4

DESCRIPTIVE STATISTICS

At this point, you should feel fairly comfortable and confident with the basic operations of Stata. The next four chapters explain how to conduct basic quantitative analyses using Stata. The strategies and techniques covered are most commonly found in introductory social statistics textbooks. Stata can perform countless additional and more advanced statistical techniques than can be covered. The commands that are discussed, however, provide a foundation for understanding virtually any specific analyses one could perform in Stata. Then the Stata Help Files section of Chapter 8 addresses ways in which interested readers can help themselves learn how to conduct techniques that are not specifically covered.

To be clear, this book does not seek to explain how or why particular statistical analyses should be used. It should not be viewed as a replacement to a thorough quantitative analysis text. Rather, it has been designed to act as a companion to such texts. Once you have a solid understanding of the mechanics underlying the basic analytic strategies, these chapters guide you through how to conduct them on real data. Most important, the interpretation of the results is extremely brief and is aimed at identifying the most important components of the output rather than explaining their meaning. Again, readers should consult dedicated quantitative analysis texts for a more comprehensive explanation of how to interpret the statistical figures presented.

Perhaps the most important first step in any quantitative research project is to understand the distribution of each of the pertinent variables. This chapter covers the commands that are used to produce univariate (i.e., single-variable) descriptive statistics, including measures of central tendency and variability. Methods for presenting these measures graphically are also addressed. All the examples that follow use the `Chapter 4 Data.dta` file, available at study.sagepub.com/longest3e. This data set includes the full NSYR Wave 3 sample of 2,532 young adults. All the missing cases have been replaced with appropriate missing codes in this data set, with `.d` referring to a response of "Don't Know," `.r` to a response of "Refused," and `.s` to cases that were legitimately skipped out of a question based on the survey design (i.e., skip pattern).

FREQUENCY DISTRIBUTIONS

For variables with a limited number of categories, the easiest way to gain an initial sense of how the variable is distributed is by producing a frequency distribution. A frequency distribution shows how many cases and the percentage of cases that belong to each category of a variable.

For example, you might conduct a project on young adults' perception and satisfaction with their physical appearance. The NSYR has a question that asked the respondents, "In general, how happy or unhappy are you with your body and physical appearance?" The responses to this question provide a good overview of young adults' body image. As was introduced in Chapter 2, the `-tab-` command is used to produce a frequency distribution of a variable. The variable for the question about body image is called `body`. Typing `tab body` into the Command window and pressing **Enter** produces the following results:

```
tab body

     (body _ w3) P:3. |
     In general, how  |
   happy or unhappy   |
  are you with your   |
     body and physic  |      Freq.     Percent        Cum.
----------------------+-----------------------------------
       Very unhappy   |         68        2.70        2.70
   Somewhat unhappy   |        389       15.42       18.11
            Neither   |        234        9.27       27.39
     Somewhat happy   |        953       37.77       65.16
         Very happy   |        879       34.84      100.00
----------------------+-----------------------------------
              Total   |      2,523      100.00
```

By now, you have seen several such distributions, but it may be helpful to review the primary components in terms of how they would relate to an actual analysis. First, the variable label is listed in the upper left-hand corner of the table. For the `body` variable (and with all variables in the NSYR data set), the default variable label is a brief description of the survey question wording. Next, the left-hand column lists all the categories for the variable to which at least one respondent belongs. There might be other possible answer choices, but they are not shown in the frequency table if no one in the data belongs to those categories.

Finally, the three columns to the right display the number of respondents in each category (`Freq.`), the percentage of respondents in each category (`Percent`), and the cumulative percentage of respondents in each category (`Cum.`). The frequency is a count of all the participants who answered in each category. For example, 68 of the respondents reported being "very unhappy" with their physical appearance. The percentage reports this frequency divided by the "Total" number of valid (i.e., nonmissing) cases on this variable. Only 2.7% of the 2,523 cases report being very unhappy with their body. This figure suggests that only a small set of young adults are very dissatisfied with their physical appearance. The cumulative percentage displays the total percentage of cases that belong to that category and the ones preceding it. For the `body` variable, 18.11% of the cases report being somewhat unhappy or very unhappy with their physical appearance. Again, you can see that the vast majority of young adults are either indifferent or at least somewhat happy with their physical appearance.

As was discussed in detail in the Data Management: Missing Data section of Chapter 3, the `-tab-` command by default does not list cases that did not provide an answer to that variable's question (i.e., are missing). Furthermore, the percentages that are displayed in the table are based on the total number of cases who provide a valid answer to the particular variable, not the total number of cases in the data set. As was discussed in more detail in the Data Management: Missing Data section of Chapter 3, often certain measures will have missing cases due to either respondents choosing not to answer a question or the survey design skipping particular cases out of specific questions. Still, there are instances where you may be interested in the percentages of respondents in each category based on the total sample, regardless of whether they provided a valid response to the question. In the NSYR the total sample is 2,532. To see the percentages of the body image questions based on this total sample, you can add the `-mis-` option to the previous -tab- command. Type tab body, mis in the Command window (or press **Page Up** while the Command window is highlighted and add ", mis" to the previously executed command) and press **Enter**. The full sample results appear as follows:

```
tab body, mis

     (body _ w3) P:3. |
    In general, how   |
  happy or unhappy    |
      are you with    |
     your body and    |
            physic    |      Freq.     Percent        Cum.
----------------------+-----------------------------------
     Very unhappy     |         68        2.69        2.69
 Somewhat unhappy     |        389       15.36       18.05
          Neither     |        234        9.24       27.29
    Somewhat happy    |        953       37.64       64.93
       Very happy     |        879       34.72       99.64
               .s     |          9        0.36      100.00
----------------------+-----------------------------------
            Total     |      2,532      100.00
```

The new frequency distribution shows that 9 cases were skipped out of this question about body image. But you can see that the percentages in each category do not change substantially (generally less than a few hundredths of a percent) when they are calculated using these 9 cases.

To see the frequency order of the categories more clearly, the -sort- option can be invoked. Type tab body, sort into the Command window, and press **Enter**. The table is now displayed as follows:

```
tab body, sort

     (body _ w3) P:3. |
    In general, how   |
  happy or unhappy    |
      are you with    |
     your body and    |
            physic    |      Freq.     Percent        Cum.
----------------------+-----------------------------------
    Somewhat happy    |        953       37.77       37.77
       Very happy     |        879       34.84       72.61
 Somewhat unhappy     |        389       15.42       88.03
          Neither     |        234        9.27       97.30
     Very unhappy     |         68        2.70      100.00
----------------------+-----------------------------------
            Total     |      2,523      100.00
```

The logical ordering of the categories is now replaced by the frequency order, with the most prevalent category listed first and the least frequent category last. This table shows even more clearly that the most popular perception among young adults is being generally happy with their body. The cumulative percentage column indicates that more than 72% of young adults report being either somewhat or very happy.

For this particular project, you might predict that there would be a relationship between being sad and perception of one's body. The NSYR asked, "How frequently do you feel sad?" Before you examine any relationship between the two variables, it is helpful to know the distribution of both variables involved. You can use the `-tab-` command as above to produce this distribution for the new `sad` variable. If you knew that you were interested in the distribution of both variables, however, you might try listing both variables after the `-tab-` command. If you use the `-tab-` command with two variables, however, a cross-tabulation will be displayed rather than separate frequency distributions. Stata offers a slightly different command that produces multiple frequency distributions using one command line. The `-tab1-` command allows for several variables to be entered at one time, and separate frequency distributions for each variable are displayed, in the order in which the variables are entered in the command line.

For example, type `tab1 body sad` into the Command window and press **Enter**. The frequency distributions of the `body` and then the `sad` variables are displayed. `-tab1-` accepts all the options that have been discussed with the `-tab-` command, but when you invoke any option with `-tab1-`, those options are applied to every variable that is listed in the command line.

A common aspect of research projects is to examine patterns of behavior within particular subgroups. For example, you might only be interested in the body image of young adult females. Often, people new to quantitative analyses are tempted to eliminate the subgroup that they are not interested in (e.g., males) from the data set entirely. This strategy is not advisable because it might be important to at least initially examine patterns for the whole sample or compare females with males, even if the ultimate research goals only concern females. Therefore, rather than dropping all the males from the data, it is more effective to use `-if-` statements along with the analytic commands. If you wanted to see the distribution of the `body` variable for females only, you would need to add an `-if-` statement, using the `gender` variable, to the end of the command line. The command would read `tab body if gender==#`. Remember, a double equal sign is needed whenever you are asking Stata to evaluate whether a statement is true.

Before you complete the command, you need to know the numeric value that is used to represent females in the `gender` variable. There are several ways to obtain this information, but perhaps the most straightforward is to perform two separate -`tab`- commands of the `gender` variable, one that invokes the -`nol`- option and one that does not. First type `tab gender` in the Command line and press **Enter**, and then type `tab gender, nol` in the Command line and press **Enter**. (If the variable of interest does not have value labels attached, the second -`tab`- command would not be necessary.) Doing so produces the following results:

```
tab gender

       tab gender
                   |
   (gender _ w3)   |
     Respondent    |
         Gender    |      Freq.       Percent        Cum.
-------------------+-----------------------------------------
           Male    |      1,232         48.66       48.66
         Female    |      1,300         51.34      100.00
-------------------+-----------------------------------------
          Total    |      2,532        100.00

 tab gender, nol

   (gender _ w3)   |
     Respondent    |
         Gender    |      Freq.       Percent        Cum.
-------------------+-----------------------------------------
              0    |      1,232         48.66       48.66
              1    |      1,300         51.34      100.00
-------------------+-----------------------------------------
          Total    |      2,532        100.00
```

Now that you know females are coded as 1 on the `gender` variable, you can produce the frequency distribution of the `body` variable for females only. In this case, you might also consider using the -`sort`- option to see if the frequency order for females is similar to that of the entire sample. The command to produce this frequency distribution is `tab body if gender==1, sort`. Notice that the -`if`- statement comes before the option. This ordering is true for all commands. -`if`- statements always precede the comma that separates the command from the options. Once you type that command line into the Command window and press **Enter**, the following results are displayed:

```
tab body if gender==1, sort

     (body _ w3) P:3 |
   In general, how |
  happy or unhappy |
      are you with |
     your body and |
            physic |      Freq.     Percent        Cum.
-------------------+-----------------------------------
    Somewhat happy |        511       39.43       39.43
       Very happy |        375       28.94       68.36
 Somewhat unhappy |        246       18.98       87.35
           Neither |        117        9.03       96.37
      Very unhappy |         47        3.63      100.00
-------------------+-----------------------------------
             Total |      1,296      100.00
```

This distribution is calculated using only female respondents. The total number of females included in this distribution is 1,296, meaning that 4 females are missing on this question (because the distribution table of `gender` showed 1,300 females present in the entire sample). Both the individual category percentages and cumulative percentages use this total in the denominator of the calculation. The results show that the majority of young adult females, 68.36%, are somewhat or very happy with their bodies.

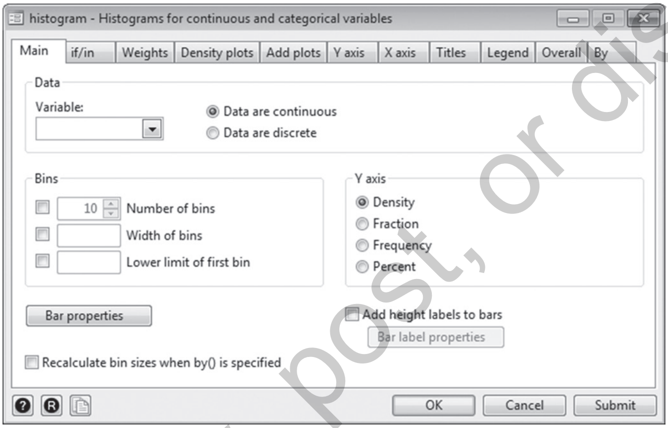
Histograms and Bar Graphs

In addition to examining the actual frequency distribution, it can be helpful to see the distribution of the variables visually. To do so, there are two primary options: histograms and bar graphs. In practice, histograms and bar graphs are essentially the same, as both use bars to illustrate the relative frequency of categories within a variable. In Stata, it is easier to produce these types of graphs using the `-histogram-` command and point-and-click box than the bar graph command. The latter will be used for examining the relationship between two variables.

Thus far, this book has focused almost entirely on running procedures by using the Command window interface. Perhaps the one area in which the point-and-click method has advantages that outweigh the Command window method is in preparing graphs. Due to the vast number of manipulations that can be made to the display of graphs, the commands to produce them can become lengthy and cumbersome. Therefore, for the sections explaining the productions of graphs, the point-and-click interface is primarily used.

To use a histogram to examine the distribution of young adults' `body` image, start by clicking on the **Graphics** menu button at the top of the Stata window and then select **Histogram**. A window like the one shown in Figure 4.1 will appear.

FIGURE 4.1 ⬢ HISTOGRAM MAIN OPTIONS WINDOW



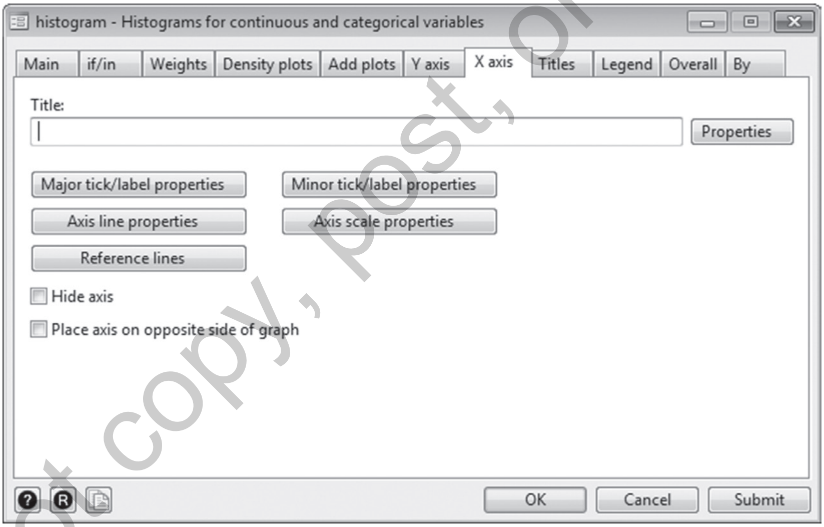
Once this window appears, you need to specify the variable that you want to produce a histogram of. Place your cursor in the **Variable** box in the upper left-hand corner. You can use the drop-down menu to find the `body` variable, or you can simply type `body` into the box.[1] Next, you can select what scale the *Y*-axis should depict. The default is set to `Density`, but it is more common to display either the **Frequency** or the **Percentage**. The shape of the histogram will not change regardless of which scale is chosen, but the interpretability of the *Y*-axis is typically enhanced by using the **Percentage** option. Simply click the radio button next to **Percent** to set the *Y*-axis scale. This window also

[1] Technically, the `body` variable contains discrete values. To create the appropriate histogram for such a variable, as opposed to a true continuous variable, click the "Data are discrete" radio button in the histogram window. For the purposes of this example, however, it is useful to treat the variable as continuous. Neither option changes the general picture of the distribution, but if this graph was being used in a project, selecting the correct discrete or continuous option is advised.

contains an option to add a “height label” to the bars. This option will display the actual number, depending on the *Y*-axis scale, that is associated with each category. For example, 2.70% would be listed above the “very unhappy” bar because 2.7% of the sample report being very unhappy with their bodies. Finally, you can select the **Bar Properties** button to alter the appearance (e.g., color and outline pattern) of the bars themselves.

At this point, if you pressed the **OK** button, a histogram would be produced. But, by default, histograms label the bars with the numeric values of the categories rather than the value labels. To alter the labeling, in order to ease the readability of the graph, click on the ***X*-axis** tab while still in the histogram window. The window as shown in Figure 4.2 will appear.

FIGURE 4.2 ⬢ HISTOGRAM *X*-AXIS OPTIONS WINDOW



histogram - Histograms for continuous and categorical variables

Title:

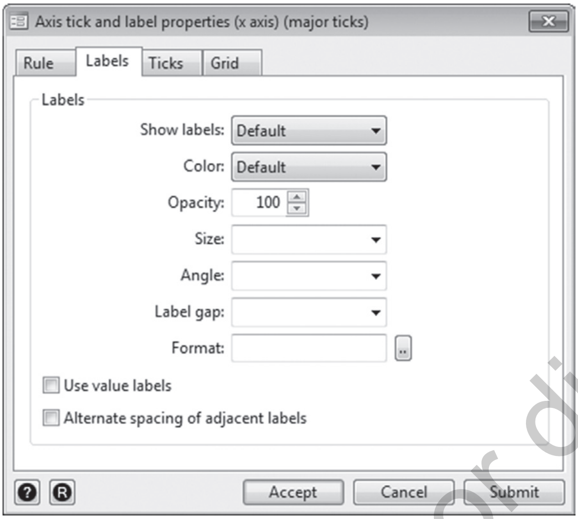
□ Hide axis

□ Place axis on opposite side of graph

This window contains all the options for the *X*-axis. For example, you could enter a title, such as “Happiness With Body and Physical Appearance,” in the **Title** box. To have the bars labeled with the appropriate value labels, press the **Major tick/label properties** button. Once the new window appears, select the **Labels** tab, which will display the window as shown in Figure 4.3.

Within this window check the radio button next to **Use value labels**, then click **Accept**, and finally click **OK**. The resulting histogram should look like the one shown in Figure 4.4.

FIGURE 4.3 ⬢ HISTOGRAM LABELS OPTIONS WINDOW



Axis tick and label properties (x axis) (major ticks)

Labels

Show labels: Default

Color: Default

Opacity: 100

Size:

Angle:

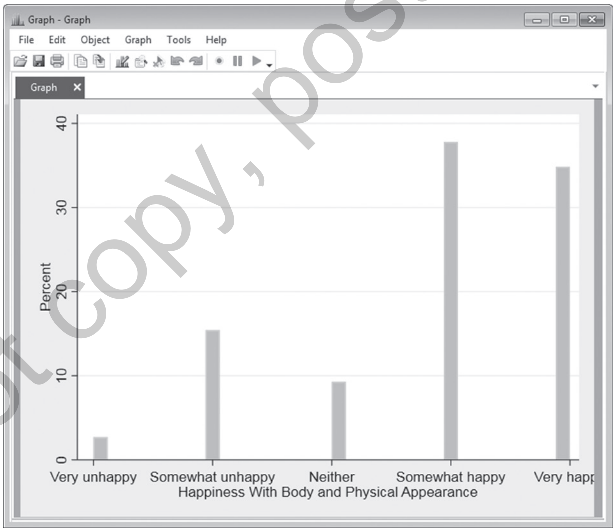
Label gap:

Format:

□ Use value labels

□ Alternate spacing of adjacent labels

FIGURE 4.4 ⬢ HISTOGRAM OF `body` VARIABLE



```
histogram body, percent
xtitle(Happiness With Body and Physical
Appearance) xlabel(, valuelabel)
```
[2]

[2] Whenever a graph is shown, the full command that would be used to produce that graph will be listed directly beneath it. Users who are interested in using these commands should see the Closer Look box "Using Commands to Create Graphs" for more information. Users who prefer or plan to only use the point-and-click method to create graphs can disregard these commands.

A CLOSER LOOK: USING COMMANDS TO CREATE GRAPHS

Although getting started with graphs through the point-and-click interface is probably easier, using the command-based interface is, as with most other commands, more effective and efficient for those planning to use Stata for full research projects. As with all operations performed using the point-and-click interface, the actual command to produce a given graph is displayed in both the Results and Review windows. This feature allows users to follow the point-and-click steps described and then use the displayed command to learn the necessary commands and options to produce the graphs via the Command window or a do file. From here on, the command that produces any displayed graph will be listed directly underneath it.

The full commands to produce these graphs can get somewhat complex given the number of options that are needed to make all of the necessary manipulations to the appearance of the graph. This situation is another example of the utility of do files, described in the What Is a Do File? section of Chapter 3. As soon as you have created a graph once, you can copy and paste the full command into a do file. From then on, you can simply replace the variables that are being graphed and/or directly alter specific options to quickly create a new or modified graph. Although all of the previously invoked options are saved within the graph window while Stata is kept open, if you want to replicate or alter a graph after starting a new Stata session, you would have to go back through and reselect and enter every option again.

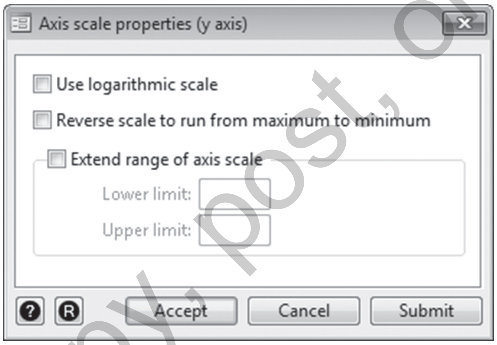
For example, as shown in Figure 4.4, the command to produce this first histogram is `histogram body, percent xtitle(Happiness With Body and Physical Appearance) xlabel(, valuelabel)`. This is the base command to create any histogram. If you copy this command, from either the Command or Review window, and paste it into a do file, you can quickly create subsequent histograms. You might replace `body` (and the accompanying title) with the `sad` variable or the `faith1` variable to examine these distributions. You can also modify some of the options. For example, you probably have already guessed that the "percent" aspect of the command controls what is graphed on the *Y*-axis and maybe can intuitively deduce that changing this to –`frequency`– would create a frequency histogram. Using a do file, therefore, allows for quick replications and manipulations, across different Stata sessions, to even very complicated graph commands by bypassing the numerous windows and radio buttons that are necessary using the point-and-click method.

This graph can be saved by clicking on the **File** menu and then **Save As**. You can change the file format into several different picture file types, although by default, it saves as a Stata graph (.gph). This graph again clearly shows that the two happy categories are far more prevalent among young adults than are the unhappy categories. Notice, however, that the *Y*-axis is scaled to the maximum percentage of the given categories, not 100%.

To alter the display of the *Y*-axis, click on the **Graphics** menu and then Histogram. All the options you have previously set are still in the **Histogram** window. If you wanted to clear all those options, you can press the Ⓡ icon located in the lower left corner. Select the ***Y*-axis** tab followed by the **Axis scale properties** button.

In the window that appears (Figure 4.5), select the **Extend range of axis scale** button, set the **Lower limit** to 0 and the **Upper limit** to 100, and click **Accept.** Now select the **Major tick/label properties** button followed by the **Suggest # of ticks** button to reveal the window shown in Figure 4.6.

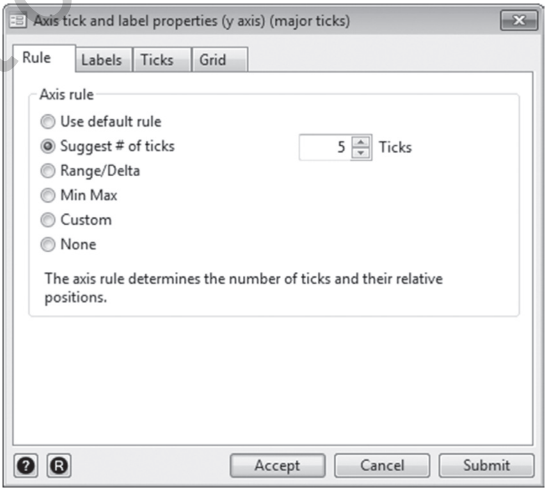
FIGURE 4.5 ⬢ HISTOGRAM *Y*-AXIS SCALE OPTIONS WINDOW



Axis scale properties (y axis)

□ Use logarithmic scale
□ Reverse scale to run from maximum to minimum
□ Extend range of axis scale
Lower limit:
Upper limit:

FIGURE 4.6 ⬢ HISTOGRAM *Y*-AXIS TICK OPTIONS WINDOW



Axis tick and label properties (y axis) (major ticks)

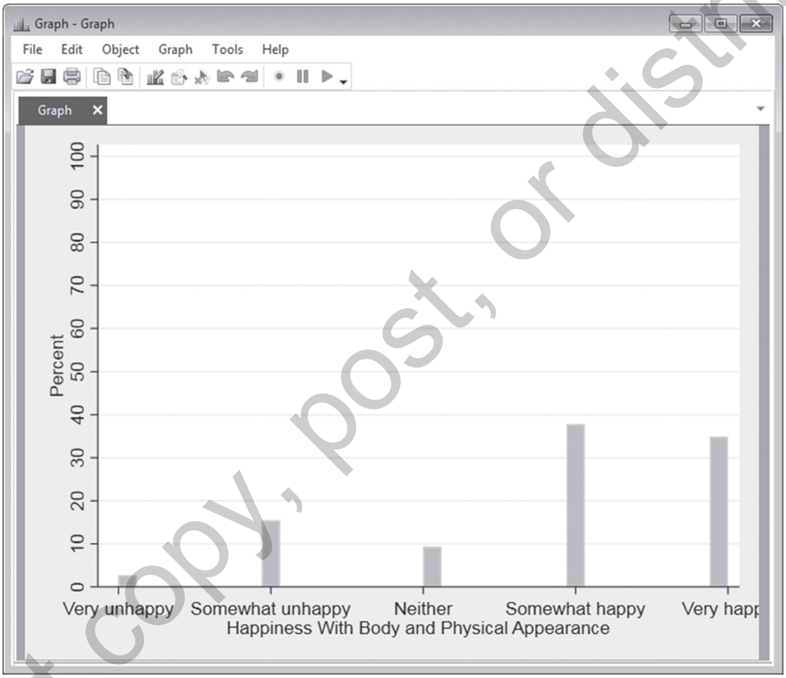
Axis rule
○ Use default rule
● Suggest # of ticks 5 Ticks
○ Range/Delta
○ Min Max
○ Custom
○ None

The axis rule determines the number of ticks and their relative positions.

Finally, change the **Ticks** value to 10, instead of 5, click **Accept**, and then **OK**.

The histogram should look like the one in Figure 4.7. The *Y*-axis is now on a 0% to 100% scale, providing a more accurate depiction of the absolute percentages of each category.

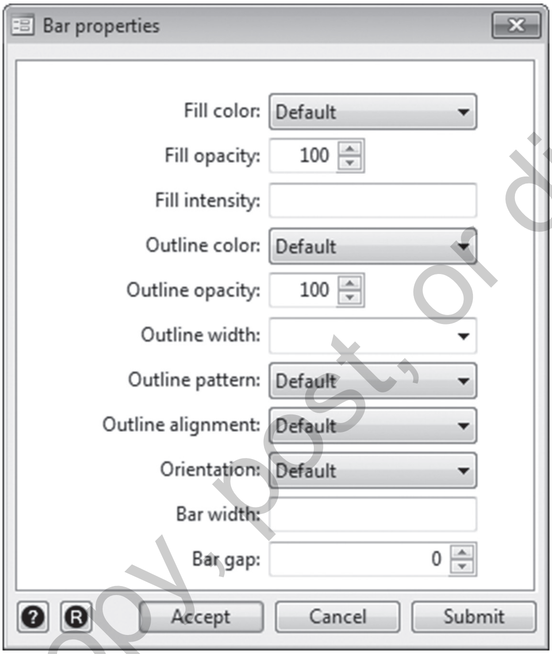
FIGURE 4.7 ⬢ HISTOGRAM OF BODY VARIABLE WITH ADJUSTED *Y*-AXIS



```
histogram body, percent yscale(range(0 100))
ylabel(#10) xtitle(Happiness With Body and
Physical Appearance) xlabel(, valuelabel)
```

If you are using the histogram simply to get a sense of the overall distribution of the variable, then the last version is probably good enough. If, however, you are producing this graph for a presentation or paper, it is probably helpful to explore the other display options to improve the appearance of the graph. There are too many such options to cover them all here, but they all work in a similar format to the options you just invoked. As one further example, you may want to alter the look of the bars. To do so, click on the **Graphics** menu and then **Histogram** again. In the lower left of the histogram window, there is a button for **Bar properties**. Clicking on that button reveals the window as shown in Figure 4.8.

FIGURE 4.8 ⬢ HISTOGRAM BAR PROPERTIES OPTIONS WINDOW



Bar properties

Fill color: Default

Fill opacity: 100

Fill intensity:

Outline color: Default

Outline opacity: 100

Outline width:

Outline pattern: Default

Outline alignment: Default

Orientation: Default

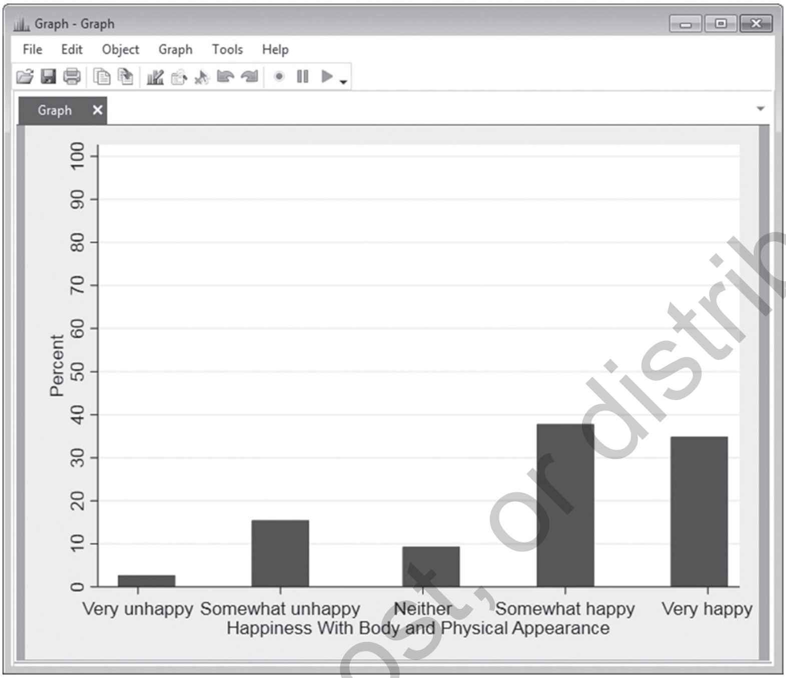
Bar width:

Bar gap: 0

This window allows you to change several properties of the bars. Click on the **Fill color** drop-down box, select **Blue**, and then do the same for the **Outline color**. Next, select the **Bar width** box and enter .4, which will widen the bars slightly. Finally, click **Accept** and **OK**. Doing so creates the graph shown in Figure 4.9.

Now the graph is a bit more visually appealing. The wider bars make the relative distribution easier to see, as well as helping to fit all the labels on the screen. Again, there are many more options that further manipulate the appearance of the graph for particular purposes. All of them can be found using the various tabs in the histogram window.

FIGURE 4.9 ⬢ FORMATTED HISTOGRAM OF BODY



```
histogram body, percent fcolor(blue) lcolor(blue)
barwidth(.4) yscale(range(0 100)) ylabel(#10)
xtitle(Happiness With Body and Physical
Appearance) xlabel(, valuelabel)
```

MEASURES OF CENTRAL TENDENCY AND VARIABILITY

A frequency distribution is a good first step in describing a variable. Such distributions are especially helpful for nominal or ordinal measures that do not have numerous categories. When working with interval-ratio measures that contain extensive categories, measures of central tendency and variability are more appropriate tools to succinctly provide important information about the variable in question.

The most prominent measures of central tendency are the mean, median, and mode. The mean refers to the arithmetic average of the variable's value, the median is the "middle" value (when the values are arranged in numerical or value order), and the mode is the most prevalent category. The range, interquartile range (IQR), variance, and standard deviation are the most commonly reported measures of variability. The range represents the difference between the maximum and the minimum value, and the IQR is the difference between the 25th and 75th percentile values. The variance is the "average" difference between each case's value and the mean value, while the standard deviation is the square root of the variance.

You actually have already learned how to obtain one of these measures. The -tab- command along with the -sort- option is the quickest way to obtain the mode. The category that is listed first, when the -sort- option is used (or the category with the highest percentage if it is not), is the modal category or value.

As introduced in Chapter 2, the -sum- command (short for -summary-) displays several of the other necessary measures. In addition to being interested in young adults' perceptions of their bodies, a researcher might be concerned with their actual body dimensions. One variable in the NSYR data that assesses body type is bmi. This variable represents the body mass index, or BMI, of the respondents. According to the National Institute of Health, BMI values greater than 30 indicate that a person is obese, values between 25 and 29.9 indicate that a person is overweight, and values 18.5 through 24.9 represent a person in the normal weight range.[3] To calculate BMI, each person's exact height and weight are used, resulting in an interval-ratio measure.

Type sum bmi into the Command window, and press **Enter**. The following results are displayed:

```
sum bmi

    Variable |     Obs        Mean    Std. Dev.       Min        Max
-------------+------------------------------------------------------
         bmi |   2,509    24.73744    5.141812   14.01495   63.49296
```

Five figures are presented with the default -sum- command. First, the number of valid observations (i.e., nonmissing) is shown (Obs.). For the bmi variable, 23 cases are missing (2,532 – 2,509 = 23). Most likely, several people did not wish to report their weight, resulting in a missing bmi

[3] www.nhlbi.nih.gov/guidelines/obesity/BMI/bmicalc.htm

value. Next, the mean is reported (Mean). A value of 24.74 suggests that, on average, young adults are at the very top of the normal weight range. Next, a standard deviation (Std. Dev.) of 5.14 suggests a reasonably wide distribution of values. The minimum (Min) and maximum (Max) values are displayed, making it easy to calculate the range by subtracting the former from the latter. For the bmi variable, there is a range of almost 50 units (63.49 – 14.01 = 49.48). As a reminder, as discussed in Chapter 2, you can use the Command window as a calculator to determine the exact range by using the -display- command (abbreviated to -di-). To do so, type di 63.49 – 14.01 into the Command window and press **Enter**. The solution, 49.48, is shown in the Results window.

The default -sum- command does not list the median, IQR, or variance. To produce these statistics, the -detail- option must be invoked.

Type sum bmi, detail in the Command window, and press **Enter** to produce the following results:

```
sum bmi, detail

           (bmi _ w3) Body Mass Index (NIH calculation)
-------------------------------------------------------------
      Percentiles      Smallest
 1%      16.9512       14.01495
 5%     18.75257       14.22837
10%     19.57563       14.64583       Obs               2509
25%     21.28223       14.76581       Sum of Wgt.       2509

50%     23.62529                      Mean          24.73744
                        Largest       Std. Dev.     5.141812
75%     26.95946       52.99345
90%     31.32101       53.21151       Variance      26.43823
95%     35.03738       56.48531       Skewness      1.536935
99%     41.97015       63.49296       Kurtosis      7.081053
```

The variance is now displayed in the lower right, and for the bmi variable, it is 26.44. The median, however, is not definitively labeled. Instead, Stata lists the major percentiles. The 50% value, or the 50th percentile, is equivalent to the median value. For bmi, the median and mean are relatively similar, suggesting a normal distribution of values. The IQR can also be easily calculated by subtracting the 25th percentile value (21.18) from the 75th percentile value (26.95), producing a figure of 5.68. In addition to numerous percentiles, the -detail- option also lists the Skewness and Kurtosis scores, which are two additional measures of variability. A normal distribution has a skewness score of 0 and a kurtosis score of 3. If the skewness score is less than 0, it suggests a negatively skewed distribution (i.e., the mean is less than the median), while a skewness score greater than 0 suggests a positively skewed distribution (i.e., the mean is greater than the median). In both cases, larger absolute values of a skewness score indicate a more extreme skewness. Kurtosis is a measure of the severity of the peak of the distribution, with larger values showing a stronger peaked distribution and smaller values indicating a flatter distribution. The `bmi` variable is somewhat positively skewed (1.54), and there is a relatively strong peak in the distribution based on the kurtosis score of 7.08.

With or without the `-detail-` option, the `-sum-` command can handle multiple variables listed in a single command line. For example, you could type `sum bmi agecats, detail` into the Command window and press **Enter**. The detailed summary statistic results would be presented, one after the other, for both variables. One drawback of this method is that it is difficult to quickly see similar figures (e.g., the mean) across several variables, and the `-detail-` option may list several statistics that are not needed.

To present a more consolidated table of the measures of central tendency and variability across numerous variables *and* control the statistics that are displayed, the `-tabstat-` command is an effective alternative. By default, without any options, the `-tabstat-` command only reports the mean. Displaying additional statistics is controlled by the `-statistics(statname)-` option (shortened `-stat(statname)-`). The "`statname`" in this option refers to a particular code for the statistic you would like to be presented. As with most of the Stata commands you have encountered, the code for each statistic is intuitively straightforward (e.g., the code for the mean is `mean`). See the "A Closer Look" box for a full list of the available statistics and their codes.

A CLOSER LOOK: STATISTICS AND THEIR CODES FOR USE WITH TABSTAT, STAT (STATNAME)

The following table lists all the statistics that can be displayed with the `-tabstat-` command and their associated codes that are typed in the `-stat(statname)-` option.

Statistic	statname Code
Mean	mean

(*Continued*)

(*Continued*)

count of nonmissing observations	count
same as count	n
Sum	sum
Maximum	max
Minimum	min
range = max – min	range
standard deviation	sd
Variance	variance
coefficient of variation (sd/mean)	cv
standard error of mean (sd/sqrt(n))	semen
Skewness	skewness
Kurtosis	kurtosis
1st percentile	p1
5th percentile	p5
10th percentile	p10
25th percentile	p25
median (same as p50)	median
50th percentile (same as median)	p50
75th percentile	p75
90th percentile	p90
95th percentile	p95
99th percentile	p99
interquartile range = p75 – p25	iqr
equivalent to specifying p25 p50 p75	q

Source: Table courtesy of Stata Manual.

To present the set of measures of central tendency and variability listed at the beginning of this section for both the `bmi` and `agecats` variable, type tabstat bmi agecats, stat(mean median min max range iqr variance sd). The following results are displayed:

```
tabstat bmi agecats, stat(mean median min max range iqr
variance sd)
```

```
         Stats |       bmi    agecats
---------------+----------------------
          mean |  24.73744   20.01817
           p50 |  23.62529         20
           min |  14.01495         17
           max |  63.49296         24
         range |  49.47801          7
           iqr |  5.677233          2
      variance |  26.43823   2.088963
            sd |  5.141812   1.445324
               |
---------------+----------------------
```

The same values for the bmi variable are listed as when the -sum- command was used, but the -tabstat- command presents them in a more concise fashion. It is also much easier to see the pertinent figures across the two variables. As when using the -detail- option, the median is labeled by the 50th percentile, or p50. The standard deviation is given by sd.

A potential problem arises when using the -tabstat- command in this manner when several variables are entered at one time. Eventually, the display becomes too wide to fit on the screen. In this case, it may be more effective to also invoke the -columns(variables/statistics)- option (shortened -col(var/stat)-). When you do not use this option, it is automatically set as -columns(variables)-, and the variables are organized in the columns, as shown above.

If you type tabstat bmi agecats, stat(mean median min max range iqr variance sd) col(stat) in the Command window (remember you can use the **Page Up** function, the Review window, or a do file to eliminate retyping this lengthy command) and press **Enter**, the results are now displayed as follows:

```
tabstat bmi agecats, stat(mean median min max range iqr
variance sd) col(stat)
```

```
          |
 Variable |     mean       p50       min       Max     range       iqr  variance        sd
----------+-------------------------------------------------------------------------------
      bmi | 24.73744   23.6252  14.01495  63.49296  49.47801  5.677233  26.43823  5.141812
  agecats | 20.01817        20        17        24         7         2  2.088963  1.445324
----------+-------------------------------------------------------------------------------
```

You can see that this modified display would be even more effective if you were examining a large number of variables at once.

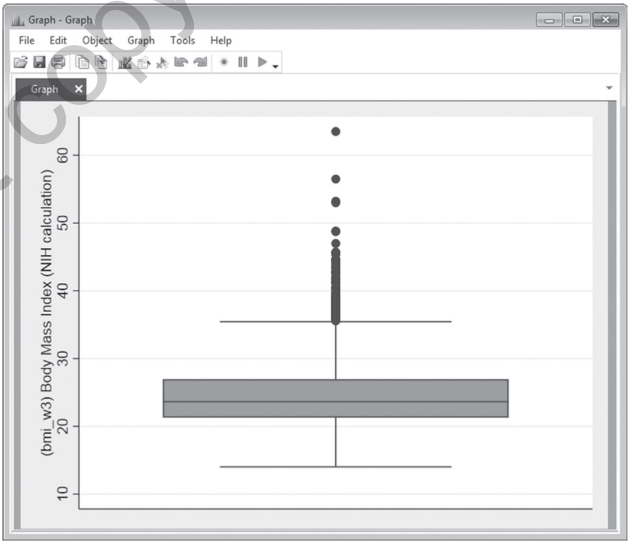
Box Plots

A box plot is a graphical presentation of all the information displayed in the -tabstat- table above. This type of plot is especially useful for examining the overall distribution of an interval-ratio variable and determining whether that distribution is skewed or affected by outliers.

A box plot is one graph that is almost as easy to produce using the Command window as it is through the point-and-click method. The command is -graph box-, followed by the variable name(s) that you want to produce a box plot of. It is possible to enter multiple variables at one time, and all the box plots will be presented on one graph. However, this type of box plot should only be used when the variables are measured with similar units. For example, it would not make a great deal of sense to place the box plots for bmi and agecats side by side because the units are not equivalent.

To produce a box plot of bmi using the Command window, type graph box bmi in the Command window, and press **Enter**. The graph as shown in Figure 4.10 will be displayed.

FIGURE 4.10 ● BOX PLOT OF bmi VARIABLE

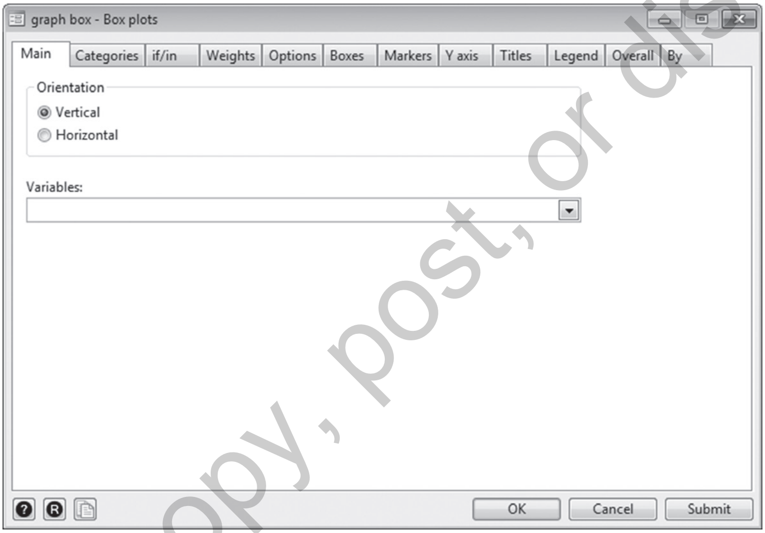


```
graph box bmi
```

The box plot of `bmi` shows that the distribution is relatively normal, as the line indicating the median is close to the center of the box, and the box itself is approximately in the middle of the range. The plot further shows that outliers may be a concern as there are several extreme values, represented by the dots, at the high end of the distribution.

Producing a similar plot using the point-and-click interface is similarly straightforward. Click on the **Graphics** menu and then **Box Plot**. The window as shown in Figure 4.11 will appear.

FIGURE 4.11 ● BOX PLOT MAIN OPTIONS WINDOW


graph box - Box plots

Orientation

● Vertical

○ Horizontal

Variables:

In the **Variables** box, either use the drop-down menu to find or type in `bmi`, and then click **OK**. The same box plot as shown above will be displayed.

A CLOSER LOOK: USING HISTOGRAMS TO EXAMINE CENTRAL TENDENCY AND VARIABILITY

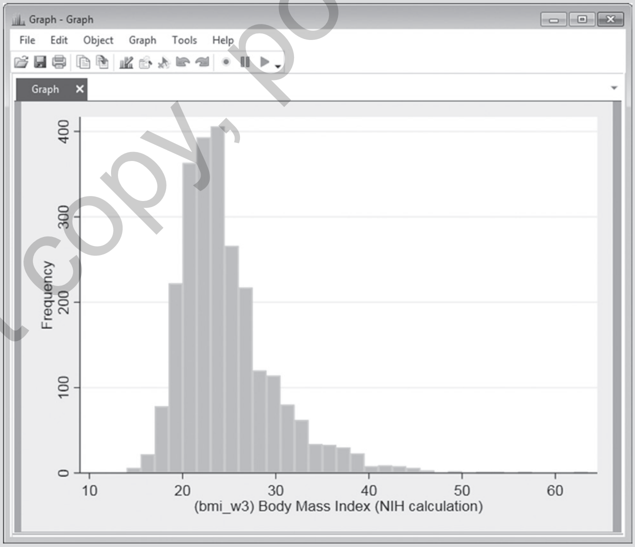
Creating histograms using the point-and-click method was discussed earlier, but these graphs are also useful for quickly examining the distribution of interval-ratio variables. Furthermore, because interval-ratio variables do not use value labels (i.e., their values are their labels), these graphs are easier to produce using the Command window interface when examining such variables.

The command to produce a histogram, as you may intuitively suppose, is `-histogram-` (shortened `-hist-`). To quickly produce a histogram of the `bmi` variable, you could type `hist bmi` in the Command window. But, as with the point-and-click method, the default *Y*-axis scale is each category's density. This scale can be changed by invoking either the `-percent-` or `-frequency-` (abbreviated `-freq-`) option.

For example, type `hist bmi, freq` into the Command window, and press **Enter**. The graph as shown in Figure 4.12 will be displayed.

```
hist bmi, freq
```

FIGURE 4.12 ● HISTOGRAM OF `BMI` VARIABLE



As you can see, the histogram very clearly illustrates the shape of the distribution. Similar to the other tools that have been used, the graph shows that the distribution is essentially normal with a few outliers at the high end of the distribution.

Summary of Commands Used in This Chapter

Frequency Distributions

```
tab body
tab body, mis
tab body, sort
tab1 body sad
tab gender
tab gender, nol
tab body if gender==1, sort
```

Measures of Central Tendency and Variability

```
sum bmi
sum bmi, detail
tabstat bmi agecats,
stat(mean median min max
range iqr variance sd)
tabstat bmi agecats,
stat(mean median min max
range iqr variance sd)
col(stat)
```

Box Plots

```
graph box bmi
```

A Closer Look: Using Histograms to Examine Central Tendency and Variability

```
hist bmi, freq
```

Exercises

Use the original `Chapter 4 Data.dta` for the following problems. (*Optional:* Complete the exercises using a do file and save the results using a log file. See Chapter 3 for an explanation of how to use these files.)

1. Produce a frequency distribution of how important religion is to young adults (`faith1`).
2. Display another frequency distribution of the `faith1` variable, including missing categories, that is arranged with the most frequent category displayed first.
3. Produce a frequency distribution for both how important religion is to young adults (`faith1`) and how much young adults care about the elderly (`crelder`) with one command.
4. Produce a frequency distribution for the variable `crelder` for respondents who think religion is extremely or very important (`faith1`).
5. Produce a percentage histogram of the `faith1` variable that uses value labels and has a *Y*-axis that ranges from 0 to 100.
6. Show the detailed measures of central tendency and variability for the number of children young adults desire (`kidwntmn`).
7. Produce a table, with the variables in the rows, that displays the mean, median, standard deviation, and variance for the `kidwntmn` variable and the number of religious retreats young adults have attended (`relretrt`).
8. Generate a box plot for the `kidwntmn` variable.
9. Produce a frequency histogram (using the Command window) of the `relretrt` variable.