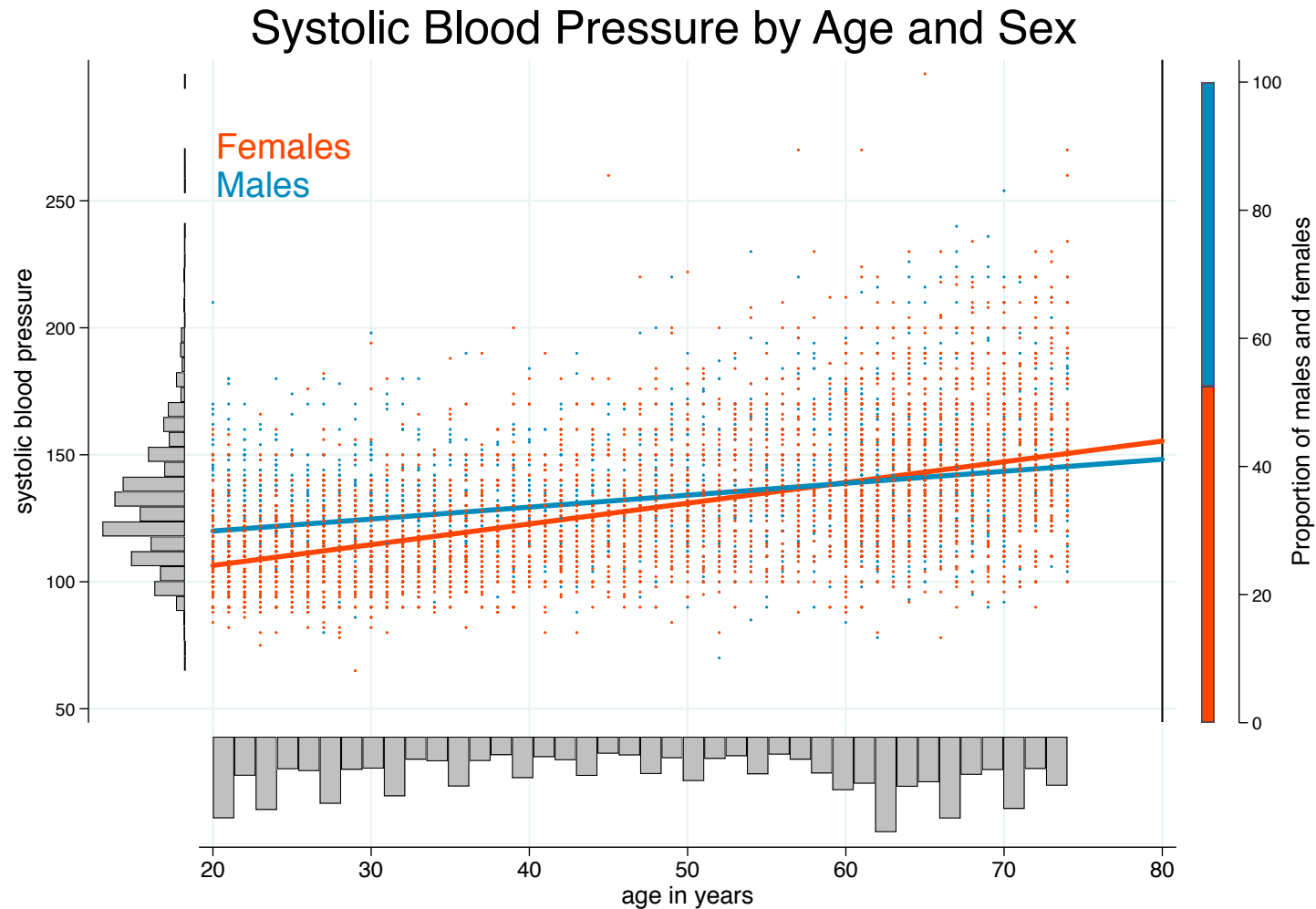
Combining graphs: `graph combine`

`graph combine graph1 graph2 graph3...`

options

- `, altshrink` - reduces size of text and markers
- `, cols(#)` - number of columns
- `, rows(#)` - number of rows

Graph combine example



Outline

- ◇ Anatomy of a plot
- ◇ Help files
- ◇ Histogram
- ◇ Bar chart
- ◇ Box plot
- ◇ Dot plot
- ◇ Coefficient plot
- ◇ Scatter
 - Linear fit

- ◇ Layering graphs
- ◇ Combining graphs
- ◇ **Saving and naming**
- ◇ Figure options
 - *Titles, captions, notes*
 - *Axis titles and legends*
 - *Colors*
 - *Region, margins*
 - *Line patterns*
 - *Marker symbols*
 - *Marker size*
 - *Text size*
 - *Line width*
 - *Schemes*

Saving, naming & exporting

, saving `graphname`, replace

-Save the graph as a Stata .gph file on

, name `graphname`

-graph is saved in memory instead of on disk. Once we quit Stata, the graph in memory is erased.

`graph export graphname, as(extension*)`

*extension options: .pdf, .jpg, .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



Outline

- ◇ Anatomy of a plot
- ◇ Help files
- ◇ Histogram
- ◇ Bar chart
- ◇ Box plot
- ◇ Dot plot
- ◇ Coefficient plot
- ◇ Scatter
 - Linear fit

- ◇ Layering graphs
- ◇ Combining graphs
- ◇ Saving and naming

- ◇ **Figure options**
 - *Titles, captions, notes*
 - *Axis titles and legends*
 - *Colors*
 - *Region, margins*
 - *Line patterns*
 - *Marker symbols*
 - *Marker size*
 - *Text size*
 - *Line width*
 - *Schemes*

Title, caption & note options

title options

`title(xyz)`

`subtitle(xyz)`

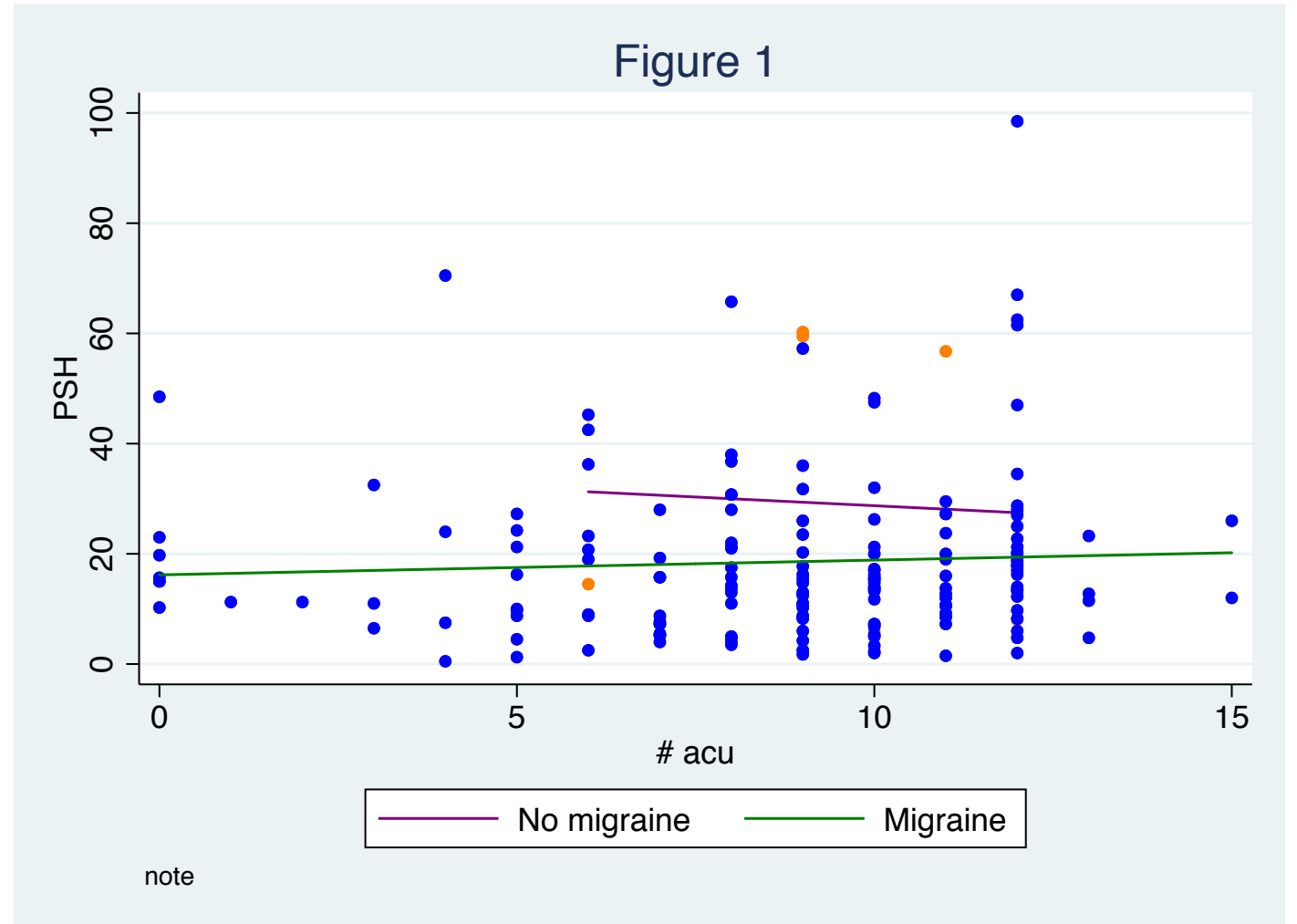
`caption(xyz)`

`note(xyz)`

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

Practice: Figure 1

- Open the 'lecture7.do' file
- Run the code for 'Figure 1'



Practice: Figure 1

- ◇ Change the title to “Figure 1: Acupuncture treatments and headache score”

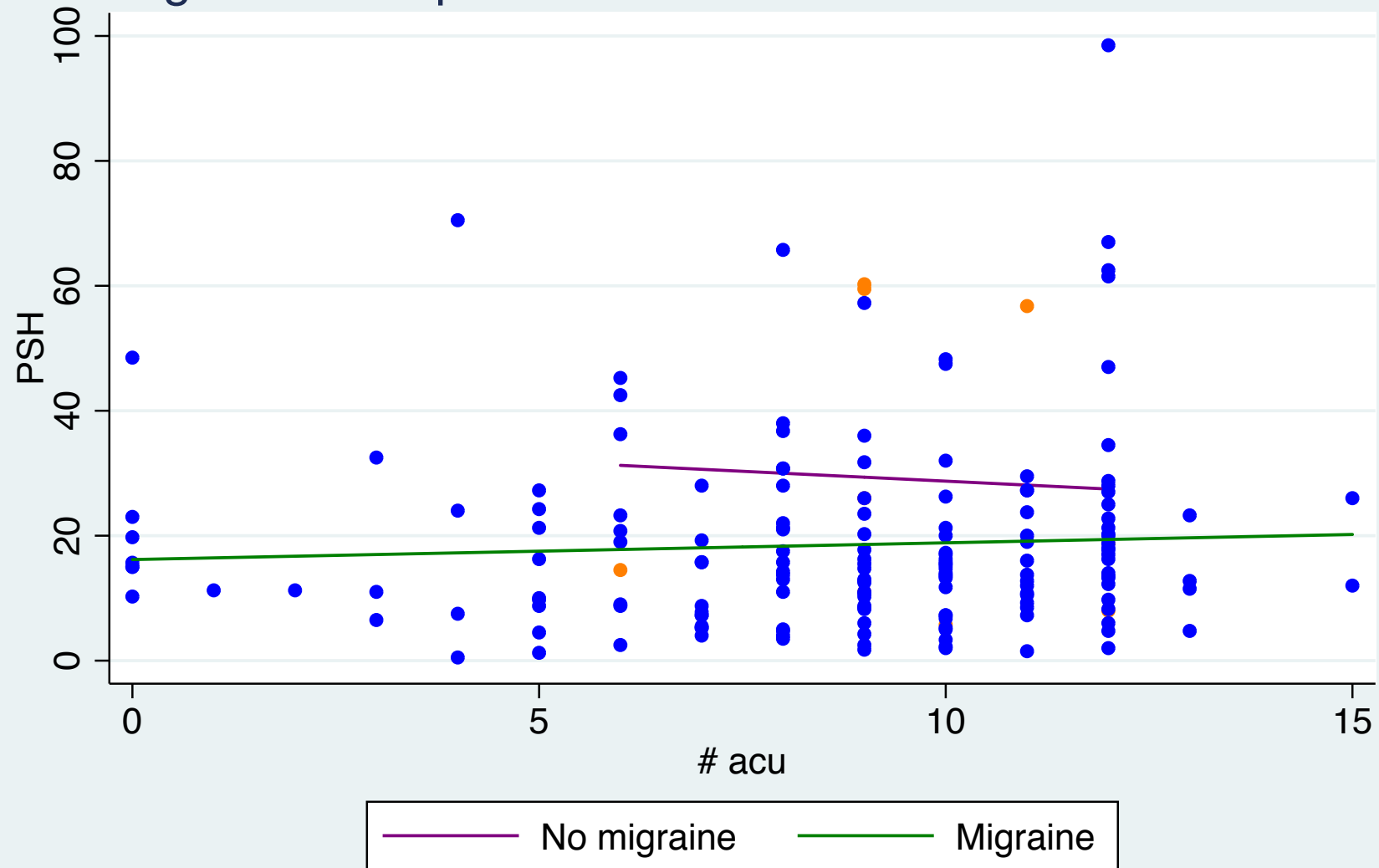
```
title (Figure 1: Acupuncture treatments and headache score)
```

- ◇ Add a note

```
note (Lecture 4 practice)
```

- ◇ Run your commands for Figure 1 to see if it worked

Figure 1 : Acupuncture treatments and headache score



Lecture 7 practice

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

- ◇ Change the y-axis to “Post-study headache score”

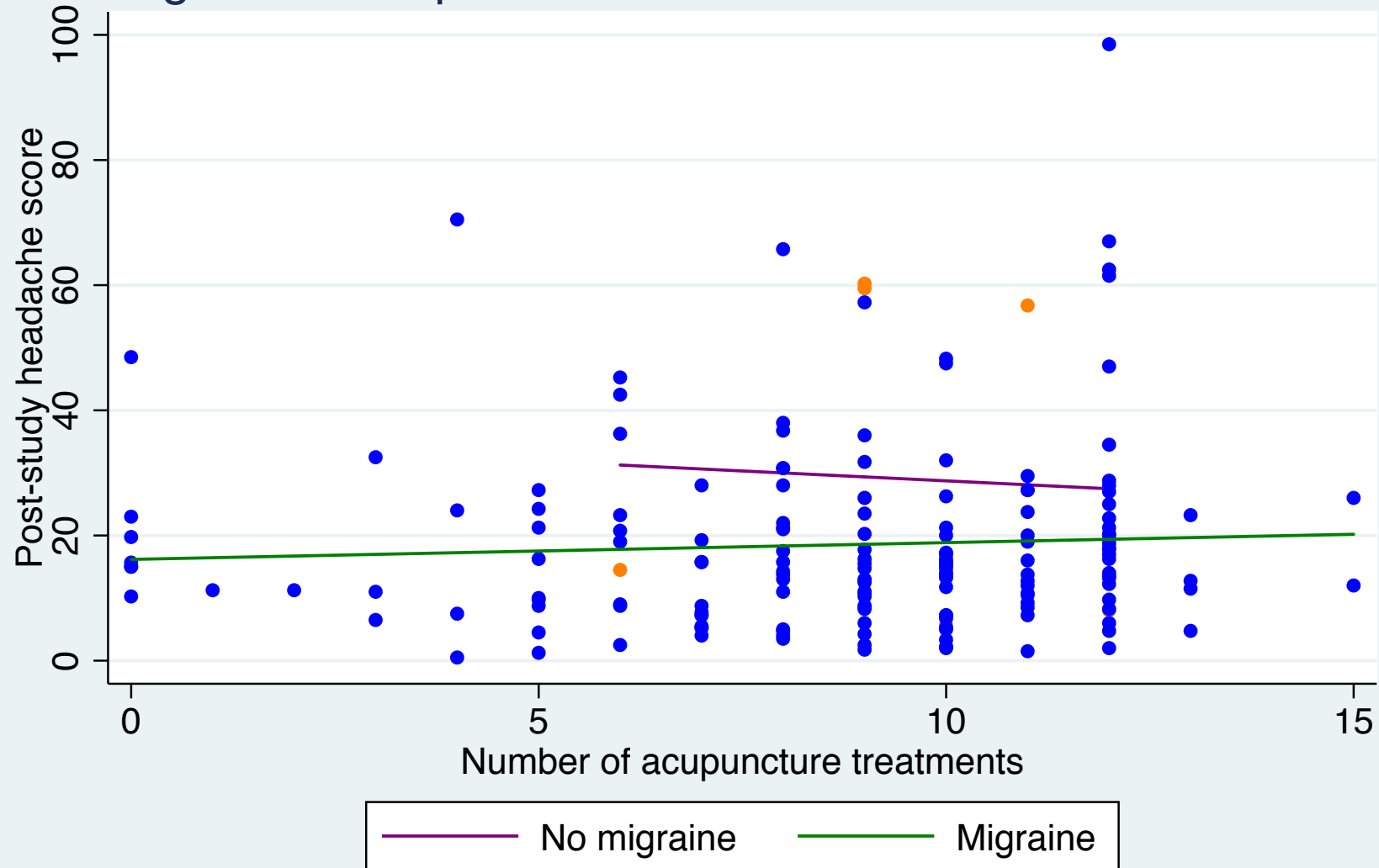
```
ytitle(Post-study headache score)
```

- ◇ Change the x-axis to “Number of acupuncture treatments”

```
xtitle(Number of acupuncture treatments)
```

- ◇ Run your commands for Figure 1 to see if it worked

Figure 1 : Acupuncture treatments and headache score



Lecture 7 practice

Color options



Marker color

`mcolor(blue)`

Marker label color

`mlcolor(blue)`

Line color

`lcolor(blue)`

Text color

`color(blue)`

Gridline color

`glcolor(blue)`

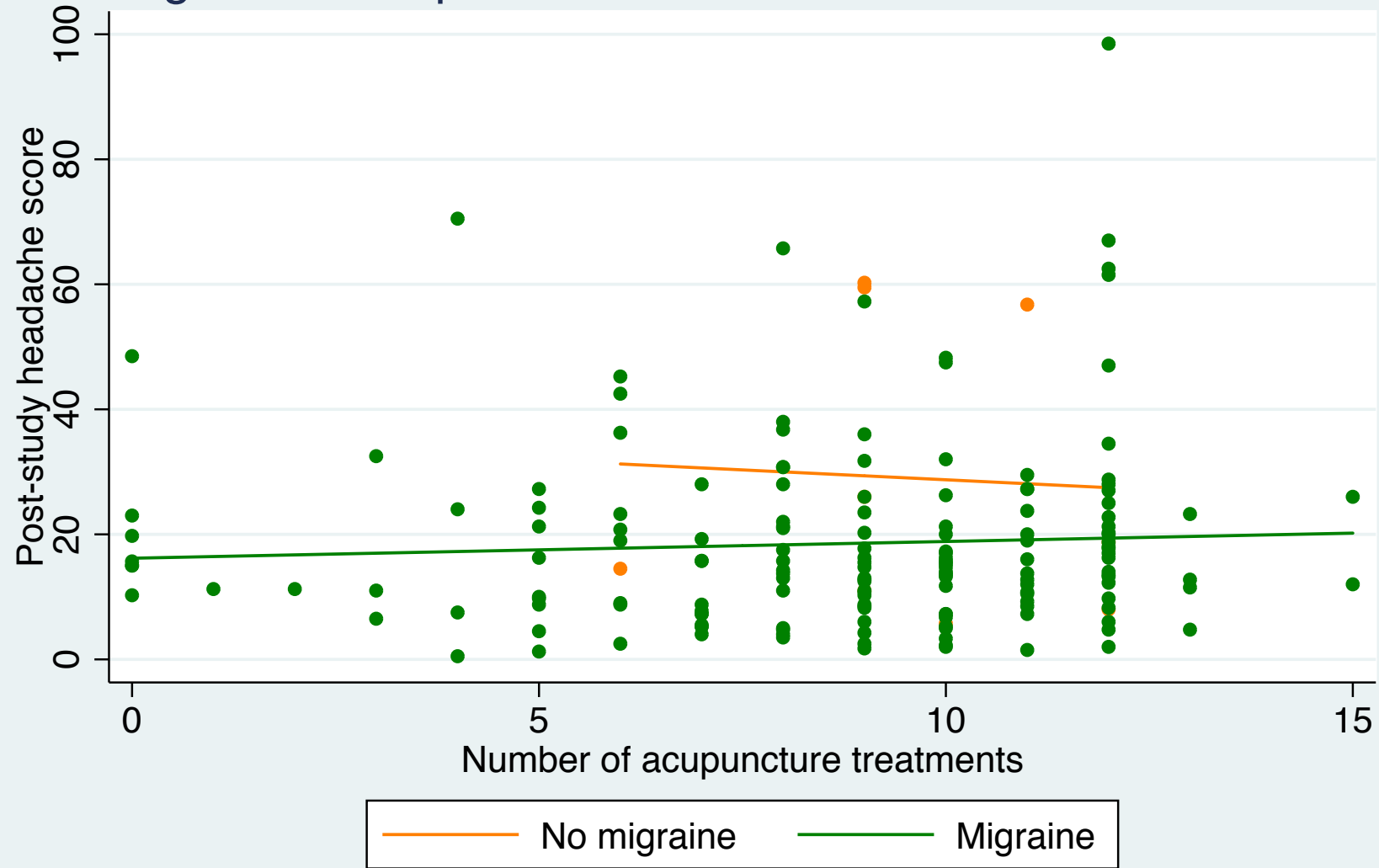
Axis label color

`labcolor(blue)`

Practice: Figure 1

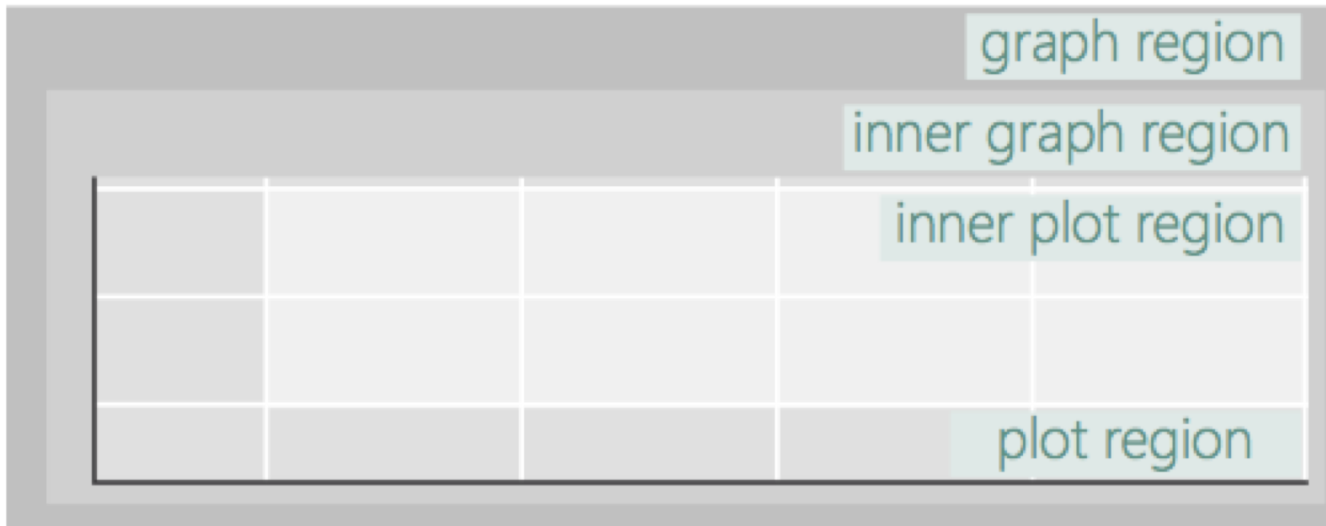
- ◇ Change the marker color to orange for the no migraine group
 - `mcolor(orange)`
- ◇ Change the line color to green for the migraine group
 - `lcolor(green)`
- ◇ Run your commands for Figure 1 to see if it worked

Figure 1 : Acupuncture treatments and headache score



Lecture 7 practice

Region options



Graph region color

`graphregion (color (blue) )`

Plot region color

`plotregion (color (blue) )`

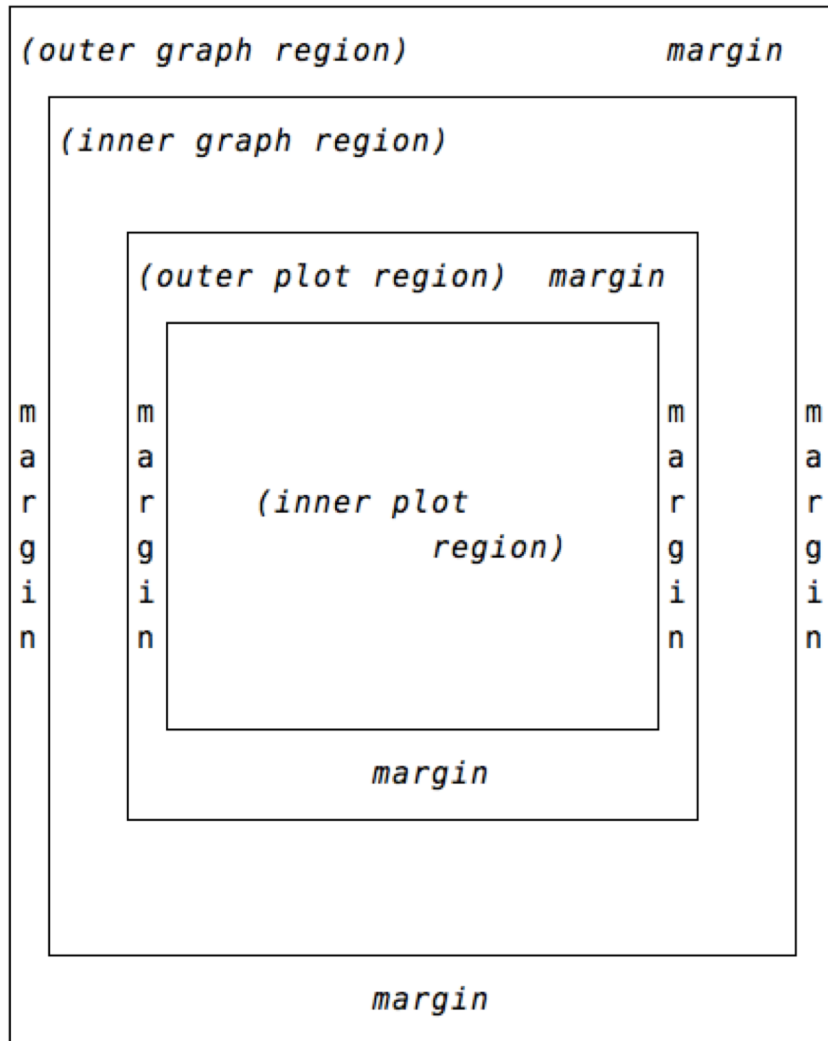
Graph region outline color

`graphregion (lcolor (blue) )`

Plot region outline color

`plotregion (lcolor (blue) )`

Margin options



Title margin

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

Graph region margin

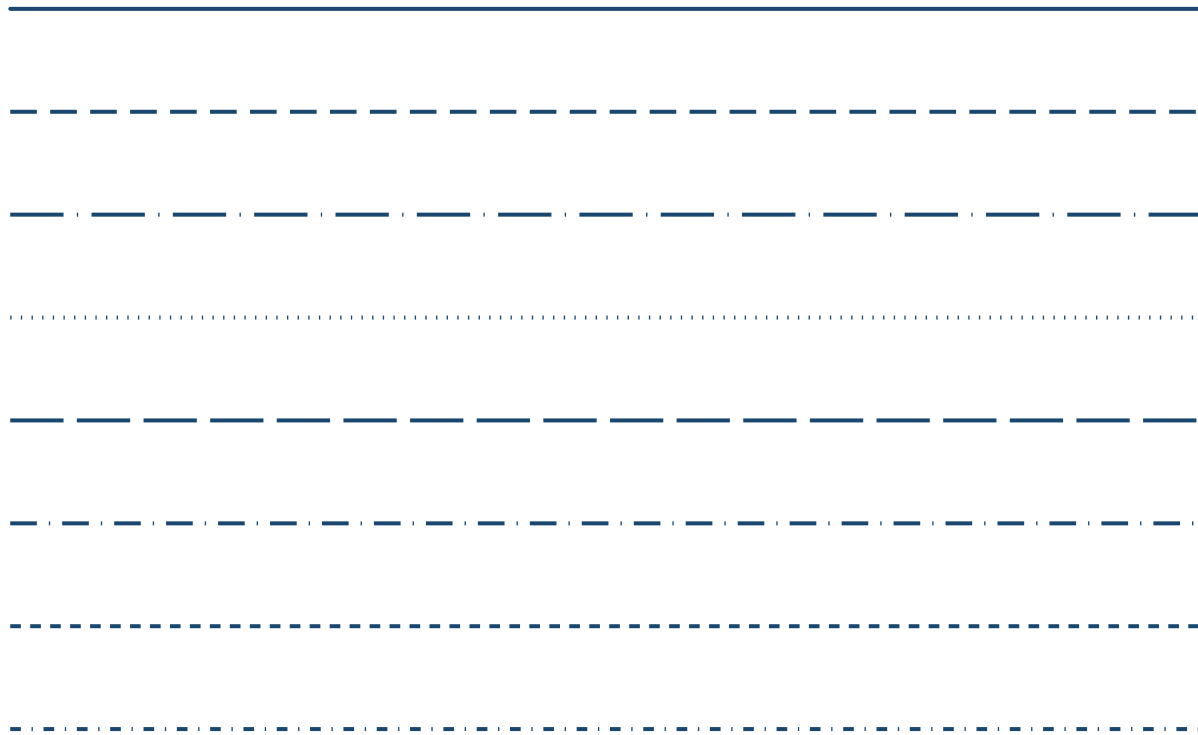
`graphregion(margin(size))`

Size options

zero
tiny
vsmall
small
medsmall
medium
medlarge
large
vlarge

Line pattern options

, lpattern(pattern)



solid

dash

longdash_dot

dot

longdash

dash_dot

shortdash

shortdash_dot

blank

, glpattern(pattern)

-grid line pattern

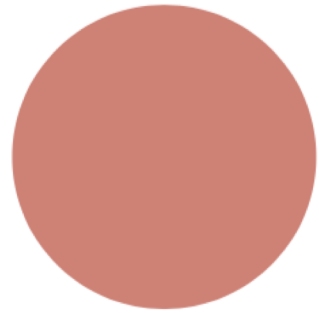
Marker symbol options

, 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

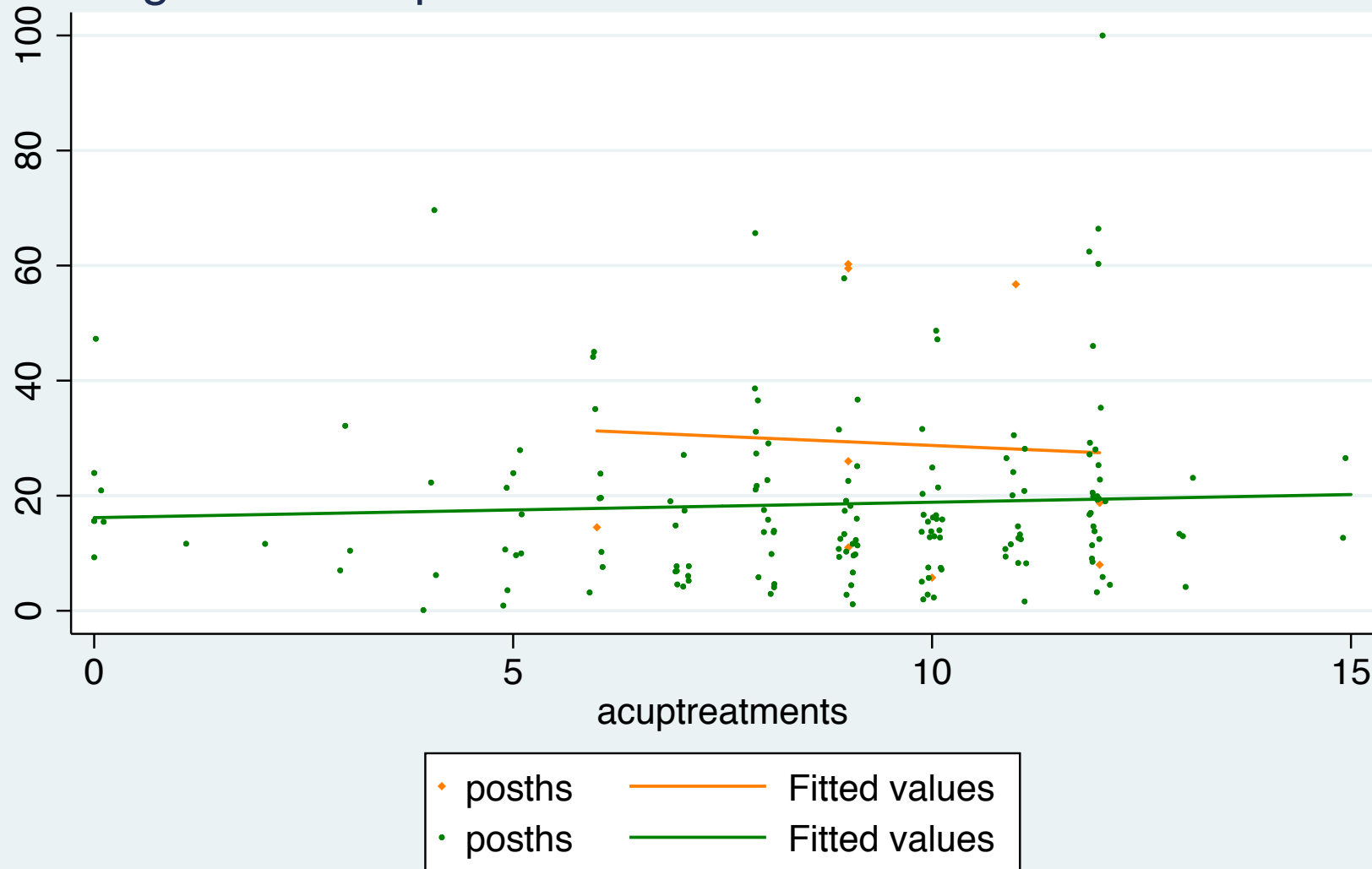
● tiny

● vtiny

Practice: Figure 1

- ◇ Change the marker 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)`
- ◇ Run your commands for Figure 1 to see if it worked

Figure 1 : Acupuncture treatments and headache score



Lecture 7 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

Line width options

, lwidth(*width*)

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

, glwidth(*width*)

– Grid line width

Schemes

, scheme (**scheme**)

Stata scheme options

s2color

- factory setting

s2mono

- s2color in monochrome

s2gcolor

- color white used in the Stata manuals

s2manual

- s2gcolor in monochrome

s1rcolor

- a plain look on black background

s1color

- a plain look

s1mono

- a plain look in monochrome

s1manual

- s1mono but smaller

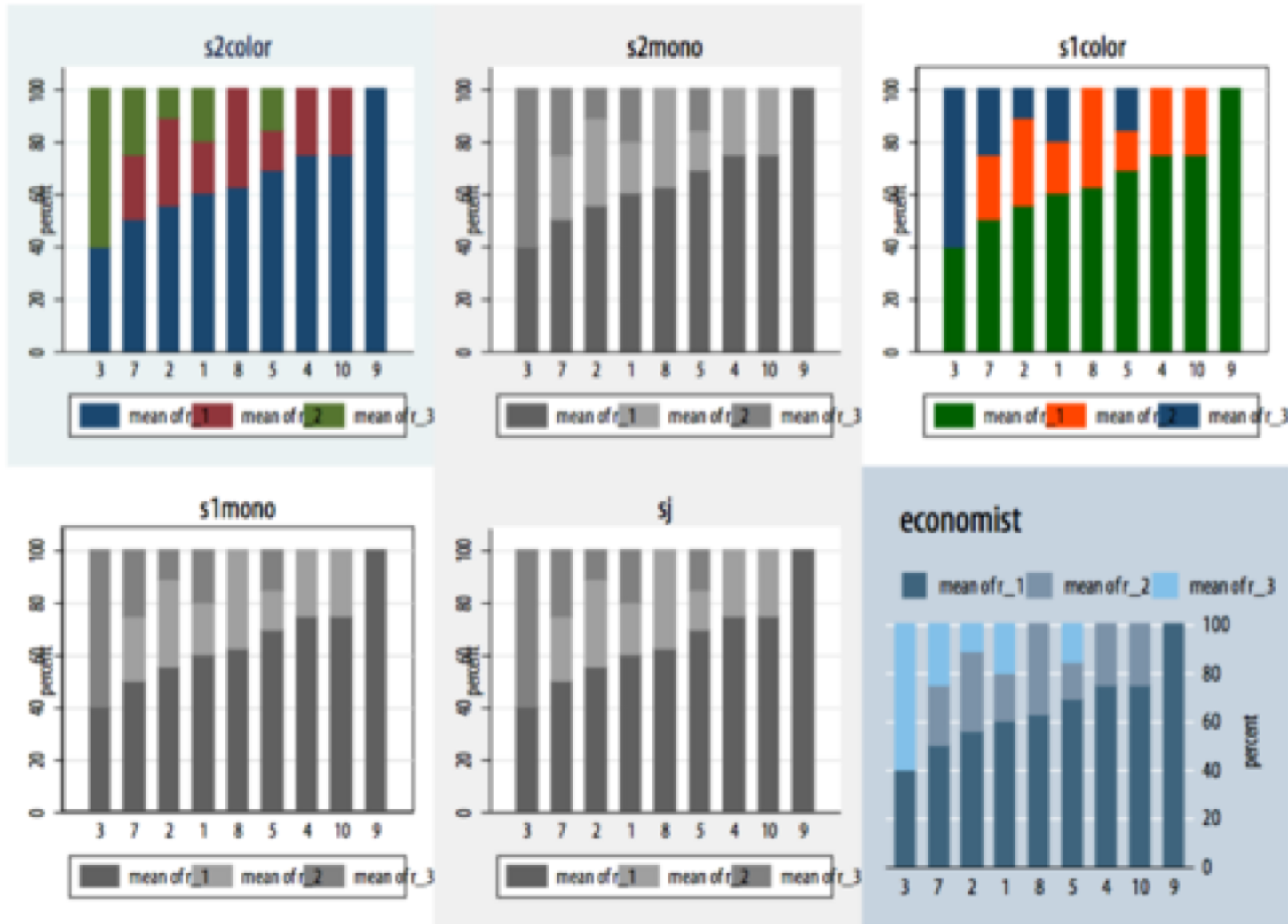
economist

- The Economist magazine

sj

- Stata Journal

Stata schemes



Scheme options

set scheme `schemename`

– set your default scheme (in .do file)

set scheme `schemename`, permanently

– set your default scheme permanently

Or:

[preferences<graph preferences]

**you can also select default text here*

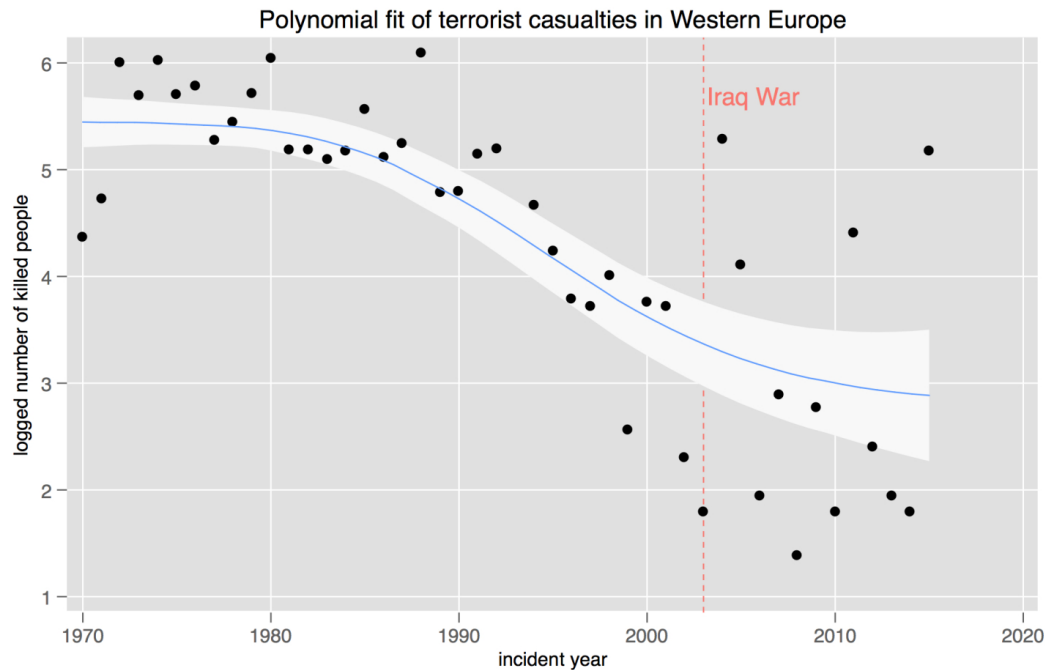
Additional options

1. Create and save your own scheme Download other schemes from the internet

Download scheme: blindschemes

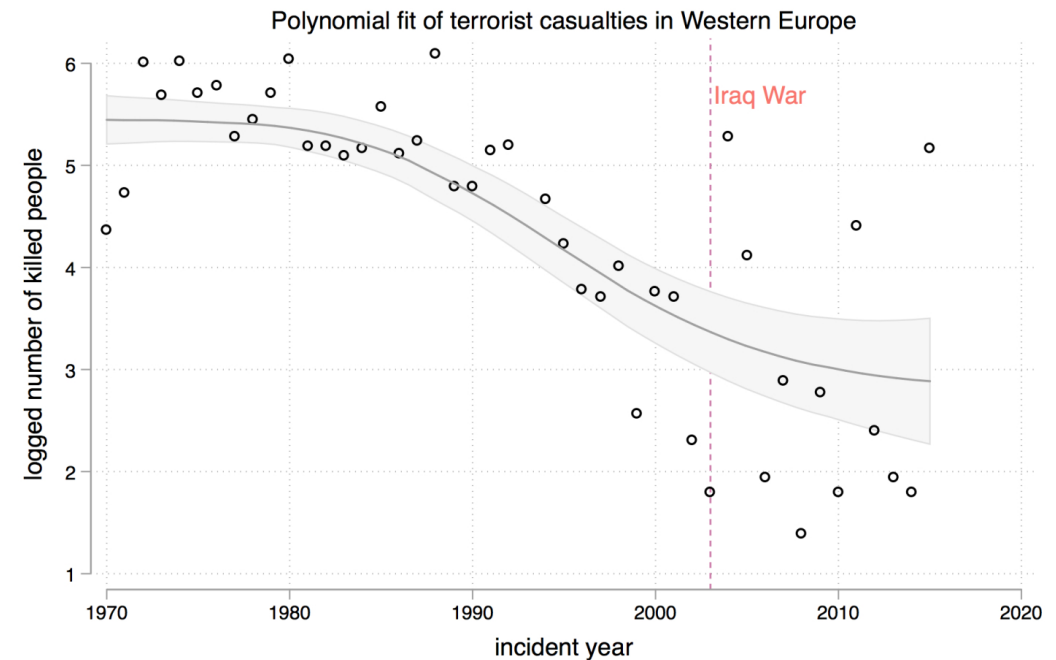
```
ssc install blindschemes
```

1. “plottig”: mimics ggplot in R



```
, scheme(plottig)
```

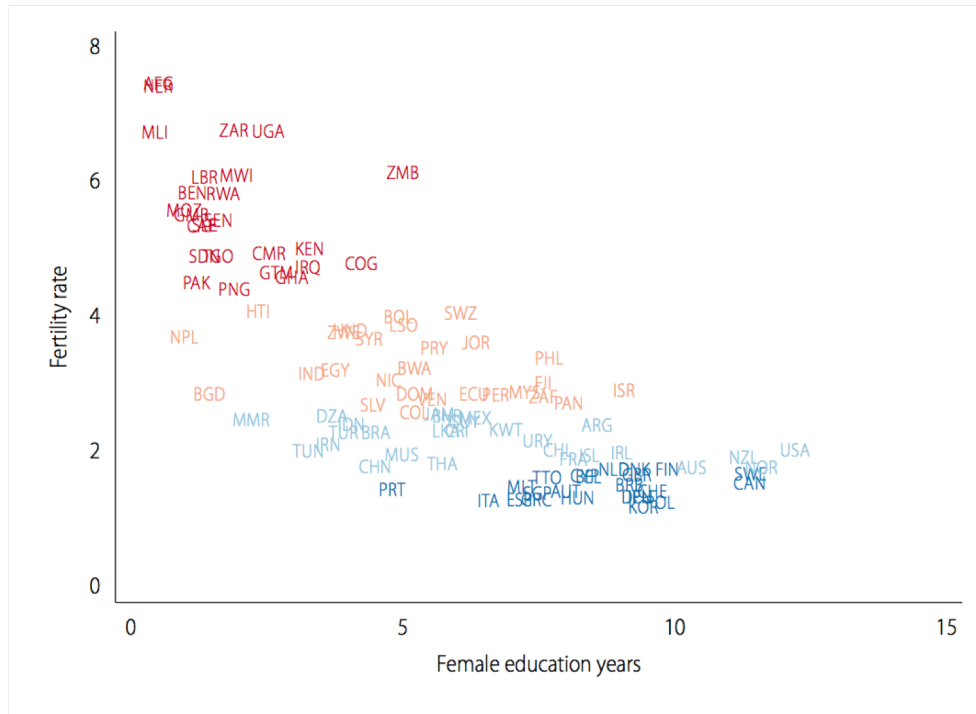
2. “plotplainblind”: color-blind friendly pallet



```
, scheme(plotplainblind)
```

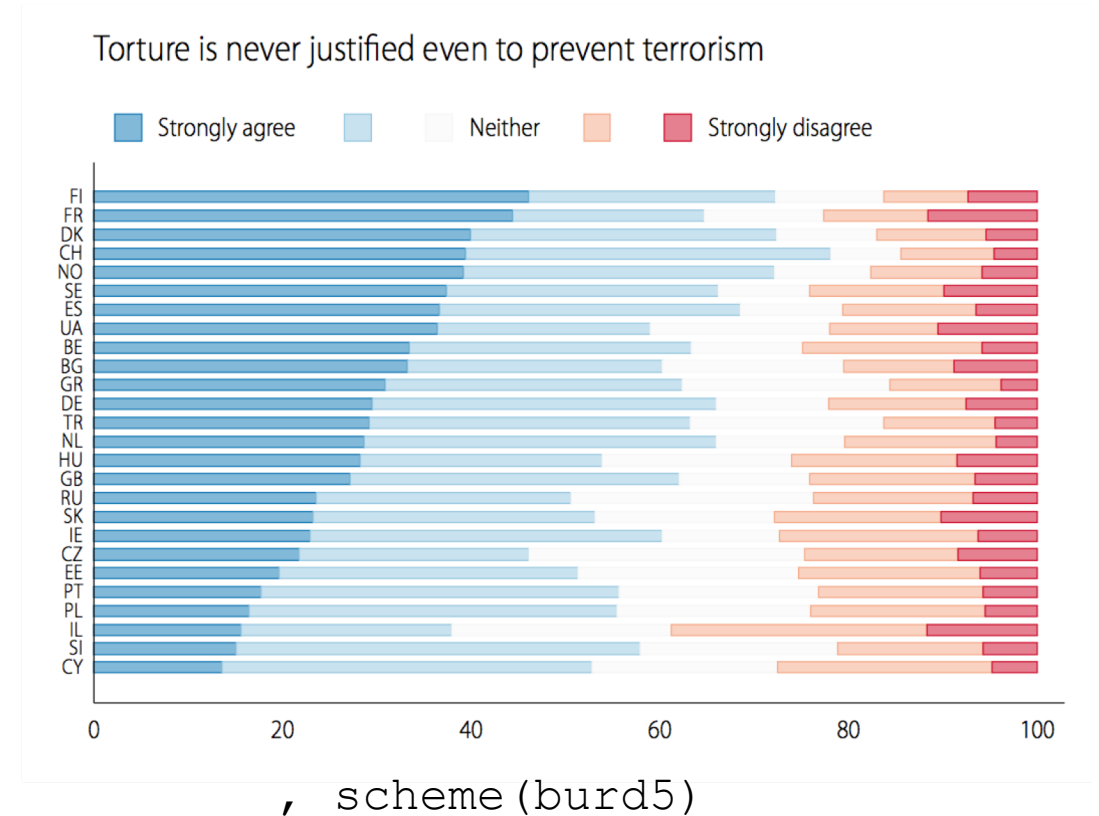
Download scheme: scheme-burd

```
ssc install scheme-burd
```



```
, scheme (burd4)
```

<https://github.com/briatte/burd>

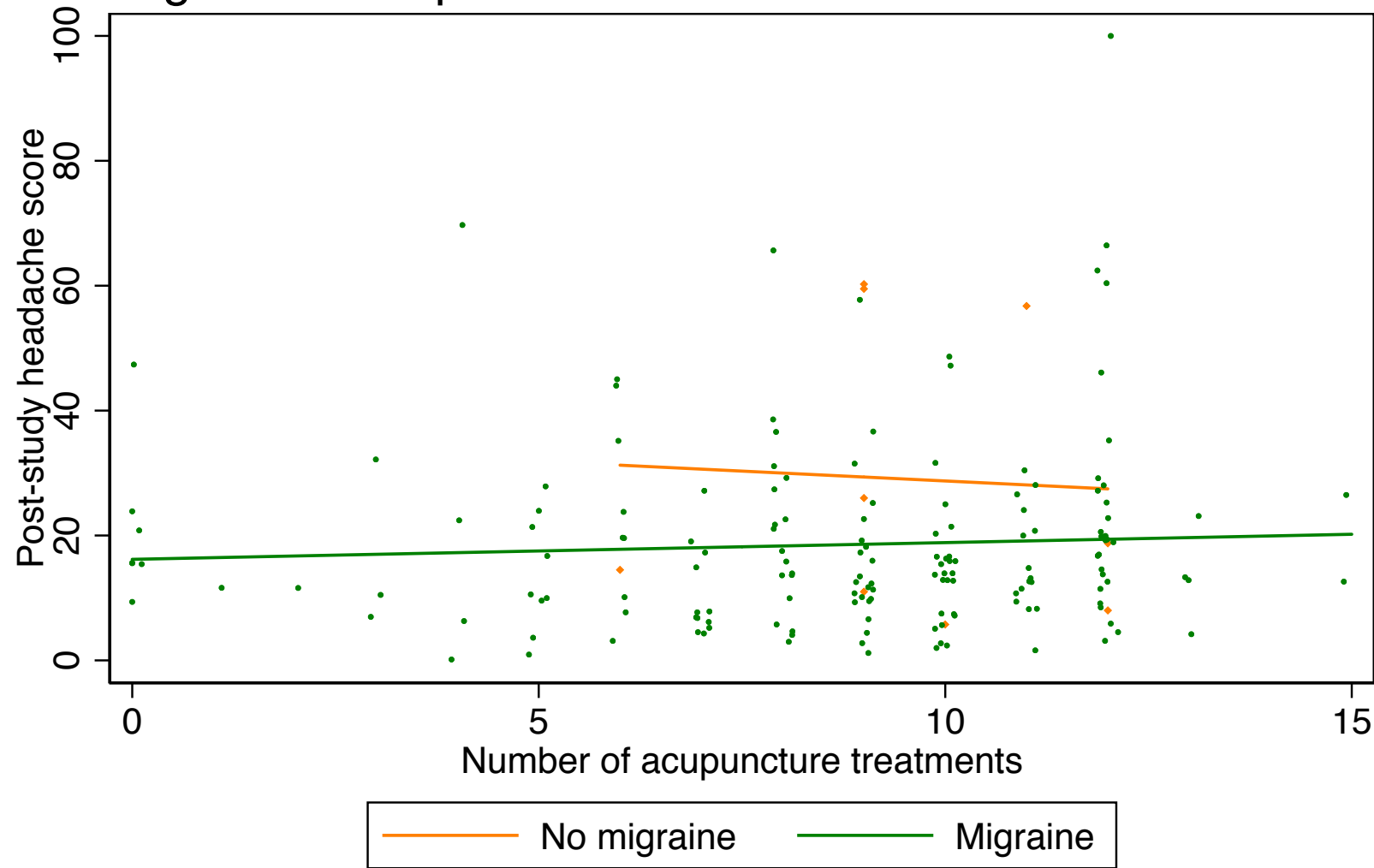


```
, scheme (burd5)
```

Practice: Figure 1

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

Figure 1 : Acupuncture treatments and headache score

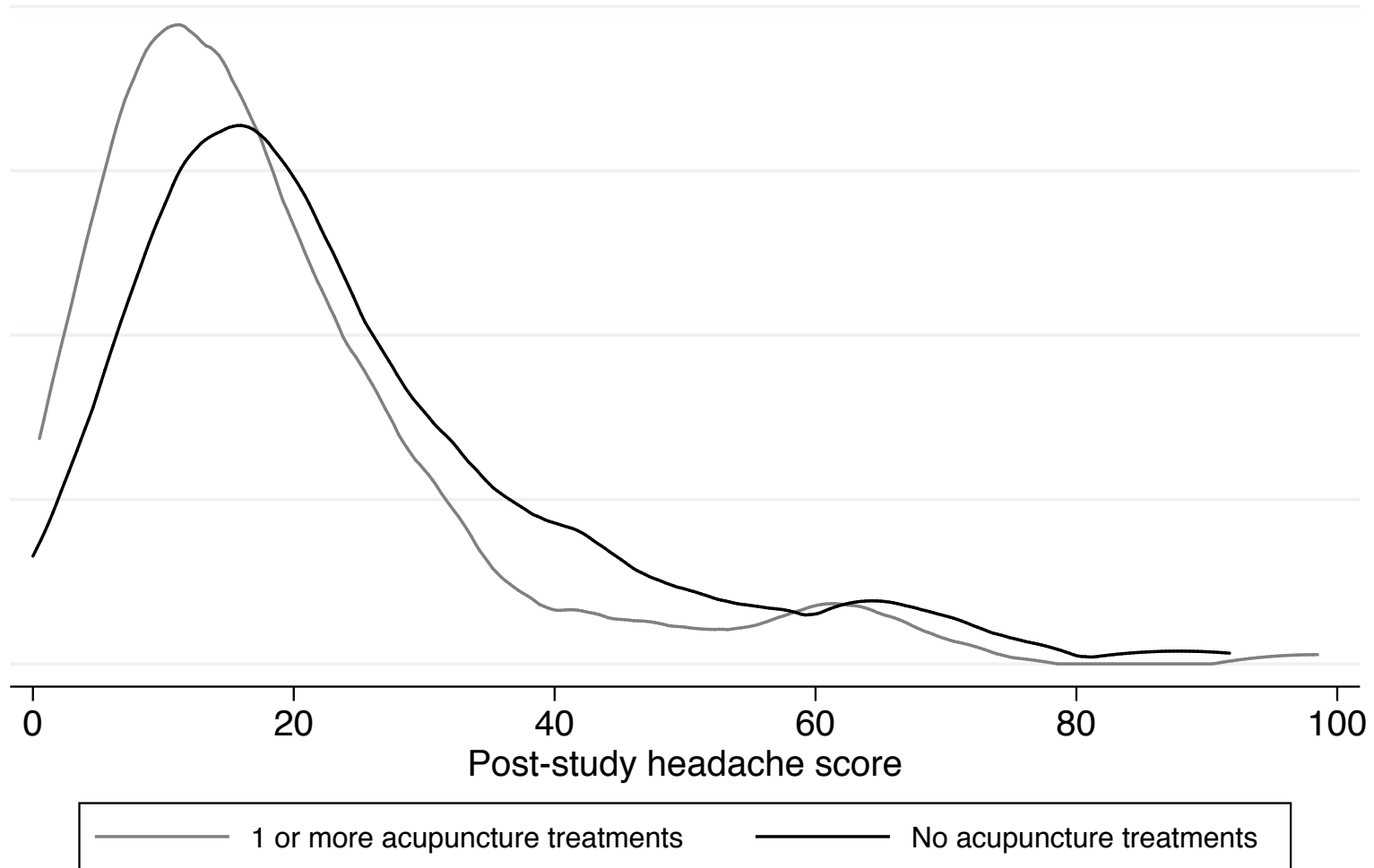


Lecture 7 practice

Practice: Figure 2

- ◇ Run the code for '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

◇ 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/>

Data visualization challenge!

Working in teams of 2 to 5:

- ◇ Visualize the effect (if any) of the acupuncture intervention (group) on post-treatment headache score (posths) or one-year follow-up headache score (oneyrhs)
- ◇ Use Stata
- ◇ Format however you like
- ◇ Write your team members' names in a note in the plot
- ◇ Upload to the course website [under “Data visualization challenge”]
- ◇ Turn in by Tuesday August 28 @ 1PM
 - .jpg or .pdf
- ◇ We have ***prizes*** for the winning team
 - For each team member: 10 extra credit points and a Stata tumbler (courtesy of Chuck Huber)