<div align="center">

Introduction to Stata
2019-20
Data Visualization – Some Basic Graphs

</div>

Summary

In this illustration, you will learn how to produce some (hopefully useful!) graphs from a Stata data set that you have imported into Stata.

Before you Begin: Be sure to have downloaded from the course website: ***framingham.dta***

<div align="center">

Introduction
Framingham Heart Study (Didactic Dataset)

</div>

The dataset you are using in this illustration **(framingham.Rdata)** is a subset of the data from the Framingham Heart Study, Levy (1999) National Heart Lung and Blood Institute, Center for Bio-Medical Communication.

The objective of the Framingham Heart Study was to identify the common factors or characteristics that contribute to cardiovascular disease (CVD) by following its development over a long period of time in a large group of participants who had not yet developed overt symptoms of CVD or suffered a heart attack or stroke. The researchers recruited 5,209 men and women between the ages of 30 and 62 from the town of Framingham, Massachusetts, and began the first round of extensive physical examinations and lifestyle interviews that they would later analyze for common patterns related to CVD development. Since 1948, the subjects have continued to return to the study every two years for a detailed medical history, physical examination, and laboratory tests, and in 1971, the study enrolled a second generation - 5,124 of the original participants' adult children and their spouses - to participate in similar examinations. In April 2002 the Study entered a new phase: the enrollment of a third generation of participants, the grandchildren of the original cohort. This step is of vital importance to increase our understanding of heart disease and stroke and how these conditions affect families. Over the years, careful monitoring of the Framingham Study population has led to the identification of the major CVD risk factors - high blood pressure, high blood cholesterol, smoking, obesity, diabetes, and physical inactivity - as well as a great deal of valuable information on the effects of related factors such as blood triglyceride and HDL cholesterol levels, age, gender, and psychosocial issues. With the help of another generation of participants, the Study may close in on the root causes of cardiovascular disease and help in the development of new and better ways to prevent, diagnose and treat cardiovascular disease.

This dataset is a HIPAA de-identified subset of the 40-year data. It consists of measurements of 9 variables on 4699 patients who were free of coronary heart disease at their baseline exam.

Coding Manual

Position	Variable	Variable Label	Codes
1.	id	Subject id	
2.	sex	Sex	1 = Men
			2 = Women
3.	sbp	Systolic blood pressure, mm Hg	
4.	scl	Serum cholesterol, mg/100 ml	
5.	age	Age in Years	
6.	bmi	Body mass index, kg/m^2	

1.
Introduction to Stata for Graphs

a.   Choose Your Scheme

The Stata command **scheme** sets the **overall appearance** of your graph. This has to do with whether or not there is a box around your plot, whether or not there is shading, the color of the lines and bars, etc.

The default scheme is s2color.

There are two ways to set the graph scheme

> **Method 1**: Using the **set scheme** command prior to specifying your graph
> set scheme *schemename*
> **Example:**   set scheme lean1
>
> **Method 2**: Using the graph option **scheme( )** as an option (after the comma) within your graph command
> **, scheme(*schemename*)**
> **Example: , scheme(lean1)**

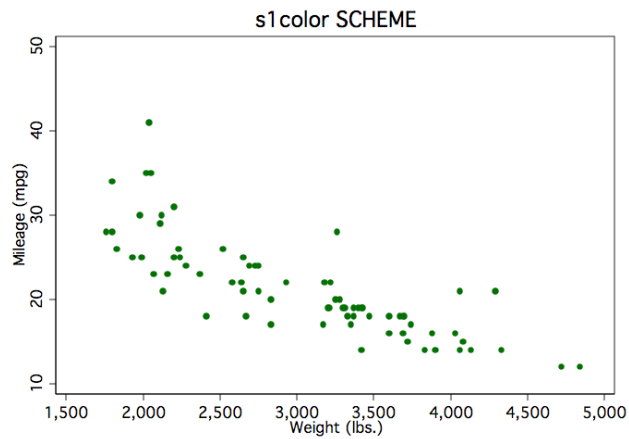Three Graph Schemes to Consider (there are lots of others, but these are for another day)

Default is s2color (no changes made yet)

```
. * DEFAULT SCHEME
. scatter mpg weight,title("DEFAULT SCHEME") xlabel(1500(500)5000) ylabel(10(10)50) msymbol(o)
```
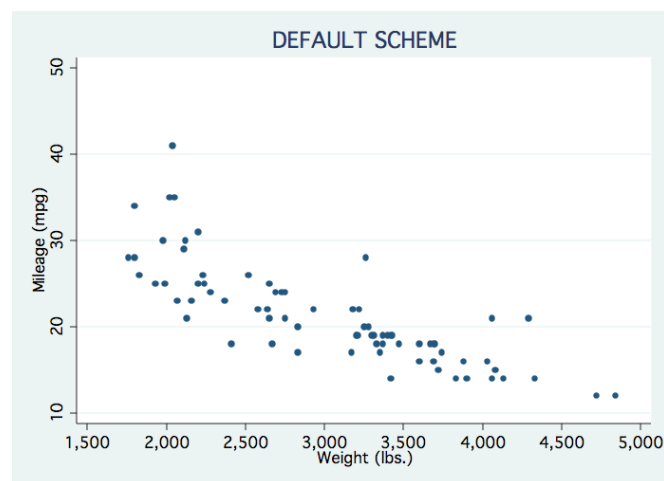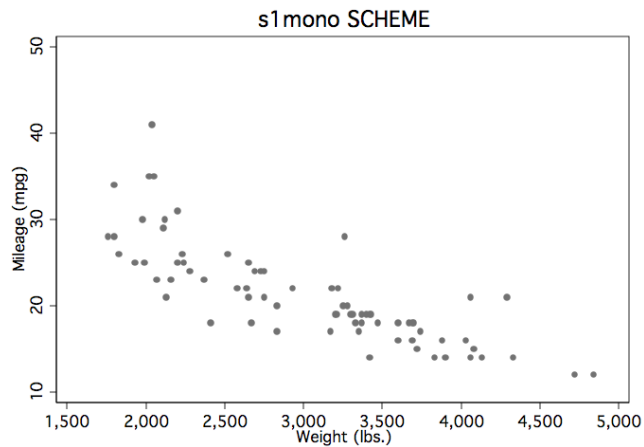
s1color

```
. * s1color SCHEME
. set scheme s1color
. scatter mpg weight,title("s1color SCHEME") xlabel(1500(500)5000) ylabel(10(10)50) msymbol(o)
```
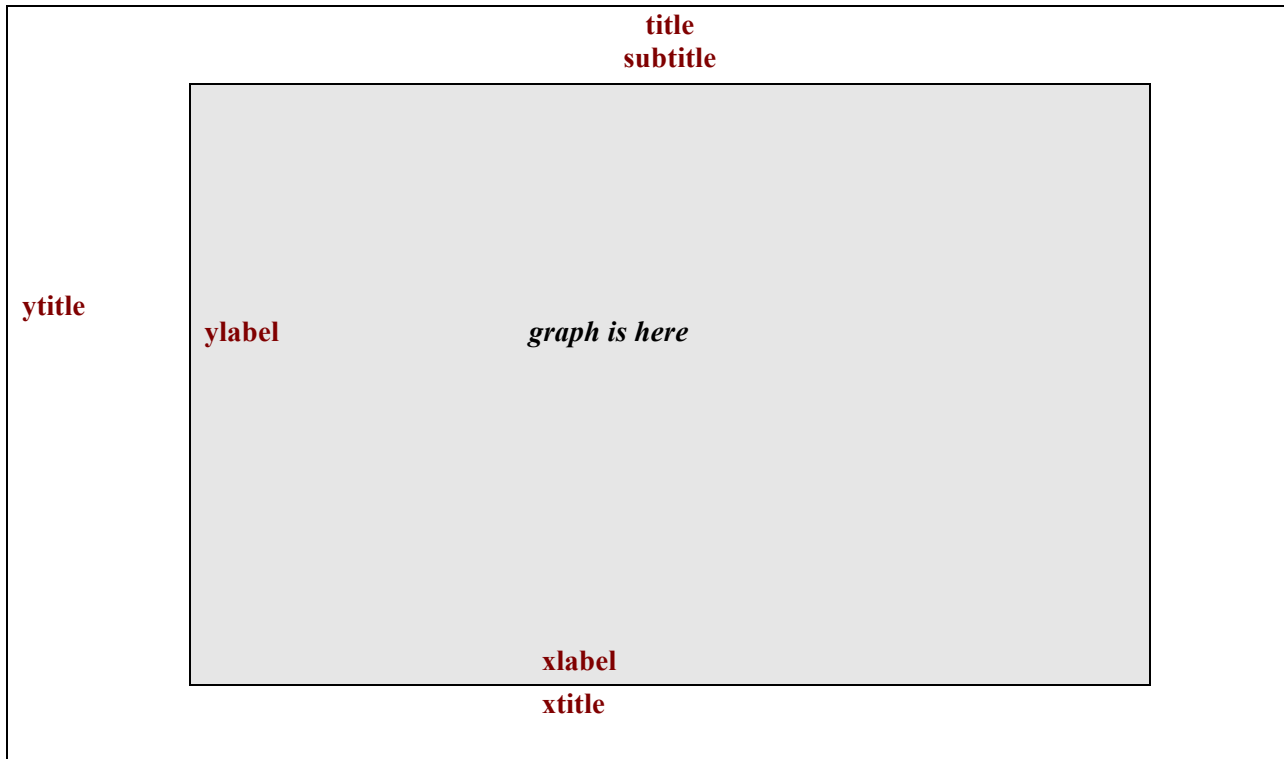

s1mono

```
. * s1mono
. set scheme s1mono
. scatter mpg weight,title("s1mono SCHEME") xlabel(1500(500)5000) ylabel(10(10)50) msymbol(o)
```

b. Architecture of Graphs in Stata

A Stata graph is comprised of: (1) the actual graph; (2) plot options (**eg – xlabel**) ; and (2) graph options (eg – title)

Schematic (*partial*) of Stata Graph Specifications

<div style="border:1px solid black; padding:20px; text-align:center;">

title
subtitle

ytitle

ylabel   *graph is here*

xlabel
xtitle

</div>

<div style="background:#f5c089; padding:20px; text-align:center;">

Tip!

Keep this page handy. When you get a little further along and are doing aesthetics (setting titles, labels, etc) this schematic will remind you of the STATA naming conventions.

</div>

c. Basic Syntax of a Stata Graph Command

.graph *graphchoice (plot_choice, plot_options) (plot_choice, plot_options)*, *graph_options*

↑   ↑   ↑

Graph options: **Note this comma!**   **Note this comma!**   **Note this comma!**

Partial listing …

             title(*"title in quotes"*)   **-   specify title**
       subtitle(*"subtitle in quotes"*)   **-   specify subtitle**
      ytitle(*"Y-axis title in quotes"*)   **-   specify Y-axis title**
      xtitle(*"X-axis title in quotes"*)   **-   specify X-axis title**
       legend (*"legend in quotes"*)   **-   specify legend**
     caption(*"caption in quotes"*)   **-   specify caption**
         note(*"note in quotes"*)   **-   specify note**

Beware! It is <u>not</u> always necessary to type "graph" as the first word in the command line. In fact, sometimes, it is incorrect. See examples below.

Example
.graph twoway (scatter mpg weight, msymbol(d)), title("Scatterplot of MPG by Weight")

       ↑   ↑   ↑   ↑   ↑   ↑

   Graph choice   plot choice   yvar   xvar   plot option   graph option

                                comma   **comma**
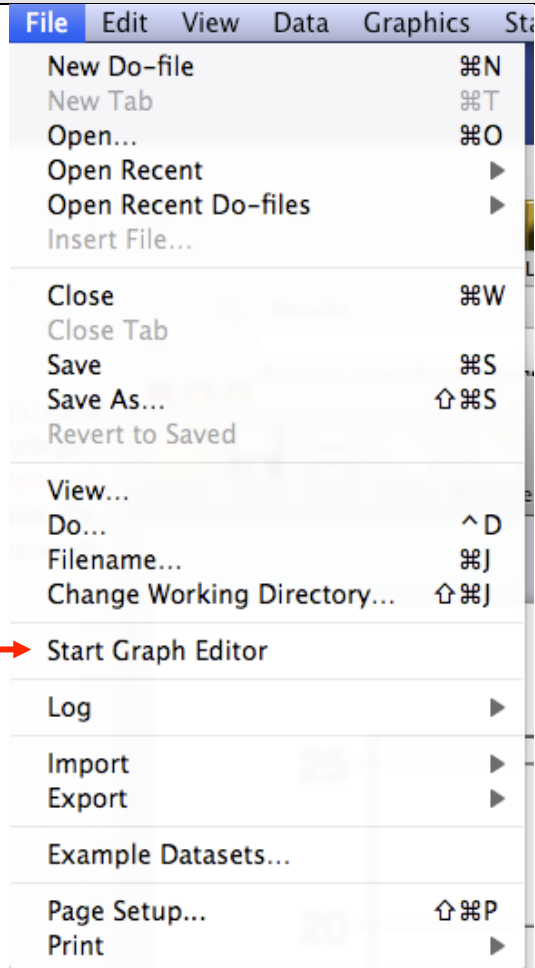
Important Tips to Remember!
Pay attention to spaces:

(1) There MUST be a space between "twoway" and the following parenthesis
(2) There must NOT be a space between "title" and the opening parenthesis that follows.

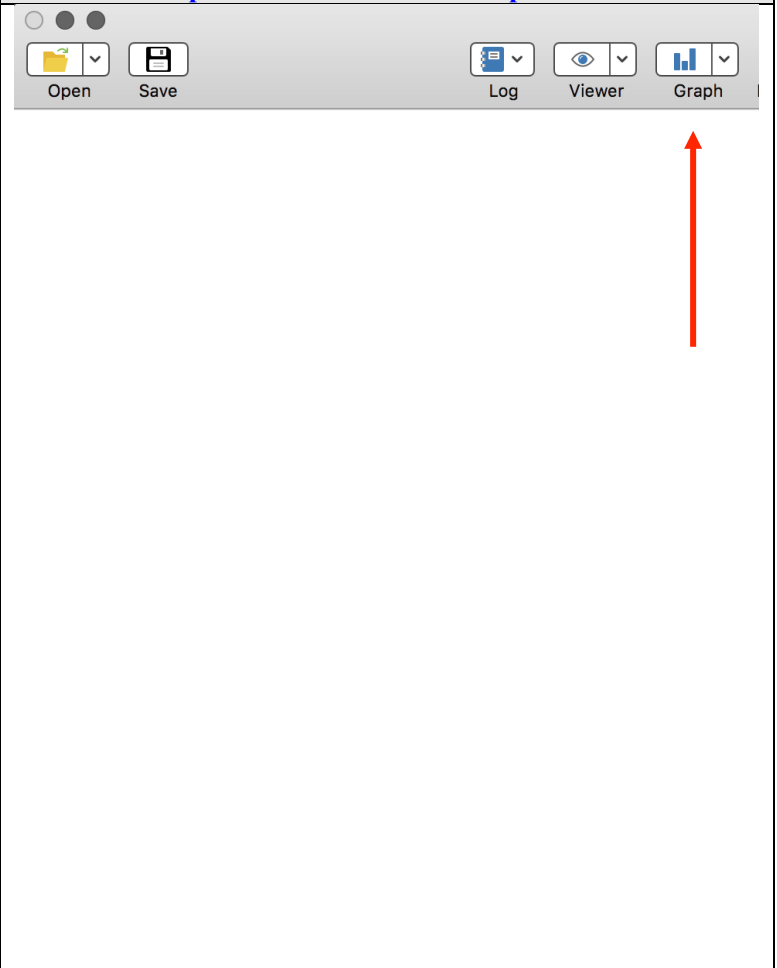### d.   Use the Graph Editor To Change the Looks of Your Graph

There are 2 ways to launch the graph editor

Method #1 - From the main menu bar:	Method #2 - From the Graph Editor Icon in the Graph Itself
File Edit View Data Graphics Sta New Do-file ⌘N New Tab ⌘T Open... ⌘O Open Recent ▶ Open Recent Do-files ▶ Insert File... Close ⌘W Close Tab Save ⌘S Save As... ⇧⌘S Revert to Saved View... Do... ^D Filename... ⌘J Change Working Directory... ⇧⌘J → Start Graph Editor Log ▶ Import ▶ Export ▶ Example Datasets... Page Setup... ⇧⌘P Print ▶	Open Save Log Viewer Graph

Key to Graph Editor Commands and Icons

Located at lower left

	Pointer Tool	Use this to select, drag, or modify the properties of an object. eg – Select your title. Then, holding the left mouse button, drag it to another position on the graph
	Add Text Tool	How to: (1) Select the "add text tool" (2) Click on the spot in your graph where you want to add text (3) A dialog box will appear (4) Type in your text. (5) If need be, use the pointer tool again to move your text to a better location.
	Add Line Tool	How to: (1) Select the "add line tool" (2) Click on the spot in your graph where you want the line to start (3) Holding the left mouse button, drag the line to where you want it to end. (4) Release the mouse.
	Add Marker Tool	Use this to add markers. The "how to" is similar to those for the "add text" and "add line" tools.
	Grid Edit Tool	Stay away from this for now….

Located at right

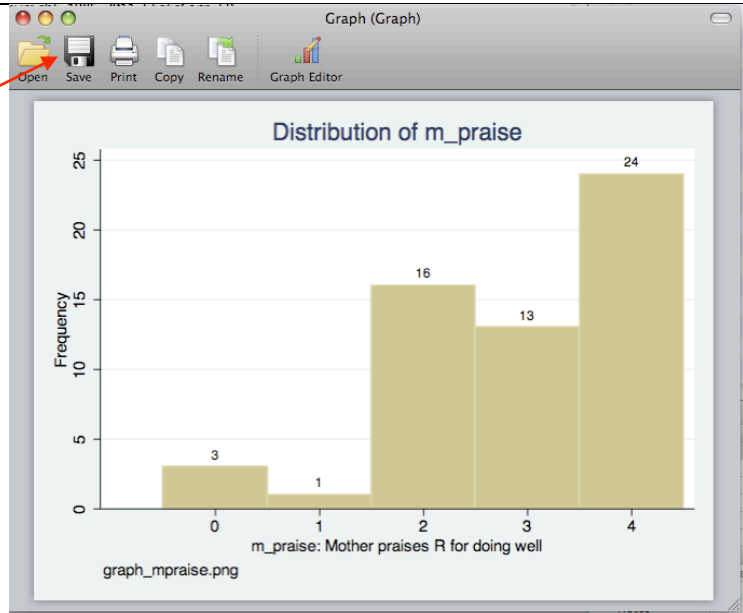|	
▼Graph ▶plotregion1 ▶yaxis1 ▶xaxis1 ▶legend ▶positional titles note caption subtitle title	This is a **series of drop down menus** from which you can modify the appearance of your plot region, titles, axes, etc.

Tip!
Use Right-Click!

You can **right click** on any object in your graph. **Try it!** When you do a drop down menu appears. It contains some very handy options, typically: (1) **hide** (2) **show** (2) **lock** (4) **unlock**

e.   Save Your Graph

Tip! Save your graph with the extension **".png"**

Step 1 **– Click anywhere in the graph to make it active. Click on SAVE icon.**	
Step 2 **– (1) At SAVE AS: type graph name without the extension, (2) At WHERE: choose directory location, (3) At FILE FORMAT drop down menu, choose "portable network graphics (recommended). Click on SAVE icon**	
Step 3 **– SAVE**	

<div align="center">

2.
Preliminaries

</div>

Before You Begin: Be sure to have downloaded from the course website: *framingham.dta*. Place in our working directory.

```
. * -----  Preliminaries -----*
. set more off

. * set working directory to desktop (yours will be different than mine) using command cd
. cd "/Users/cbigelow/Desktop"
/Users/cbigelow/Desktop

. * check working directory specification using command pwd
. pwd
/Users/cbigelow/Desktop

. * -----  Read in Stata data set framingham.dta using drop down menus  ---*
. * FILE > OPEN .. navigate to desktop .. select framingham.dta.  Click OPEN
. * You should then see in the command window
. use "/Users/cbigelow/Desktop/framingham_1000.dta"


. * Check
. codebook, compact

Variable    Obs Unique    Mean    Min   Max  Label
-------------------------------------------------------------------------------------------------
sex        1000      2    1.557     1     2  Sex
sbp        1000     87   132.35    80   270  Systolic Blood Pressure
scl         996    182  227.8464   115   493  Serum Cholesterol
age        1000     36   45.922    30    66  Age in Years
bmi         998    186  25.56623  16.4  43.4  Body Mass Index
id         1000   1000  2410.031     1  4697  Subject id
-------------------------------------------------------------------------------------------------


. * Descriptives on the discrete variables used in this illustration
. * Following assumes that you have already done (one time) ssc install fre
. fre sex

sex -- Sex
-------------------------------------------------------------
                |    Freq.    Percent    Valid     Cum.
----------------+--------------------------------------------
Valid  1 Men    |      443     44.30     44.30     44.30
       2 Women  |      557     55.70     55.70    100.00
       Total    |     1000    100.00    100.00
-------------------------------------------------------------

. * Selected descriptives on continuous variables used in this illustration
. tabstat bmi age, col(stat) statistics(n mean min max)

    variable |        N      mean       min       max
-------------+----------------------------------------
         bmi |      998  25.56623      16.4      43.4
         age |     1000    45.922        30        66
-----------------------------------------------------
```

3.
# Single Variable Graphs

__3a.   Discrete: Bar Chart

```
. * Basic
. histogram sex, discrete
(start=1, width=1)

. * Fancy
. * Notes:  (1) I set the scheme to s1color because I like it better; (2) in xlabel I tricked things to
. * obtain centering; and (3) I used a caption so as to show the name I gave to the graph

. set scheme s1color
. histogram sex, discrete bcolor(blue) frequency gap(10) xlabel(0 " " 1 "Men" 2 "Women" 3 " ")
title("Framingham Heart Study") subtitle("Bar Chart of SEX") caption("bar_fancy.png")
(start=1, width=1)

. * Save graph using drop down menu.   You should then see in the command window:
. graph export "/Users/cbigelow/Desktop/bar_fancy.png", as(png) name("Graph")
(file /Users/cbigelow/Desktop/bar_fancy.png written in PNG format)
```

Basic	Fancy
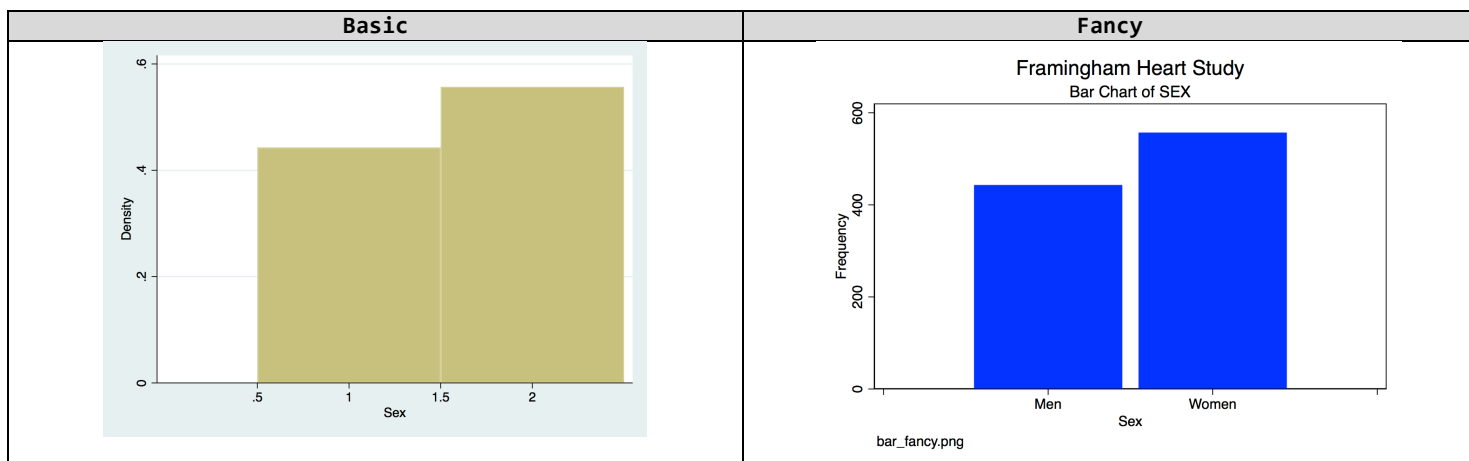
__3b.   Continuous: Histogram (I added an overlay normal for fun!)

```
. * BASIC
. histogram bmi
(bin=29, start=16.4, width=.93103455)


. * FANCY
. histogram bmi, width(1) bcolor(blue) frequency normal xlabel(15(5)45) title("Framingham Heart Study")
subtitle("Histogram of Body Mass Index") caption("histogram_fancy.png")
(bin=28, start=16.4, width=1)

. * Save graph using drop down menu.   You should then see in the command window:
. graph export "/Users/cbigelow/Desktop/histogram_fancy.png", as(png) name("Graph")
(file /Users/cbigelow/Desktop/histogram_fancy.png written in PNG format)
```

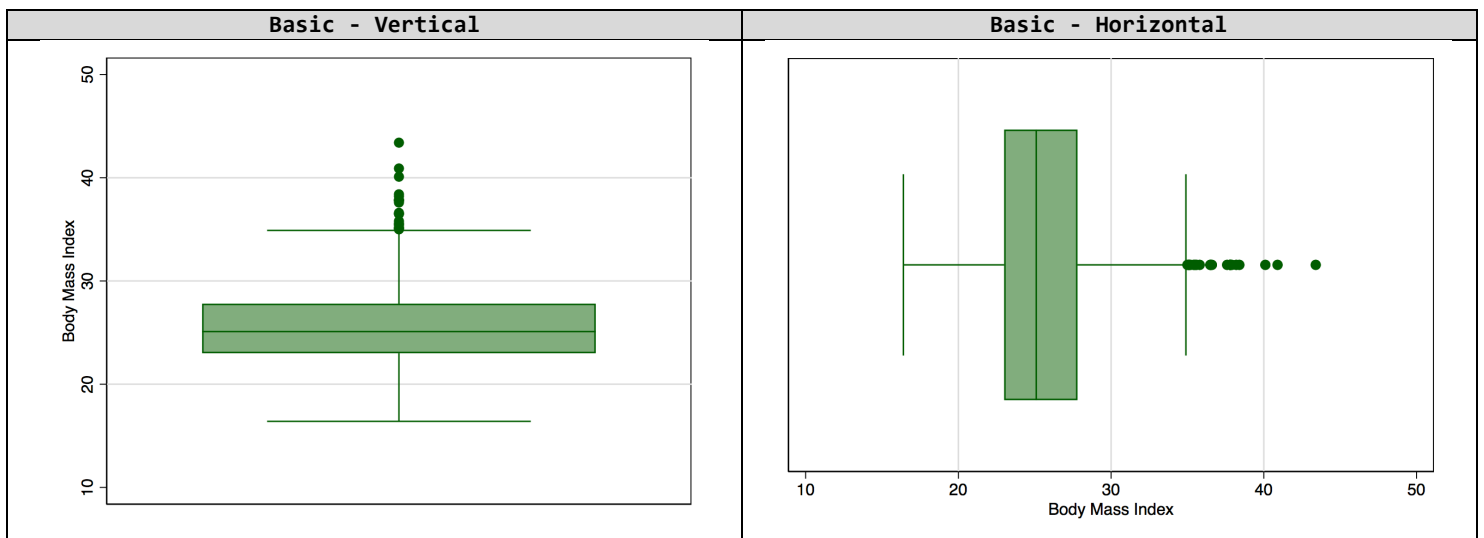| Basic	Fancy

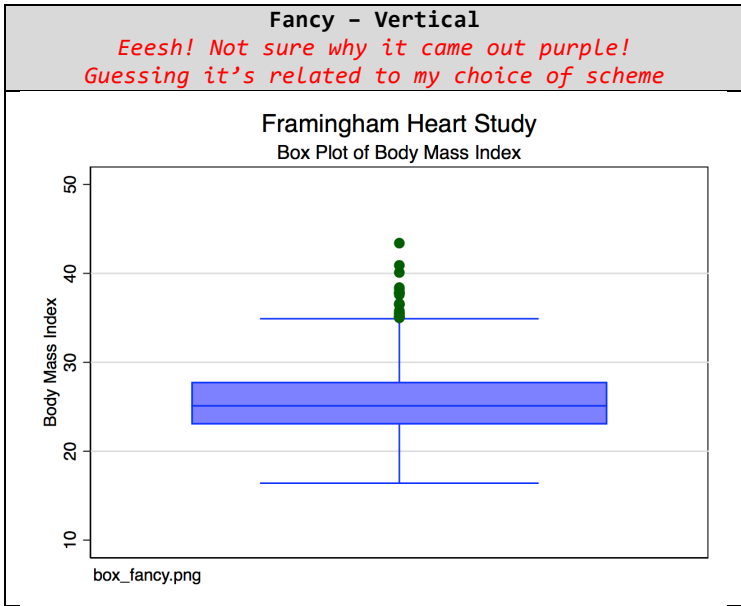histogram_fancy.png
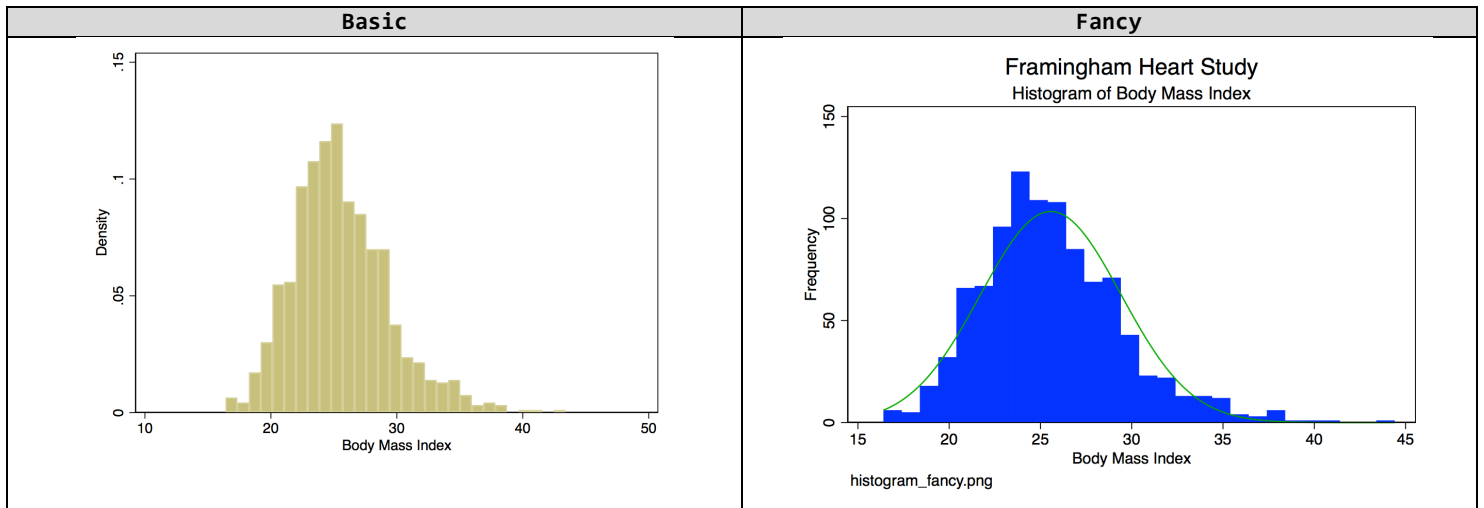
__3c. Continuous: Box Plot

```
. * BASIC - Vertical
. graph box bmi

. * BASIC - Horizontal
. graph hbox bmi

. * FANCY - Vertical
. graph box bmi, box(1,color(blue)) title("Framingham Heart Study") subtitle("Box Plot of Body Mass Index")
caption("box_fancy.png")
```

Basic - Vertical	Basic - Horizontal

Fancy – Vertical
Eeesh! Not sure why it came out purple!
Guessing it's related to my choice of scheme

Framingham Heart Study

Box Plot of Body Mass Index

box_fancy.png

<div align="center">

4.
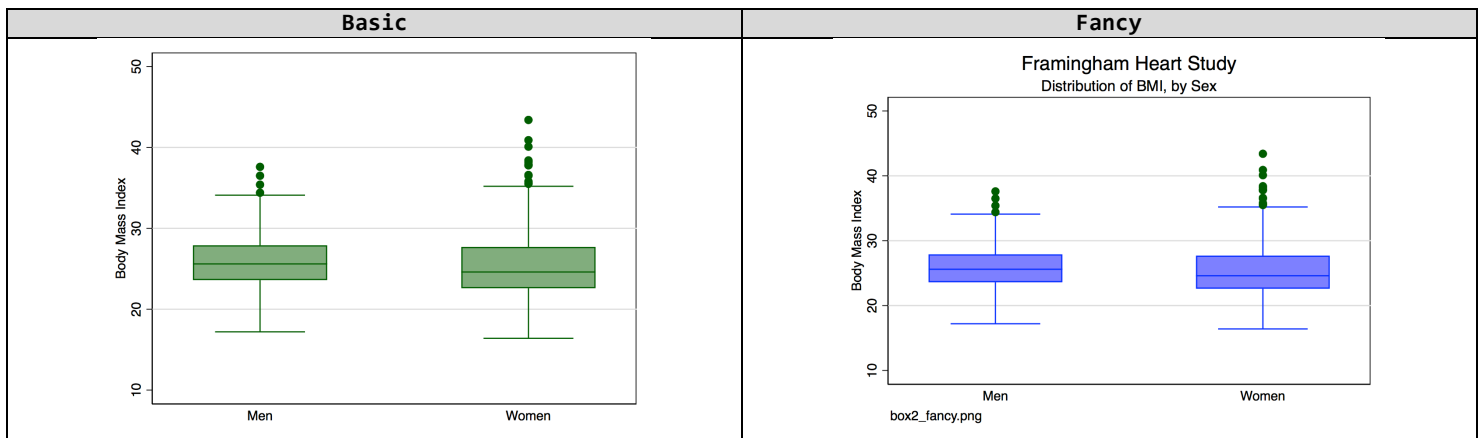**Multiple Variable Graphs**

</div>

__4a.   Continuous, by Group (Discrete): Side-by-Side Box Plot

`. sort sex`

`. * BASIC`
`. graph box bmi, over(sex)`

`. * FANCY`
`. graph box bmi, over(sex) box(1,color(blue)) title("Framingham Heart Study") subtitle("Distribution of BMI, by Sex") caption("box2_fancy.png")`
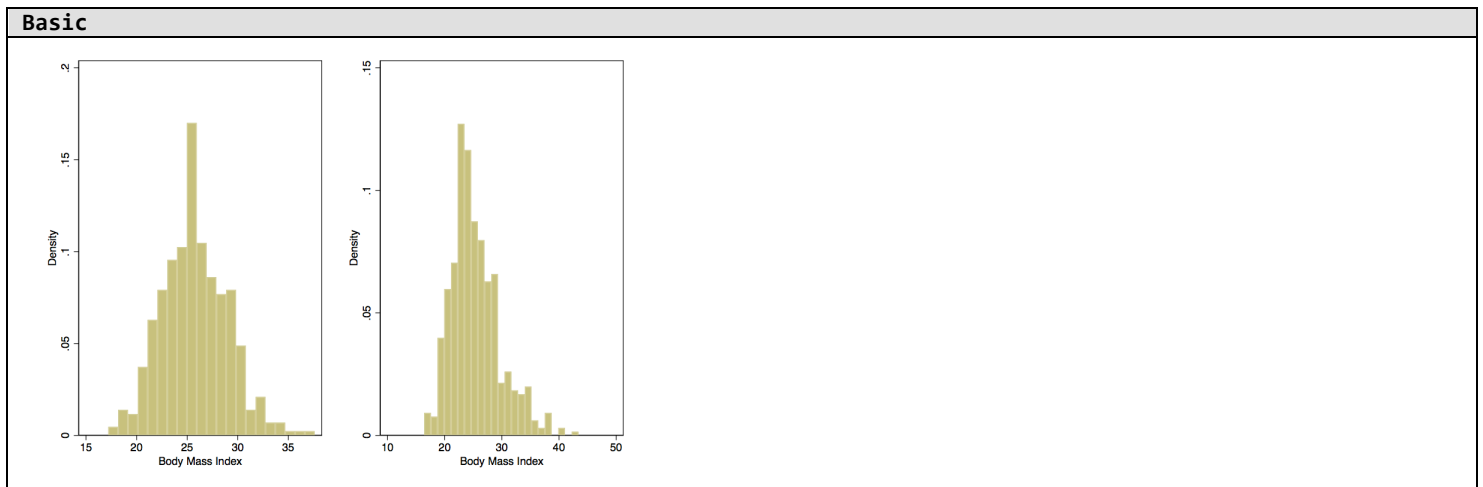
Basic	Fancy

_4b. Continuous, by Group (Discrete): Side-by-Side Histogram

. * **BASIC** NOTE: This "basic" is really a poor choice because if you look: the axes are not the same
. **histogram bmi if sex==1, name(men1,replace)**
(bin=21, start=17.200001, width=.97142846)
. **histogram bmi if sex==2, name(women1, replace)**
(bin=23, start=16.4, width=1.1739131)
. **graph combine men1 women1**

```
Basic
```


. * **FANCY** IMPORTANT: Don't forget to define your X and Y axes exactly the same!
. **histogram bmi if sex==1, width(1) bcolor(blue) frequency normal xlabel(15(5)45) ylabel(0(20)80)**
subtitle("Men") name(men2, replace)
(bin=21, start=17.200001, width=1)
. **histogram bmi if sex==2, width(1) bcolor(blue) frequency normal xlabel(15(5)45) ylabel(0(20)80)**
subtitle("Women") name(women2, replace)
(bin=28, start=16.4, width=1)
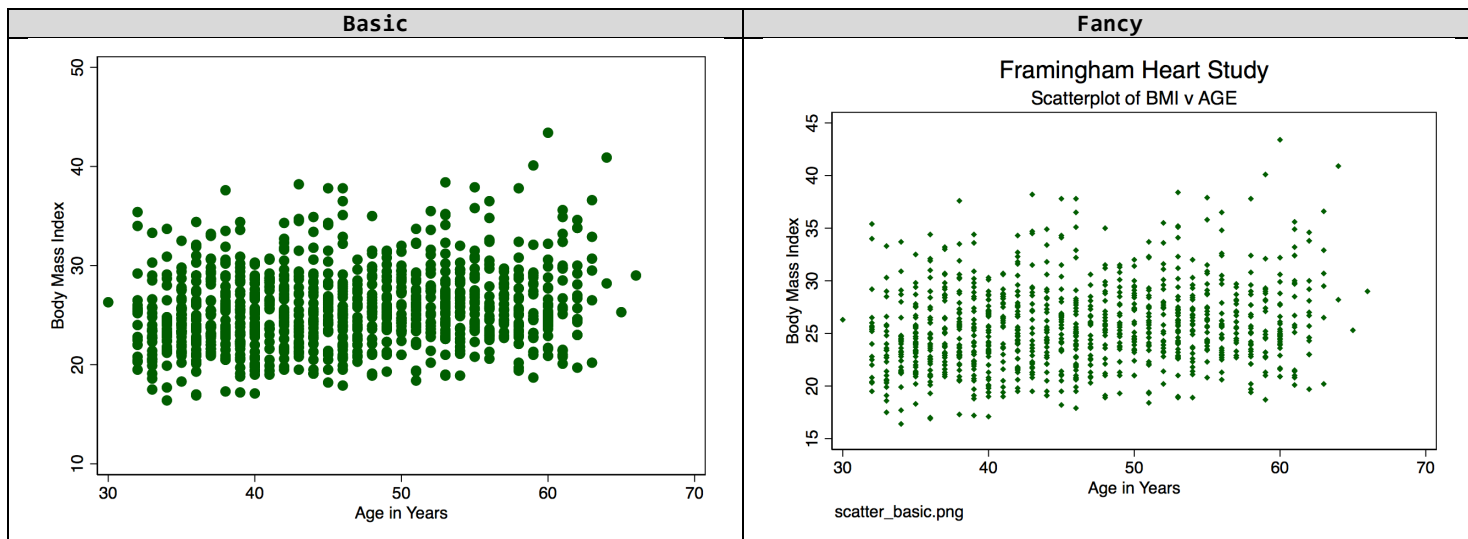. **graph combine men2 women2, title("Framingham Heart Study: Distribution of Body Mass Index")**

```
Fancy
```
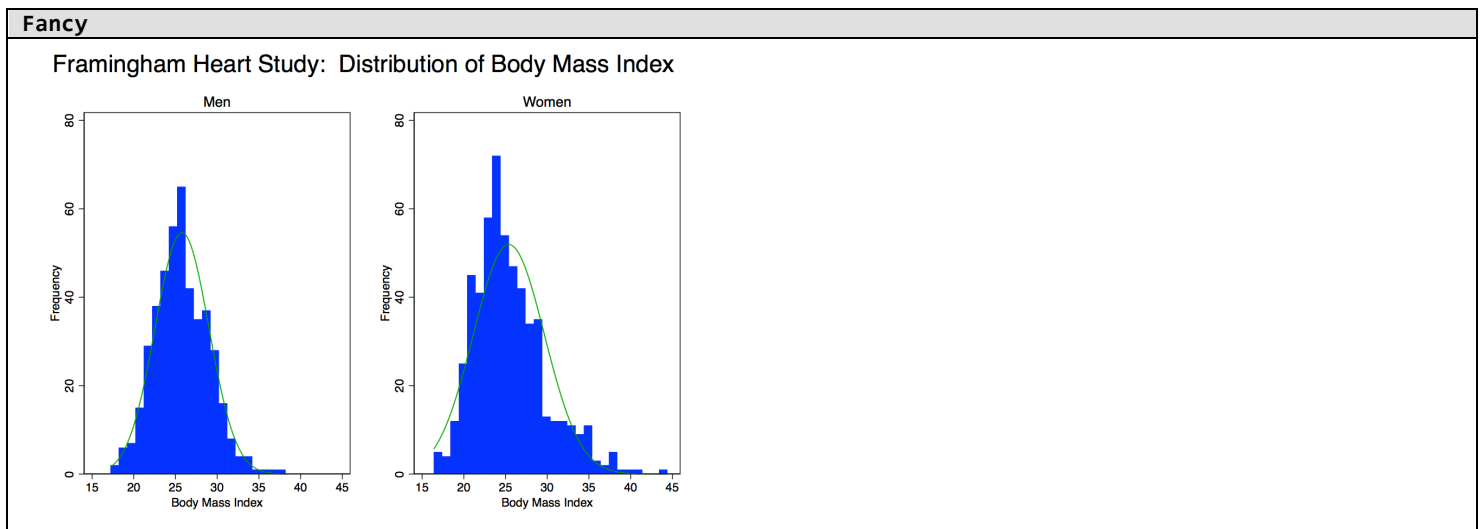
__4c.   Continuous: X-Y Plot (Scatterplot)

```
. * BASIC
. graph twoway (scatter bmi age)
```

```
. * FANCY
. graph twoway (scatter bmi age, symbol(d) msize(vsmall)), title("Framingham Heart Study") xlabel(30(10)70)
ylabel(15(5)45) subtitle("Scatterplot of BMI v AGE") caption("scatter_basic.png")
```


## __4d.   Continuous: X-Y Plot (Scatterplot), with Overlay Linear Regression Model Fit

```
. * IMPORTANT TIP!
. * When doing overlay plots, take care to plot the data points last so that they appear on top of the fit
```

```
. * BASIC
. graph twoway (lfitci bmi age) (scatter bmi age)
```

```
. * FANCY
. graph twoway (lfitci bmi age) (scatter bmi age, symbol(d) msize(vsmall)), title("Framingham Heart Study")
xlabel(30(10)70) ylabel(15(5)45) subtitle("Scatterplot of BMI v AGE w Fitted Linear Regression and 95% CI")
legend(off) caption("scatterline_fancy.png")
```

Basic	Fancy

scatterline_fancy.png
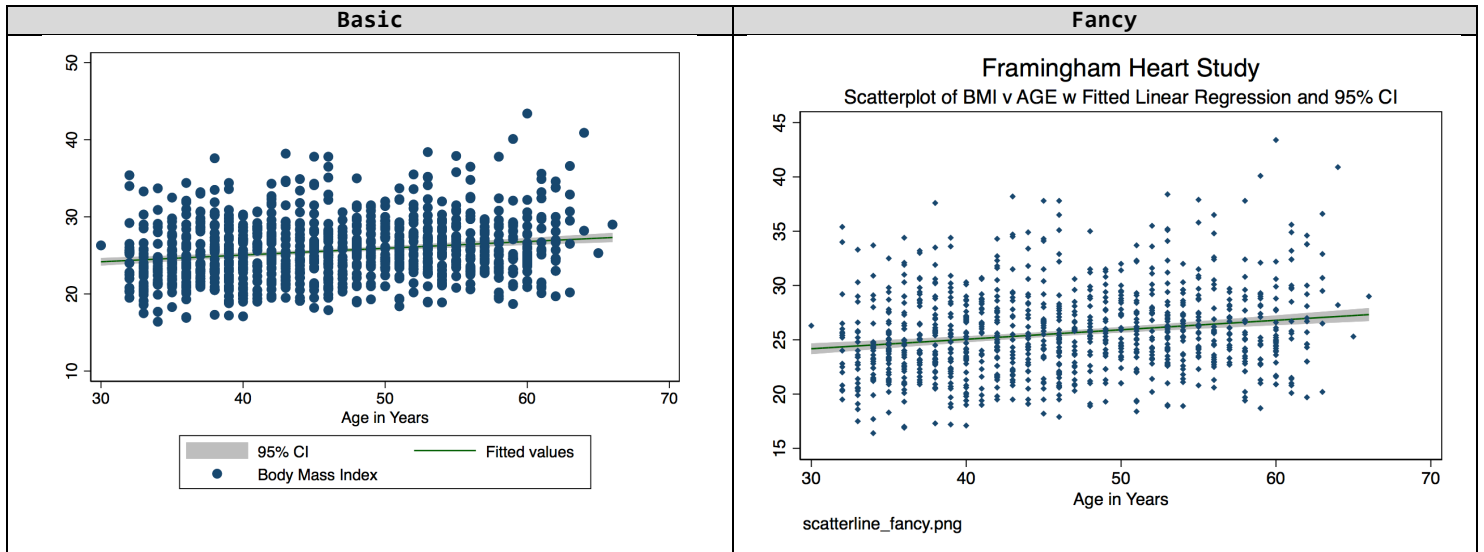
__4e.   Continuous: X-Y Plot, by Group (Discrete)

. * FANCY only
. graph twoway (scatter bmi age if sex==1, symbol(D) mcolor(navy) msize(vsmall)) (scatter bmi age if sex==2, symbol(Oh) mcolor(red) msize(vsmall)), title("Framingham Heart Study") xlabel(30(10)70) ylabel(15(5)45) legend(label(1 Men) label(2 Women)) subtitle("Scatterplot of BMI v AGE") caption("scatter2_fancy.png")

scatter2_fancy.png