Stata Program Notes
Biostatistics: A Guide to Design, Analysis, and Discovery
Chapter 3: Descriptive Statistics

Before discussing the program notes for chapter 3, we present data on 40 participants from the Digoxin clinical trial. We refer to the dataset as the **DIG40** dataset throughout the Program Notes. The **DIG40** dataset is presented in Table 3.1 of the textbook. The command below was used to call the data directly into Stata:

Stata Commmand:

```
use http://www.biostat-edu.com/files/dig40.dta, clear
```

From here on, we can use the Stata command **list (l)** to display variable values as shown below. Later in the Program Notes, we show how to obtain the web address where the data is stored on our website.

Stata Command:

```
list id trtmt age race sex bmi creat sysbp
```

Stata Output:

```
     +-------------------------------------------------------------+
     |   id   trtmt   age   race   sex      bmi   creat   sysbp |
     |-------------------------------------------------------------|
  1. | 2289       0    76      1     1   30.586     1.7     130 |
  2. | 6745       0    45      1     1    22.85   1.398     130 |
  3. | 1322       1    45      1     2   43.269      .9     115 |
  4. |  538       1    31      1     1   27.025   1.159     120 |
  5. |  999       1    47      1     2   30.506   1.386     120 |
     |-------------------------------------------------------------|
  6. | 3103       0    60      1     1   29.867   1.091     140 |
  7. | 1954       1    77      1     1   26.545   1.307     140 |
  8. | 5750       1    76      1     1   39.837   1.455     140 |
  9. | 1109       0    68      1     2   27.532   1.534     144 |
 10. | 4787       1    46      1     1   28.662   1.307     140 |
     |-------------------------------------------------------------|
 11. |  666       0    65      1     1   28.058       2     120 |
 12. | 6396       0    83      1     1   26.156   1.489     116 |
 13. | 5753       1    75      1     1    37.59     1.3     138 |
 14. | 1882       0    50      1     1   25.712   1.034     140 |
 15. | 5663       0    59      2     1   27.406   1.705     152 |
     |-------------------------------------------------------------|
 16. | 6719       1    34      1     1   20.426   1.886     116 |
 17. | 4995       0    55      1     1   19.435     1.6     150 |
 18. | 4055       0    71      1     1   22.229   1.261     100 |
 19. | 4554       1    58      1     2   28.192   1.352     130 |
 20. | 2217       1    65      1     1   23.739   1.614     170 |
     |-------------------------------------------------------------|
 21. |  896       0    50      1     1   27.406     1.3     140 |
 22. | 5368       1    38      1     1   30.853      .9     134 |
 23. | 3403       0    55      1     2    21.79    1.17     130 |
 24. | 1426       0    70      1     1    19.04    1.25     150 |
 25. |  764       1    63      2     2   28.731      .9     122 |
```

```
     |---------------------------------------------------------|
 26. | 5668       0    74      1      1   29.024   1.227      116 |
 27. | 1653       1    63      1      1   28.399     1.1      105 |
 28. | 1254       1    73      1      1   26.545     1.3      144 |
 29. | 2312       0    78      2      1   22.503   2.682      104 |
 30. | 2705       1    66      1      2   28.762      .9      150 |
     |---------------------------------------------------------|
 31. | 4181       0    44      2      2    26.37   1.148      124 |
 32. | 3641       0    64      1      1   21.228      .9      130 |
 33. | 2439       1    49      1      1   15.204   1.307      140 |
 34. | 3640       0    79      1      1   18.957   2.239      150 |
 35. | 6646       0    61      1      1   27.718   1.659      128 |
     |---------------------------------------------------------|
 36. |  787       0    58      2      2   27.369    .909      100 |
 37. | 5407       1    50      1      2   24.176       1      130 |
 38. | 5001       1    70      1      1   19.044     1.2      110 |
 39. | 4375       0    61      1      1   32.079   1.273      128 |
 40. | 4326       0    65      1      1    29.34     1.2      170 |
     +---------------------------------------------------------+
```

To display a portion of the data, for example only the first 5 participants, we can use the following Stata commands:

Stata Command:

```
* First 5 patients in DIG40 dataset
list id trtmt age race sex bmi creat sysbp in 1/5
```

Stata Output:

```
     +---------------------------------------------------------+
     |   id   trtmt   age   race   sex      bmi   creat   sysbp |
     |---------------------------------------------------------|
  1. | 2289       0    76      1      1   30.586     1.7      130 |
  2. | 6745       0    45      1      1    22.85   1.398      130 |
  3. | 1322       1    45      1      2   43.269      .9      115 |
  4. |  538       1    31      1      1   27.025   1.159      120 |
  5. |  999       1    47      1      2   30.506   1.386      120 |
     +---------------------------------------------------------+
```

Or, we can display the last 5 participants with the Stata commands:

Stata Command:

```
list id trtmt age race sex bmi creat sysbp in 36/40
```

Stata Output:

```
     +---------------------------------------------------------+
     |   id   trtmt   age   race   sex      bmi   creat   sysbp |
     |---------------------------------------------------------|
 36. |  787       0    58      2      2   27.369    .909      100 |
 37. | 5407       1    50      1      2   24.176       1      130 |
```

```
 38. | 5001       1    70      1     1   19.044     1.2     110 |
 39. | 4375       0    61      1     1   32.079   1.273     128 |
 40. | 4326       0    65      1     1    29.34     1.2     170 |
     +---------------------------------------------------------+
```

Use the Stata command **help list** to explore more details of the list command.

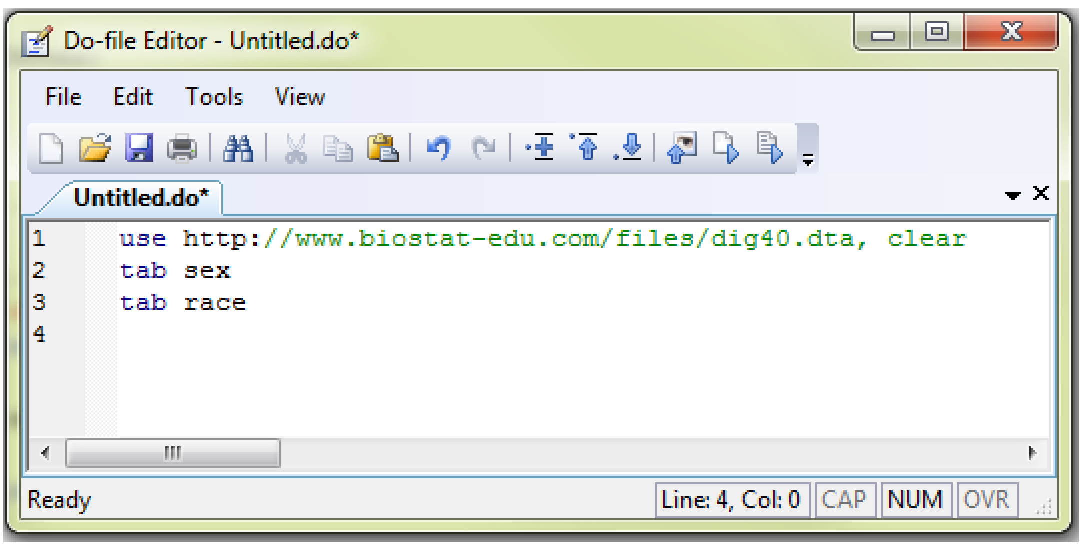
Program Note 3.1 – **Tabulating data**

As already shown above, the **list** command is useful when examining characteristics of specific individuals. By displaying the entire **DIG40** data set, we are able to see the specific values of the variables: treatment, age, race, sex and so on- basically, the characteristics of the entire set of forty participants. However, there is also a need to summarize the information provided in the data set. For example, we may want to know how many males and females are in the **DIG40** data set. This would simply require the creation of a table displaying the frequency of males and females which is referred to as a one-way table. The **tabulate (tab)** command in Stata is used to create one- and two-way tables. For example, we are able to obtain the information presented in Table 3.2 using the Stata commands below:

Stata commands:

```
use http://www.biostat-edu.com/files/dig40.dta,clear
tab sex
tab race
```

Below, we provide a screen shot of the commands and their corresponding output. The commands are provided in Stata's **Do-file Editor**.



Stata Output:

```
. tab sex

        sex |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |         30       75.00       75.00
          2 |         10       25.00      100.00
------------+-----------------------------------
      Total |         40      100.00

. tab race
```

```
        race |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |         35       87.50       87.50
          2 |          5       12.50      100.00
------------+-----------------------------------
      Total |         40      100.00
```

We don't recommend that each line of commands in the **Command Window** be typed in- simply because the **Do-file Editor allows** you to save your commands making them readily available the next time you'd like to use them. To learn more about the **Do-file Editor** and Stata in general, we recommend that you go to UCLA's Academic Technology Services website. You can get to the site using any search engine like Yahoo search, Bing, or Google; or the Statistical Resources page of our website.

Moving on; in some cases, we may need to create a categorical variable from a continuous variable. For example in Table 3.4, the continuous variable bmi is presented as a categorical variable. A new variable can be created in Stata by using the **generate (gen)** command. Since we are generating a new variable, bmi_cat, whose value is 0 if bmi is less than 18.5 kg/m^2, we also need to use the **replace** command along with the **if** statement to change the value of bmi_cat to 1 if bmi is greater than or equal to 18.5 kg/m^2 but less than 25 kg/m^2. The Stata commands are shown below.

```
Stata Commands:

use http://www.biostat-edu.com/files/dig40.dta,clear
gen bmi_cat = 0 if bmi < 18.5
replace bmi_cat = 1 if bmi >= 18.5 & bmi < 25
replace bmi_cat = 2 if bmi >= 25 & bmi < 30
replace bmi_cat = 3 if bmi >= 30
```

Notice that four ordered categories that are created. To find the cross tabulation of bmi and sex, we use the Stata command **tab** along with the option **column (col)** to display column percentages. Row percentages can be displayed using the **row (r)** option. The **cell** option displays the relative frequency of each cell in the table.

```
Stata Commands:

tab bmi_cat sex, col


Stata Output:

+-------------------+
| Key               |
|-------------------|
|     frequency     |
| column percentage |
+-------------------+

           |          SEX
  bmi_cat1 |         1          2 |     Total
-----------+----------------------+----------
         0 |         1          0 |         1
```

```
           |      3.33       0.00 |      2.50 
-----------+----------------------+----------
         1 |        10          2 |        12 
           |     33.33      20.00 |     30.00 
-----------+----------------------+----------
         2 |        14          6 |        20 
           |     46.67      60.00 |     50.00 
-----------+----------------------+----------
         3 |         5          2 |         7 
           |     16.67      20.00 |     17.50 
-----------+----------------------+----------
     Total |        30         10 |        40 
           |    100.00     100.00 |    100.00
```

Use the Stata command **help tab** to explore more details of the tabulate command.

Program Note 3.2 – Creating Line graphs and Bar charts

1. Line graphs

In Table 3.6, we present health expenditures data as a percentage of GDP by year for Canada, the United Kingdom, and the United States. The data are entered as shown below:

Stata Commands:

```
input Canada UK US year
5.4 3.9 5.1  1960
5.6 4.1 6.0  1965
7.0 4.5 7.0  1970
7.0 5.5 8.4  1975
7.1 5.6 8.8  1980
8.0 6.0 10.6 1985
9.0 6.0 12.0 1990
9.2 7.0 13.4 1995
9.2 7.3 13.3 2000
end
```

After the data have been entered, we can use the **twoway** command and the **connected** option in Stata to make line graphs. Other useful options available with the **twoway** command are:

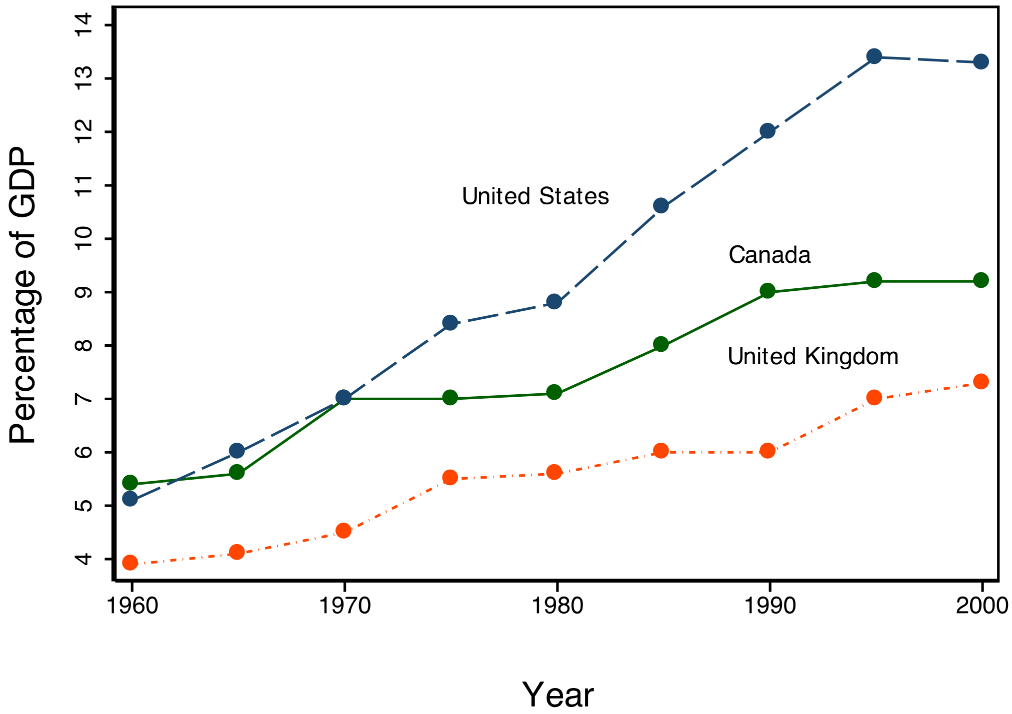
lpattern( ) allows you to specify the line pattern
legend(off) removes the appearance of a legend
titlegap( ) allows you to increase the spacing between the axes titles and the axes
text( ) allows you to insert text within the plot area
scheme( ) allows you to modify the graphs display

The Stata commands below create the line graphs in Figure 3.1:

Stata Commands:

```
twoway (connected Canada year, lpattern(solid))                          ///
       (connected UK     year, lpattern(shortdash_dot))                  ///
       (connected US     year, lpattern(longdash)),                      ///
       scheme(s1color) yscale(titlegap(4)) xscale(titlegap(7)) legend(off) ///
       text(10.8  1979 "United States")                                  ///
       text(9.7 1990 "Canada")                                           ///
       text(7.8 1992 "United Kingdom")                                   ///
       ytitle("Percentage of GDP",size(large))                           ///
       xtitle("Year",size(large))                                        ///
       ylabel(4(1)14)
```

Stata Output:



Use the Stata command **help twoway** to explore more details of the twoway command.

2. Bar charts

The Stata commands **graph bar** and **graph hbar** can be used to create vertical or horizontal bar charts. For example, the horizontal bar chart in Figure 3.5 displays the proportion of diabetes by age group for individuals in the **DIG200** data set. First, we present the Stata commands required to do the data management that creates the age categories of interest. Notice that we are naming the new variable: age_cat.

Stata Commands:

```
use http://www.biostat-edu.com/files/DIG200.dta,clear
gen age_cat = 0 if age < 40
replace age_cat = 1 if age >= 40 & age < 50
replace age_cat = 2 if age >= 50 & age < 60
replace age_cat = 3 if age >= 60 & age < 70
replace age_cat = 4 if age >= 70
```

Then we use the tab command to find the percent of diabetes for each age category. The **tab** command gives us information that could be incorporated into the graph using the **text** option.

```
Stata Commands:

tab diabetes age_cat, col
```

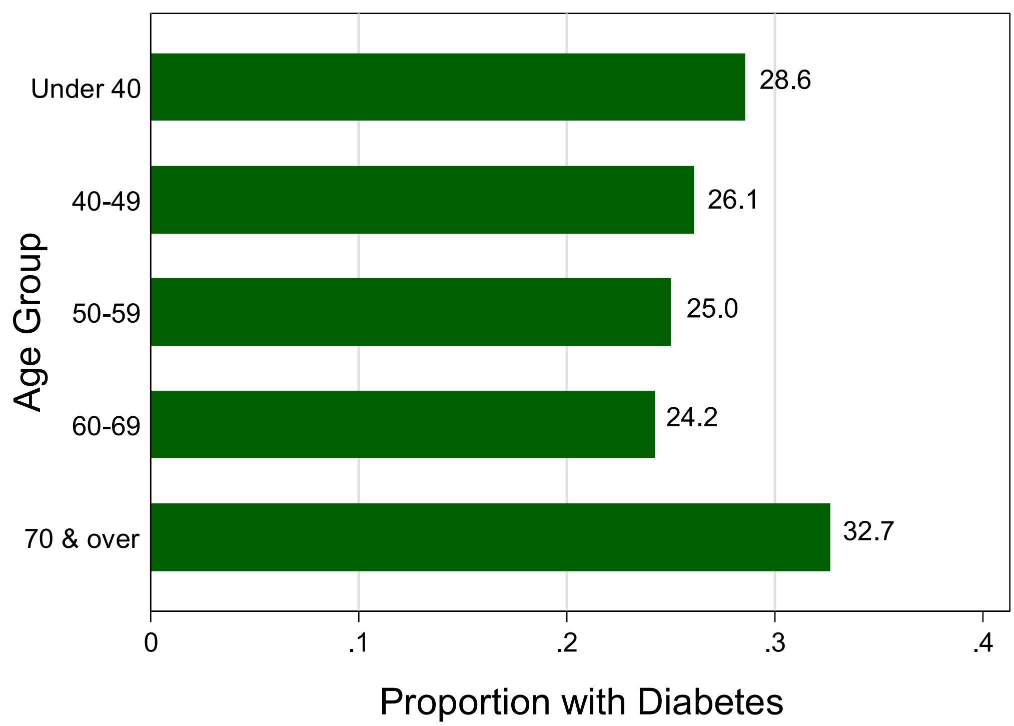
The Stata commands below allowed us to label the values of the age_cat variable:

```
Stata Commands:

* Label the values of the age categories
label define labage 0 "Under 40" 1 "40-49" 2 "50-59" 3 "60-69" 4 "70 & over"
label values age_cat labage
```

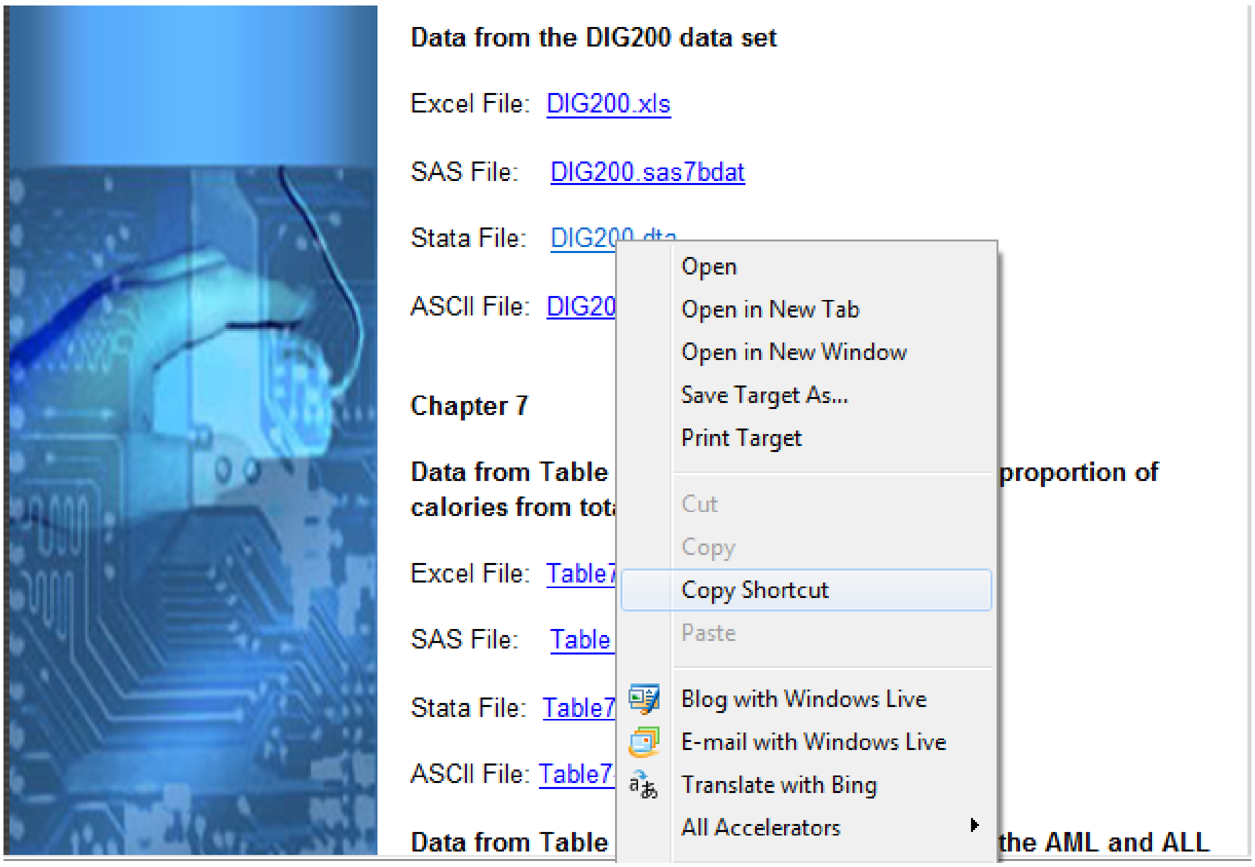
Finally, the Stata commands shown below were used to create Figure 3.5.

```
Stata Commands:

graph hbar diabetes, over(age_cat)                        ///
ytitle("Proportion with Diabetes", size(large))             ///
yscale(titlegap(4)) scheme(s1color) ylabel(0(.1).4)       ///
text(-.06 48 "Age Group",size(large) orientation(vertical))  ///
text(.305 93 "28.6") text(.28 71 "26.1") text(.27 51 "25.0") ///
text(.26 31 "24.2") text(.345 10 "32.7")
```

Stata Output:



Stata can be used to create more complicated graphs like the stacked and clustered bar charts displayed in Figures 3.6 and 3.7. Those graphs also correspond to data found in the **DIG200** dataset. As an aside, to find the web location of the **DIG200** dataset, see the screen shot below. Simply right-click on the Stata File: **DIG200.dta**, then left-click on **Copy Shortcut**. Proceed by going into your **Do-file Editor** and paste in the web address (using a right-click or Ctrl+V).



The command should look as follows after the web address for the data has been pasted in:

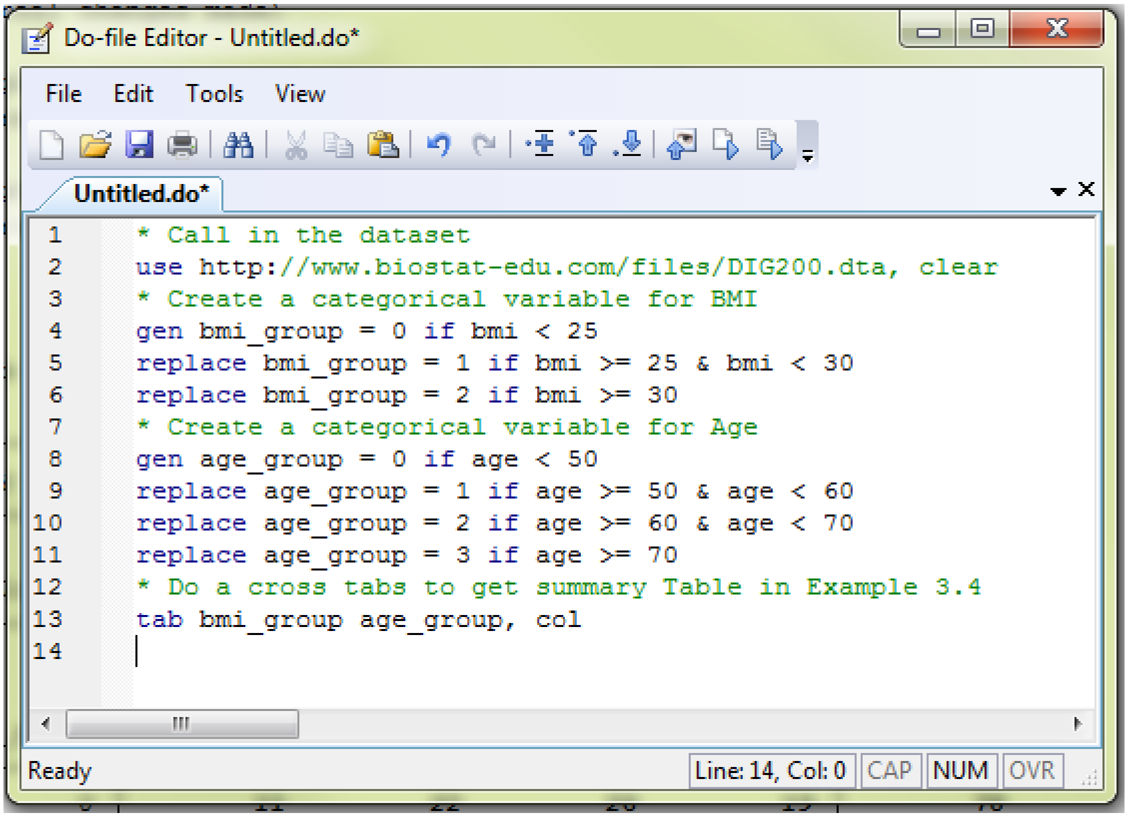
```
use http://www.biostat-edu.com/files/DIG200.dta, clear
```

Now back to creating bar charts. To create a stacked bar chart, Stata requires the **stack** option to be included in the command lines. If the **stack** option is not included then a clustered bar chart is produced. Before getting to Figures 3.6 and 3.7, we begin by showing the commands to create a new variable that categorizes the continuous variable age. This is similar to what was done for Figure 3.5. However, in this case, we aren't using as many categories for age. Also, notice that this time, we are calling the new variable: age_group.

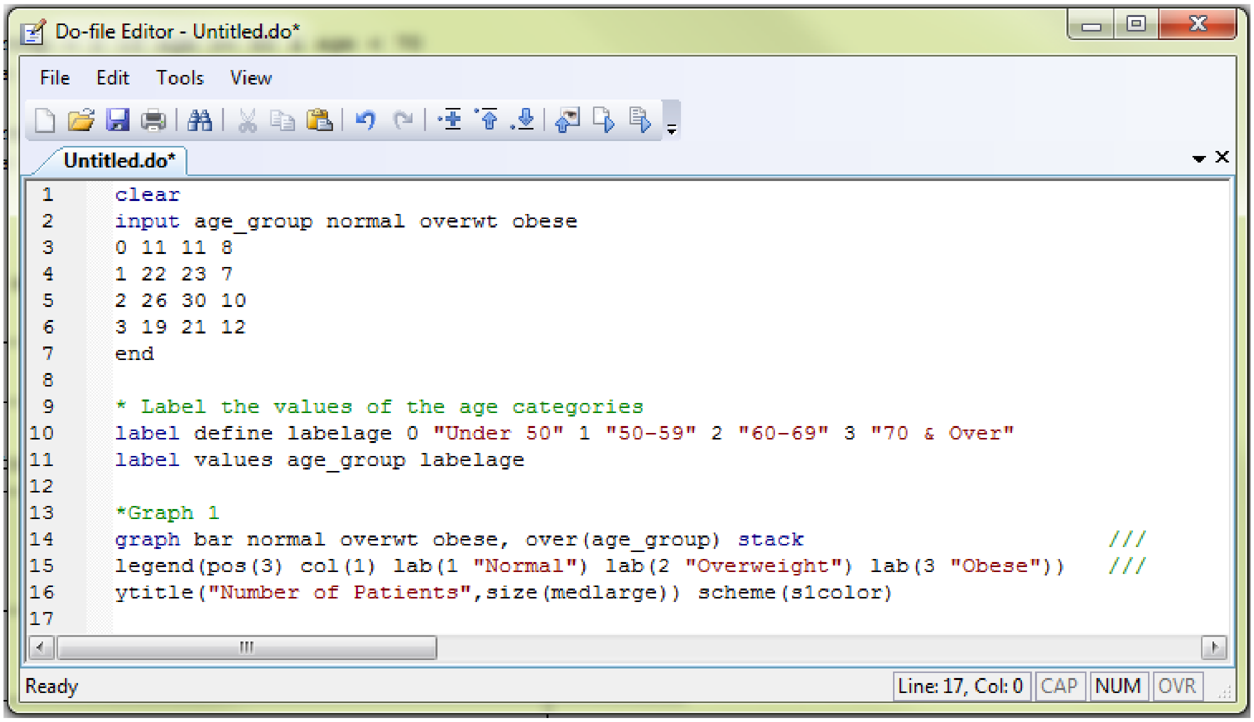
Stata Commands:

```
gen age_group = 0 if age < 50
replace age_group = 1 if age >= 50 & age < 60
replace age_group = 2 if age >= 60 & age < 70
replace age_group = 3 if age >= 70
```

The screen shot below shows the commands used to obtain the information provided in the summary table presented in Example 3.4 (page 29) of the textbook.



Once the summary data is retrieved, another copy of Stata can be launched, and the screen shot below shows the subsequent commands that were used to create Figure 3.6.



Pay close attention to the Stata commands below that allowed us to label the values of the age_group variable:

Stata Commands:

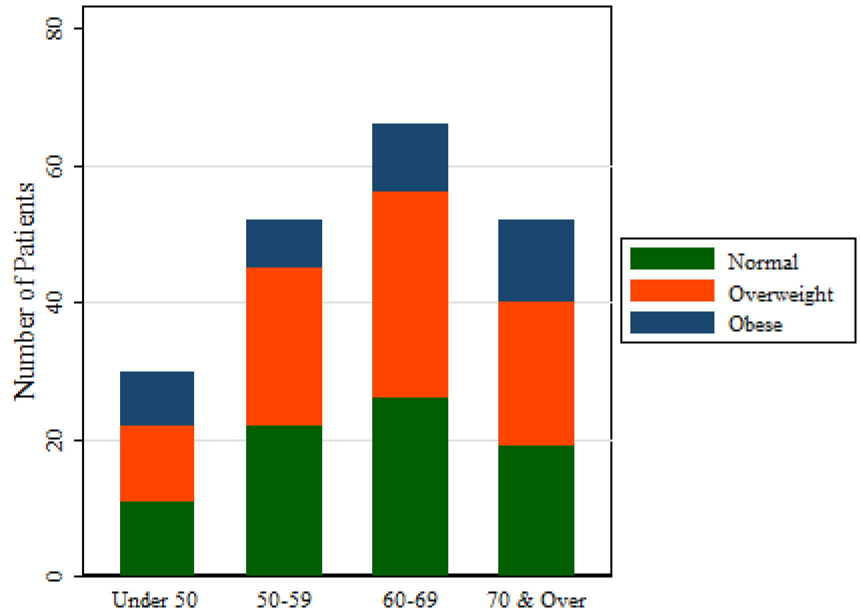
```
* Label the values of the age categories
label define labelage 0 "Under 50" 1 "50-59" 2 "60-69" 3 "70 & Over"
label values age_group labelage
```

And, the Stata commands below were used to position and label the legend corresponding to the figure.

Stata Commands:

```
*Graph 1
graph bar normal overwt obese, over(age_group) stack                    ///
legend(pos(3) col(1) lab(1 "Normal") lab(2 "Overweight") lab(3 "Obese"))   ///
ytitle("Number of Patients",size(medlarge)) scheme(s1color)
```

Stata Output:

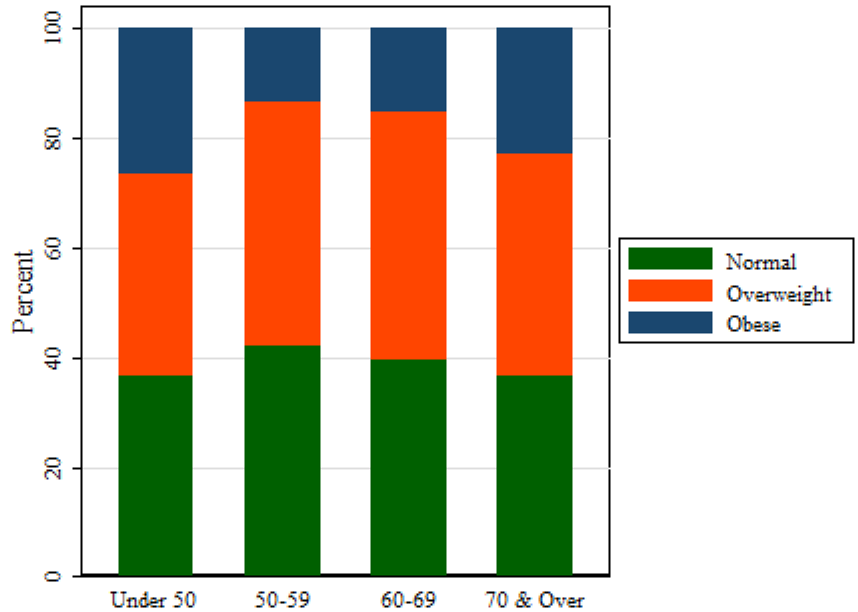


The Stata commands below created the second graph in Figure 3.6.

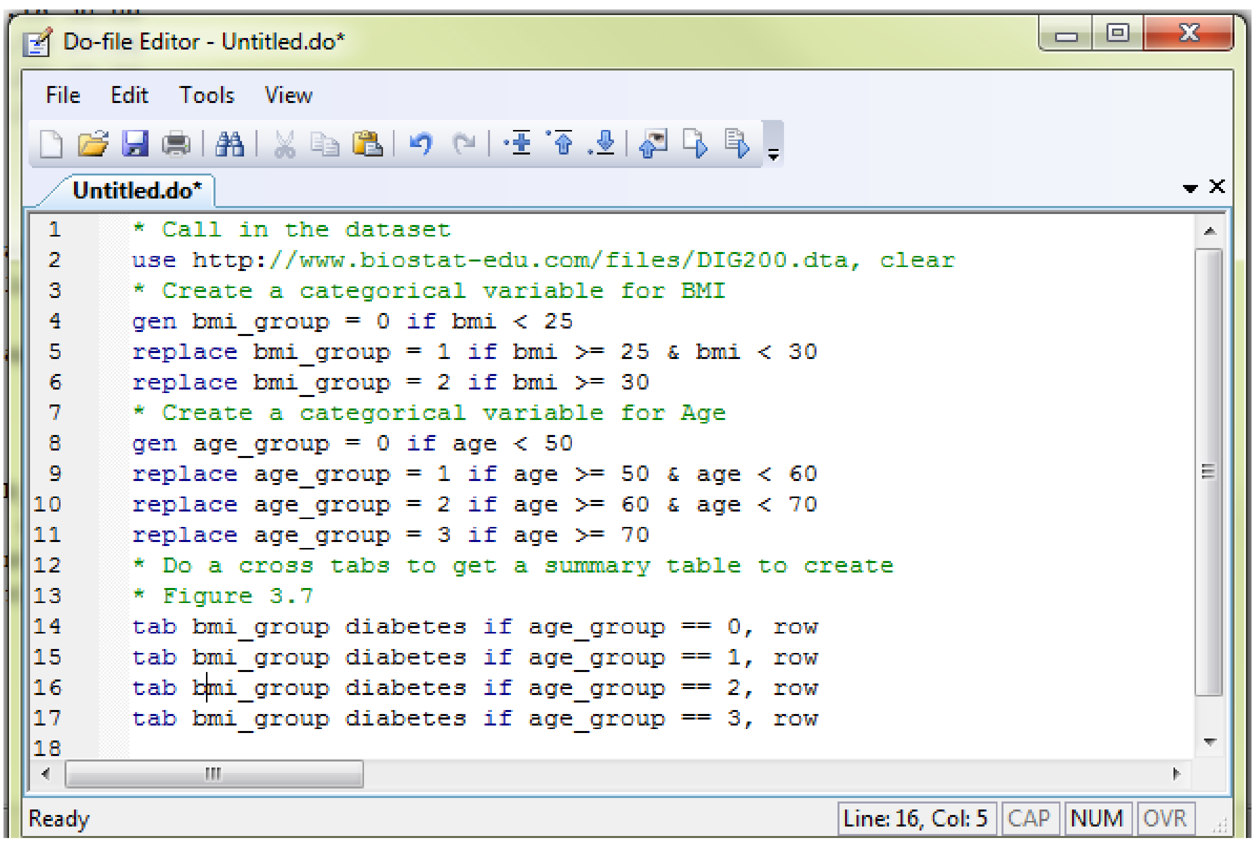
Stata Commands:

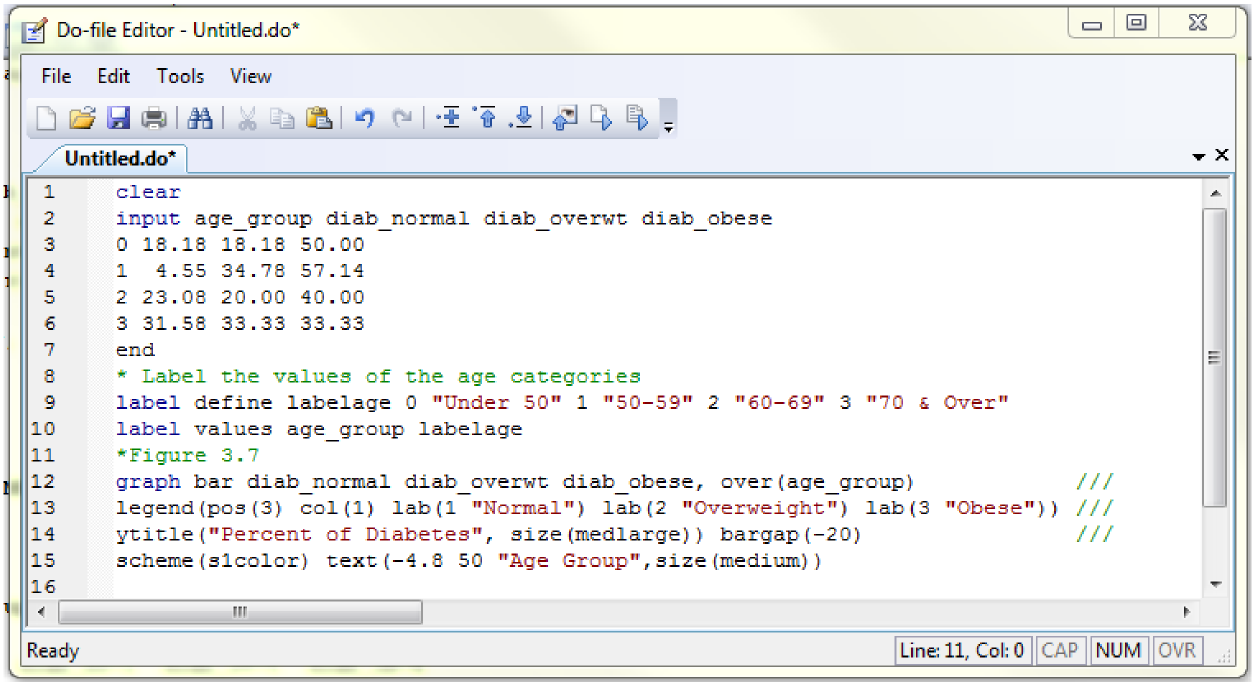
```
*Graph 2
graph bar normal overwt obese,percent over(age_group) stack                 ///
legend(pos(3) col(1) lab(1 "Normal") lab(2 "Overweight") lab(3 "Obese"))    ///
ytitle("Percent",size(medlarge)) scheme(s1color)
```

Stata Output:



The Stata commands below create the clustered bar chart in Figure 3.7. The first two screen shots indicate the Stata commands that were used to acquire the data then produce the clustered bar chart.





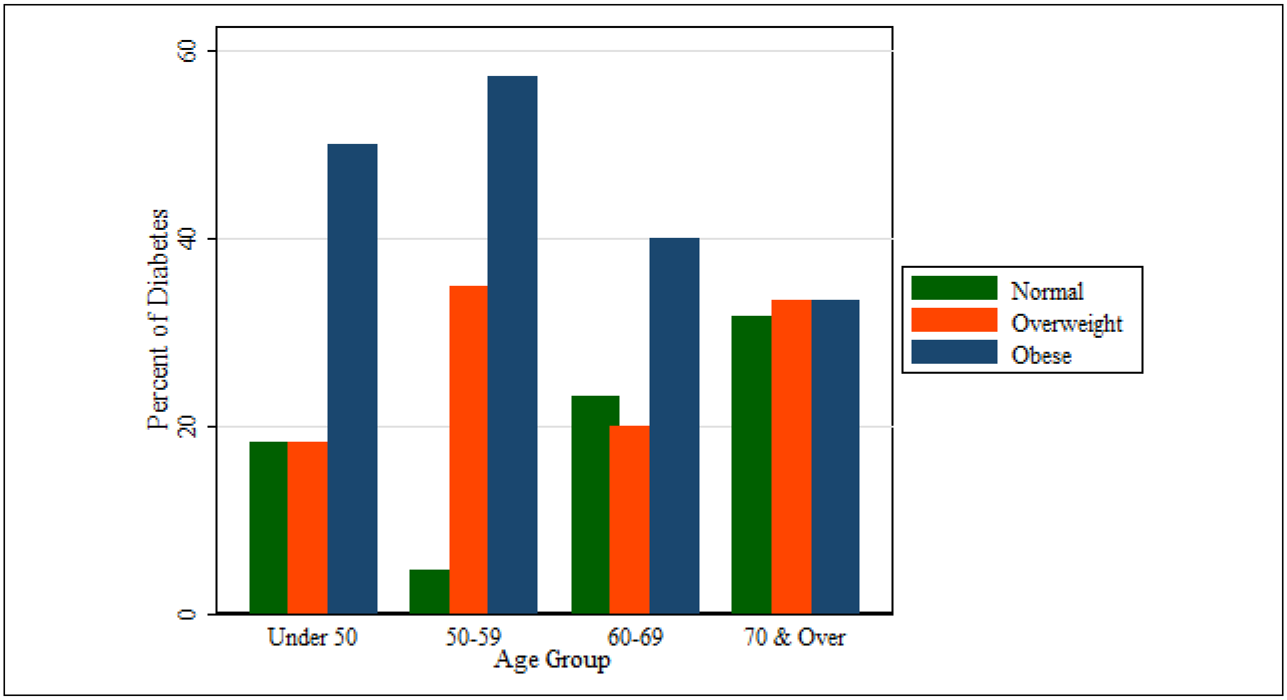
Carefully review the commands below and make modifications to tailor the figure to your liking.

Stata Commands:

```
graph bar diab_normal diab_overwt diab_obese, over(age_group)          ///
legend(pos(3) col(1) lab(1 "Normal") lab(2 "Overweight") lab(3 "Obese")) ///
ytitle("Percent of Diabetes", size(medlarge)) bargap(-20)             ///
scheme(s1color) text(-4.8 50 "Age Group",size(medium))
```

The figure looks as follows:

Stata Output:



Use the Stata command **help graph bar** to explore more details about graphing bar charts.

Program Note 3.3 – **Creating histograms**

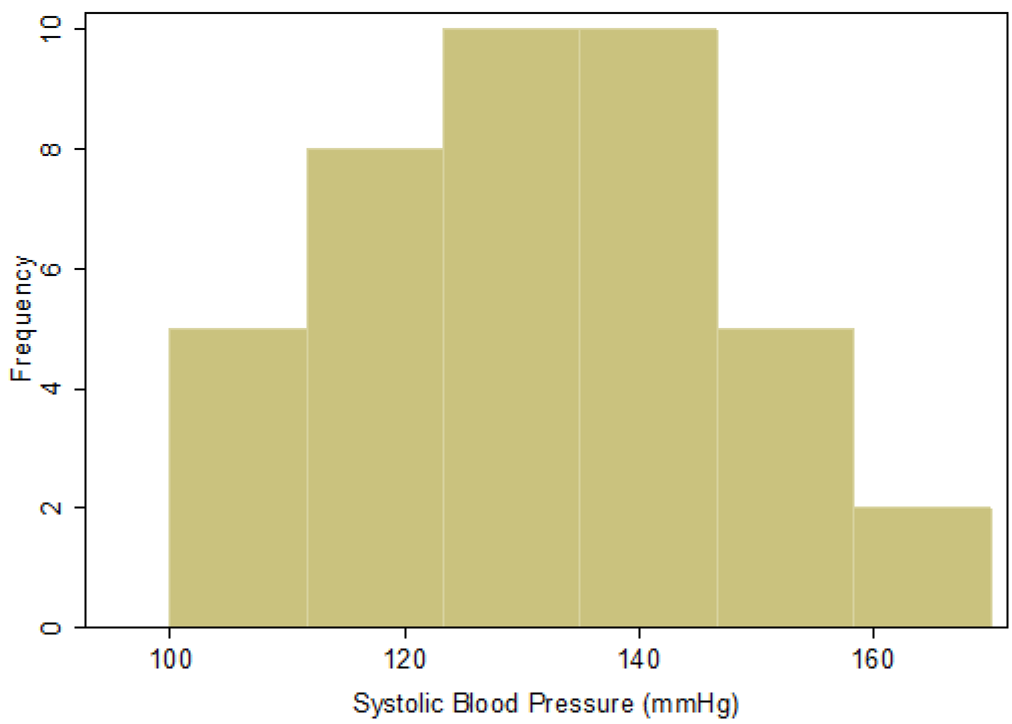
The **histogram (hist)** command can be used to create histograms in Stata. The **bin( )** option specifies the number of intervals or bins that are to be used. The **width( )** option specifies the bin width. By default, Stata uses **bin(k)** where **k = min{sqrt(N), 10*ln(N)/ln(10)}** and **N** is the number of observations. The **freq** option tells Stata that the vertical axis is to be labeled in frequency units rather than fractional units. By default, Stata uses the **density** option. We can also use **xlabel( )** and **ylabel( )** options to provide more information on the x and y-axes.

As an example, we use the **hist** command to display a histogram of the systolic blood pressure readings of participants in the **DIG40** data set.

Stata Commands:

```
use http://www.biostat-edu.com/files/dig40.dta, clear
histogram sysbp, freq                                         ///
xtitle("Systolic Blood Pressure (mmHg)") xscale(titlegap(3))  ///
scheme(s1color)
```

Stata Output:



To obtain more information including other options available for creating histograms use the Stata command **help hist**.

Program Note 3.4 – **Creating stem and leaf plots and scatter plots**

1. Stem and Leaf plots

The Stata command **stem** can be used to create stem and leaf plots. Using the **DIG40** data set, a stem and leaf plot for systolic blood pressure readings can be created with the commands below:

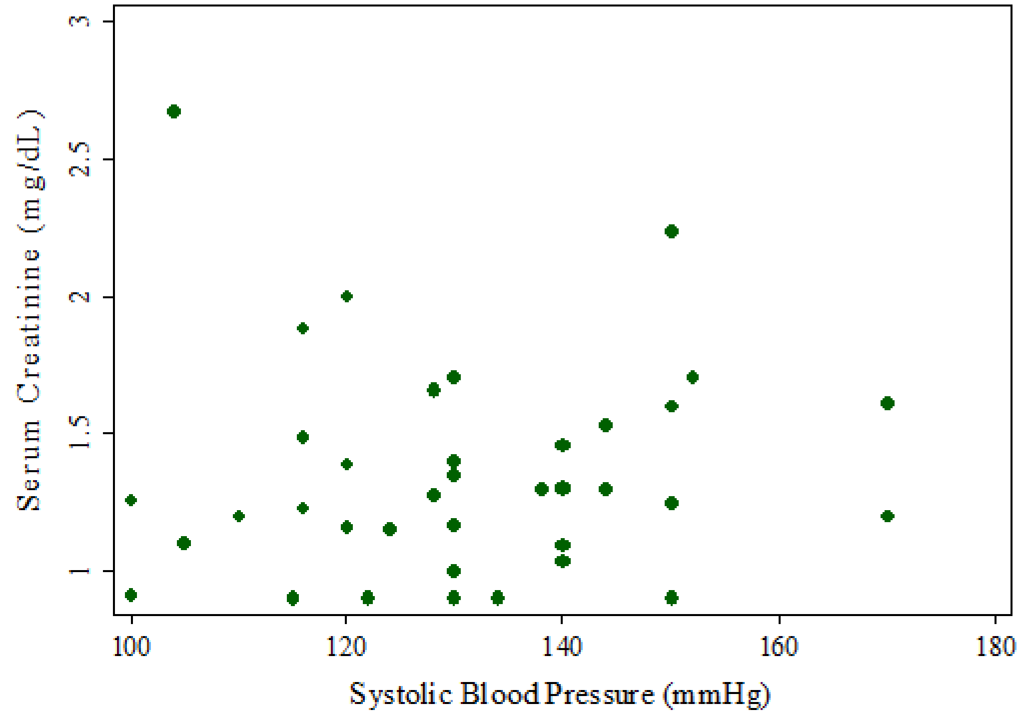
```
Stata Output:

use http://www.biostat-edu.com/files/dig40.dta, clear
stem sysbp

Stata Output:

Stem-and-leaf plot for sysbp (SYSBP)

  10* | 0045
  11* | 05666
  12* | 0002488
  13* | 00000048
  14* | 000000044
  15* | 00002
  16* |
  17* | 00
```

Use the Stata command **help stem** to explore more details about stem and leaf plots.

2. Scatter plots

The Stata command **twoway** introduced in the section on line graphs can also be used to create scatter plots. For line graphs, we used the command connected, but for scatter plots, we will use the command **scatter**. In Figure 3.12, we use a scatter plot to examine the relationship between serum creatinine and systolic blood pressure using the **DIG40** data set.

```
Stata Commands:

use http://www.biostat-edu.com/files/dig40.dta, clear
twoway (scatter creat sysbp), scheme(s1color)                  ///
ytitle("Serum Creatinine (mg/dL)",size(medlarge))              ///
xtitle("Systolic Blood Pressure (mmHg)",size(medlarge))        ///
xscale(titlegap(3)) yscale(titlegap(3))


Stata Output:
```



The plot of the quadratic relationship shown in Example 3.20 can be easily created using the Stata command **scatter** along with identifying the *y* and *x* variables needed to construct the plot. The commands below were used to input the data directly into Stata:

Stata Commands:

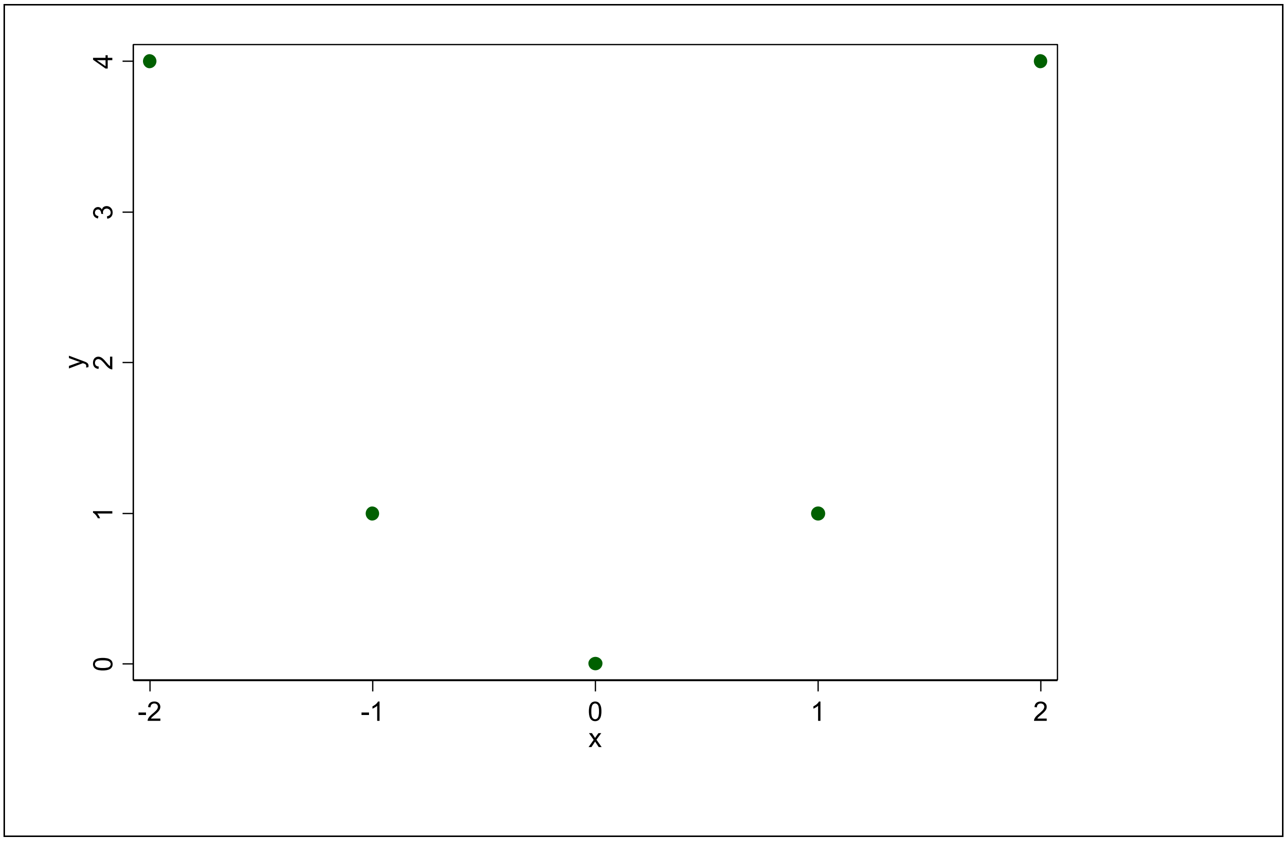
```
input x y
-2 4
-1 1
 0 0
 1 1
 2 4
end
```

Using the scatter command shown below, we created Figure 3.30.

Stata Commands:

```
scatter y x, scheme(s1color)
```

Stata Output:



Use the Stata command **help scatter** to explore more details about scatter plot options.

Program Note 3.5 – Descriptive statistics and creating box plots

1. Descriptive Statistics

The **tabstat** command in Stata can be used to obtain descriptive statistics on continuous variables. For example, we can use the Stata commands below to get the mean, standard deviation, and range for systolic blood pressure for patients from the **DIG40** data set.

```
Stata Commands:

use http://www.biostat-edu.com/files/dig40.dta, clear
tabstat sysbp, statistics(mean sd min max)


Stata Output:

    variable |      mean        sd       min       max
-------------+----------------------------------------
       sysbp |     131.4  16.87024       100       170
------------------------------------------------------
```

Other statistics that can be calculated include (Note-this information can be obtained by using the **help tabstat** command):

mean displays the mean
count displays the count of nonmissing observations
n is the same as count
sum displays the sum
max displays the maximum value
min displays the minimum value
range displays the range or (max – min)
sd displays the standard deviation
variance displays the variance
cv displays the coefficient of variation (sd/mean)
semean displays the standard error of mean (sd/sqrt(n))
skewness displays the skewness
kurtosis displays the kurtosis
p1 displays the 1st percentile
p5 displays the 5th percentile
p10 displays the 10th percentile
p25 displays the 25th percentile
median displays the median (same as p50)
p50 displays the 50th percentile (same as median)
p75 displays the 75th percentile
p90 displays the 90th percentile
p95 displays the 95th percentile
p99 displays the 99th percentile
iqr displays the interquartile range = p75 - p25
q is equivalent to specifying p25 p50 p75

Another Stata command that gives similar information is the **summarize** command. Using the Stata commands below, we were able obtain the number of observations, the mean , the standard deviation, the minimum and the maximum for the systolic blood pressure readings of the **DIG40** participants:

```
Stata Commands:

use http://www.biostat-edu.com/files/dig40.dta, clear
summarize sysbp


Stata Output:

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       sysbp |        40       131.4    16.87024        100        170
```

We used the **detail** option to have information on key percentiles. For example, using the same commands as the ones above but adding the **detail** option results in more statistics being presented.

```
Stata Commands:

use http://www.biostat-edu.com/files/dig40.dta, clear
summarize sysbp, detail


Stata Output:

                            SYSBP
-------------------------------------------------------------
      Percentiles      Smallest
 1%          100            100
 5%          102            100
10%        107.5            104       Obs                  40
25%          120            105       Sum of Wgt.          40

50%          130                      Mean              131.4
                        Largest       Std. Dev.      16.87024
75%          140            150
90%          150            152       Variance       284.6051
95%          161            170       Skewness       .1429545
99%          170            170       Kurtosis       2.823232
```

2. Box plots

The Stata command **graph box** or **graph hbox** can be used to create box plots. Box plots can be displayed vertically with the **graph box** command or horizontally with the **graph hbox** command. As an example, we show the commands that were used to create Figure 3.15 using the **DIG40** data set starting with the data management commands first.

Stata Commands:

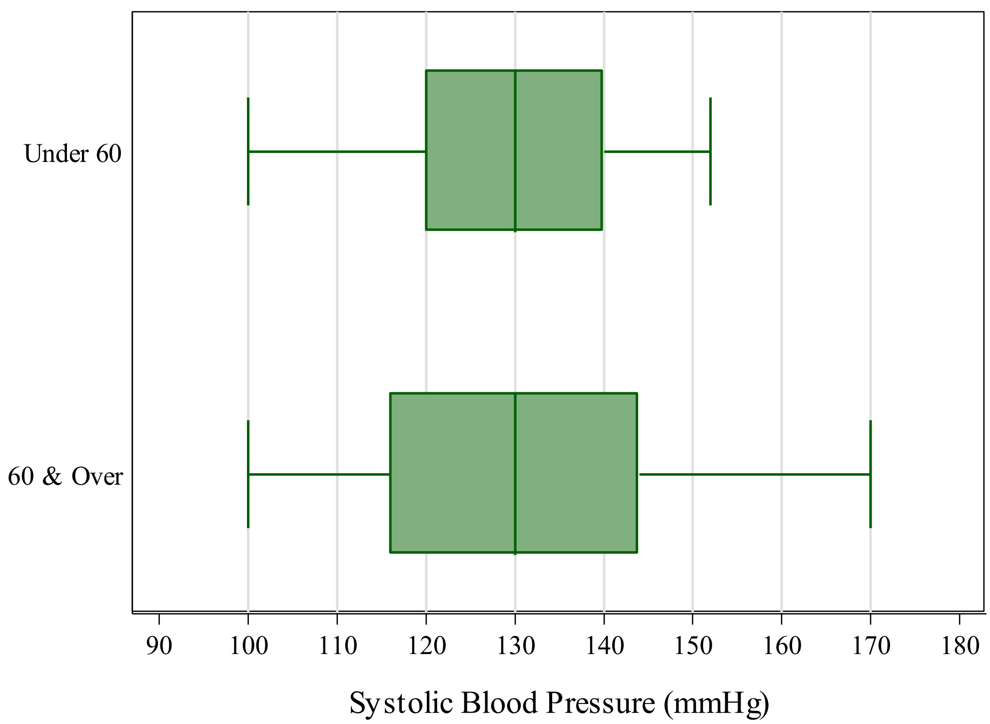
```
use http://www.biostat-edu.com/files/dig40.dta, clear
gen age_cat = 0 if age<60
replace age_cat = 1 if age>= 60

label define labage 0 "Under 60" 1 "60 & Over"
label values age_cat labage
```

In the command lines, we used the **over** option to draw box plots for systolic blood pressure at each level of the variable in parentheses- specifying the two age categories (Under 60 and 60 & Over).

Stata Commands:

```
graph hbox sysbp, over(age_cat) scheme(s1color)            ///
ytitle("Systolic Blood Pressure (mmHg)", size(medlarge))   ///
ylabel(90 (10) 180) yscale(titlegap(4))
```

Stata Output:



Use the Stata command **help box** to explore more details about box plots.

Program Note 3.6 – Calculating Pearson and Spearman correlation coefficients

Stata uses the **correlate (corr)** command to calculate the Pearson correlation coefficient, or the **spearman** command to calculate the Spearman Rank Correlation Coefficient. We can use the data in Example 3.18 to show the usefulness of the **corr** and **spearman** commands. The data can be entered as shown below:

Stata Commands:

```
input sysbp diabp
120 60
118 60
130 68
140 90
140 80
128 75
140 94
140 80
120 60
128 80
124 70
135 85
end
```

The commands below give the Pearson correlation coefficient:

Stata Commands:

```
corr sysbp diabp
```

Stata Output:

```
(obs=12)

             |    sysbp    diabp
-------------+------------------
       sysbp |   1.0000
       diabp |   0.8936   1.0000
```

The commands below give the Spearman rank correlation coefficient:

Stata Commands:

```
spearman sysbp diabp
```

Stata Output:

```
Number of obs =      12
Spearman's rho =     0.8660

Test of Ho: sysbp and diabp are independent
Prob > |t| =       0.0003
```

Use the Stata command **help correlate** or **help spearman** to get more details about calculating correlation coefficients.