

2. Examining Distributions

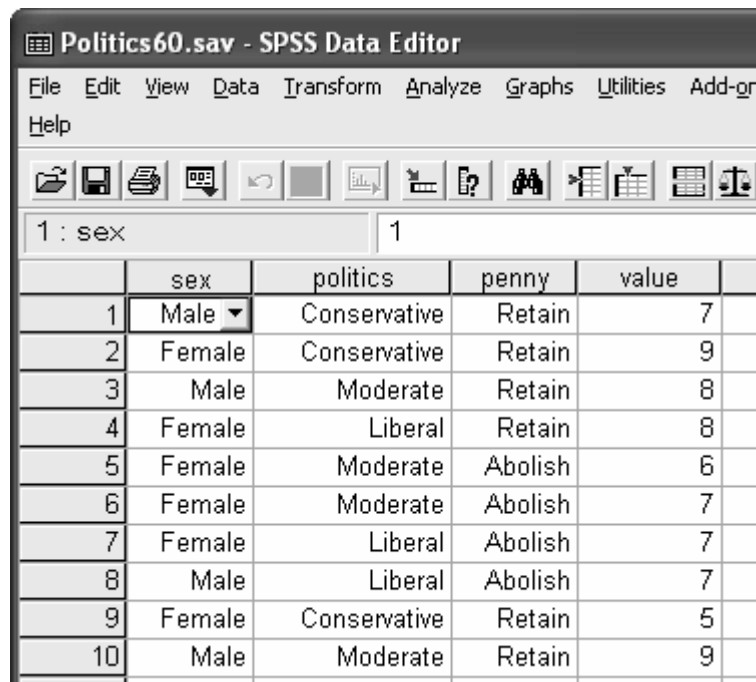
This chapter describes procedures for examining categorical and quantitative variables graphically and numerically. The procedures in this chapter are for examining variables one at a time. The next chapter describes procedures for looking at the relationship between two variables.

2.1 Methods for categorical variables

For categorical variables, a frequency table summarizes the number and percent of responses in each category. A bar graph (called a bar chart in SPSS) is a graphical display of a frequency table.

Frequency tables

Example 2-1: Student Views. Consider a data set obtained from 23 students in an introductory statistics class. It consists of four variables: *sex* (categorical: male or female), *politics* (categorical: liberal, moderate, or conservative), *penny* (categorical: retain or abolish), and *value* (quantitative: value of statistics). The variables *sex*, *politics*, and *penny* were entered as numerical values and then value labels were used to attach labels to the numerical categories (Section 1.9). For example, *politics* was coded as 1, 2, 3 (Conservative, Moderate, Liberal) and *penny* was coded as 0,1 (Retain, Abolish). The first few cases are shown in Figure 2-1 as they appear in the Data Editor with the option **View... Value Labels** selected. The value labels for *politics* are shown in Figure 2-2.



	sex	politics	penny	value
1	Male	Conservative	Retain	7
2	Female	Conservative	Retain	9
3	Male	Moderate	Retain	8
4	Female	Liberal	Retain	8
5	Female	Moderate	Abolish	6
6	Female	Moderate	Abolish	7
7	Female	Liberal	Abolish	7
8	Male	Liberal	Abolish	7
9	Female	Conservative	Retain	5
10	Male	Moderate	Retain	9

Figure 2-1

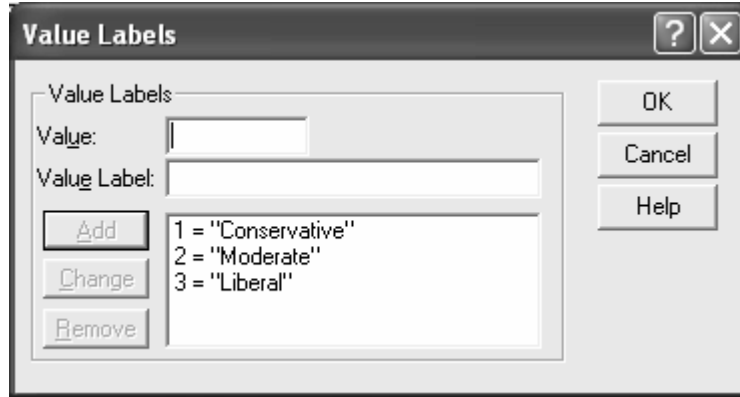



Figure 2-2

To obtain a frequency table in SPSS for a single categorical variable such as *politics*, follow these steps.

1. Click **Analyze**, click **Descriptive Statistics**, then click **Frequencies**. The SPSS window in Figure 2-3 appears.
2. Click *politics*, then click  to move *politics* into the “Variable(s)” box.
3. Click **OK**.

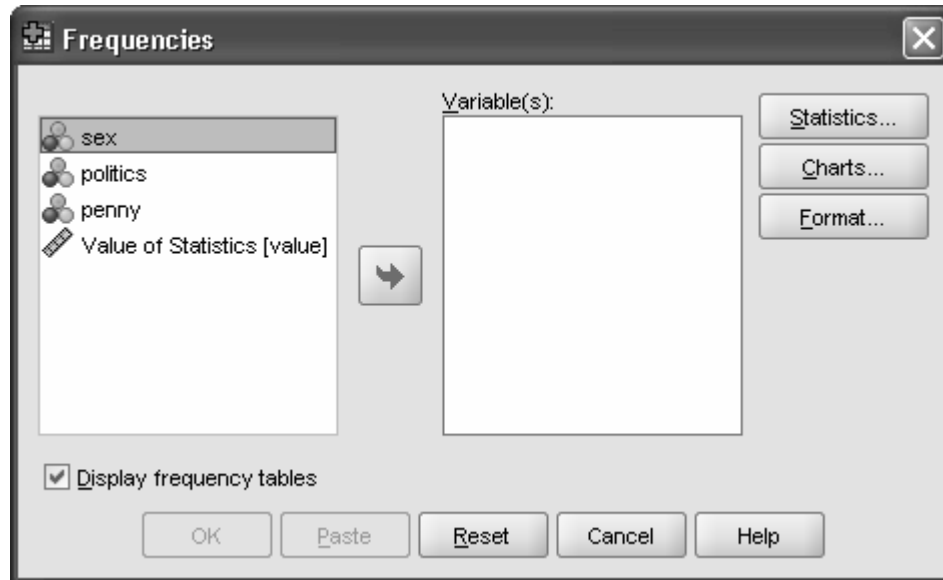


Figure 2-3


Table 2-1 is part of the resulting SPSS output. It indicates that 5 of the students, or 21.7% of the 23 students responding, identified themselves as Conservative, 12 of the students, or 52.2%, identified themselves as Moderate, and 6 of the students, or 26.1%, identified themselves as Liberal.

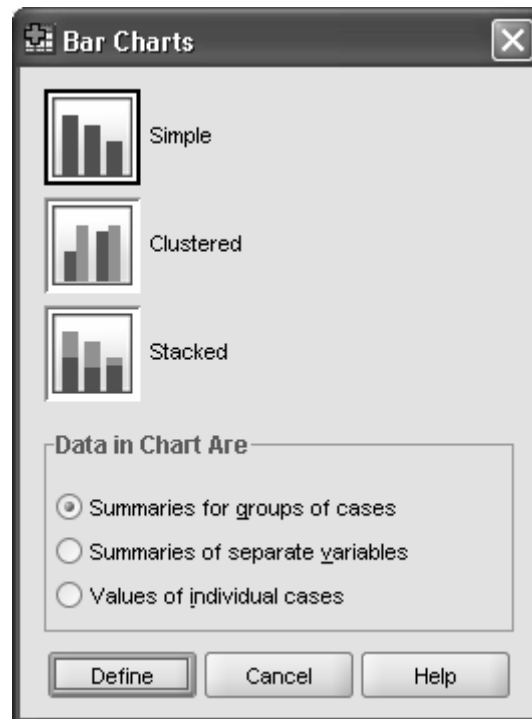
POLITICS

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Conservative	5	21.7	21.7	21.7
	Moderate	12	52.2	52.2	73.9
	Liberal	6	26.1	26.1	100.0
	Total	23	100.0	100.0	

Table 2-1**Bar graphs**

A bar graph (called a bar plot in SPSS) is a graphical display of a categorical variable. In **Example 2-1: Student Views** (page 16), follow these commands to obtain a bar graph of the variable *politics*.

1. Click **Graphs**, **Legacy Dialogs**, and then click **Bar**. The SPSS window in Figure 2-4 appears. Note that the Simple bar chart is highlighted by default; click on it if it is not.
2. Click **Define**. The SPSS window in Figure 2-5 appears.
3. Click *politics*, then click  to move *politics* into the “Category Axis” box.
4. By default, the bars represent the number of cases. If you are interested in having the y axis labeled as “Percent” rather than “Count”, click **% of cases** in the “Bars Represent” box. This will not change the shape of the bar graph, only the labeling of the y axis.
5. Click **OK**.

**Figure 2-4**

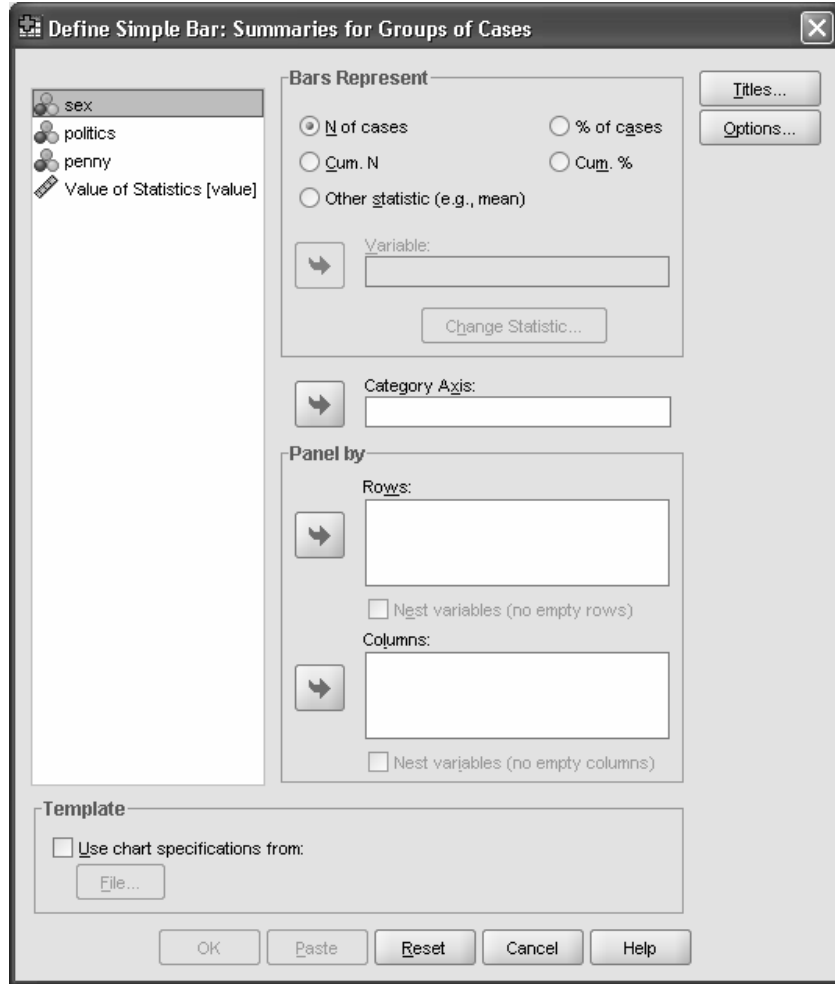


Figure 2-5

Figure 2-6 is the resulting output (with the default title omitted). Note: the bars will probably be red in versions of SPSS 11.5 or earlier when you view them on your monitor, but obviously, they will not print red if you do not use a color printer. In SPSS versions 12.0 - 16.0, the bars will be light brown.

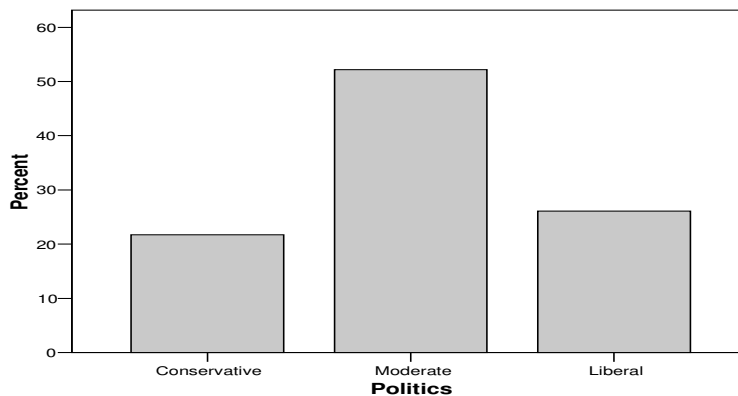


Figure 2-6

Editing bar graphs

You can change many aspects of a bar graph by using the SPSS Chart Editor. To start the editor, double-click on the bar chart in the SPSS Output Window.

For SPSS versions 12.0 – 16.0: The “Chart Editor” window in Figure 2-7 appears.

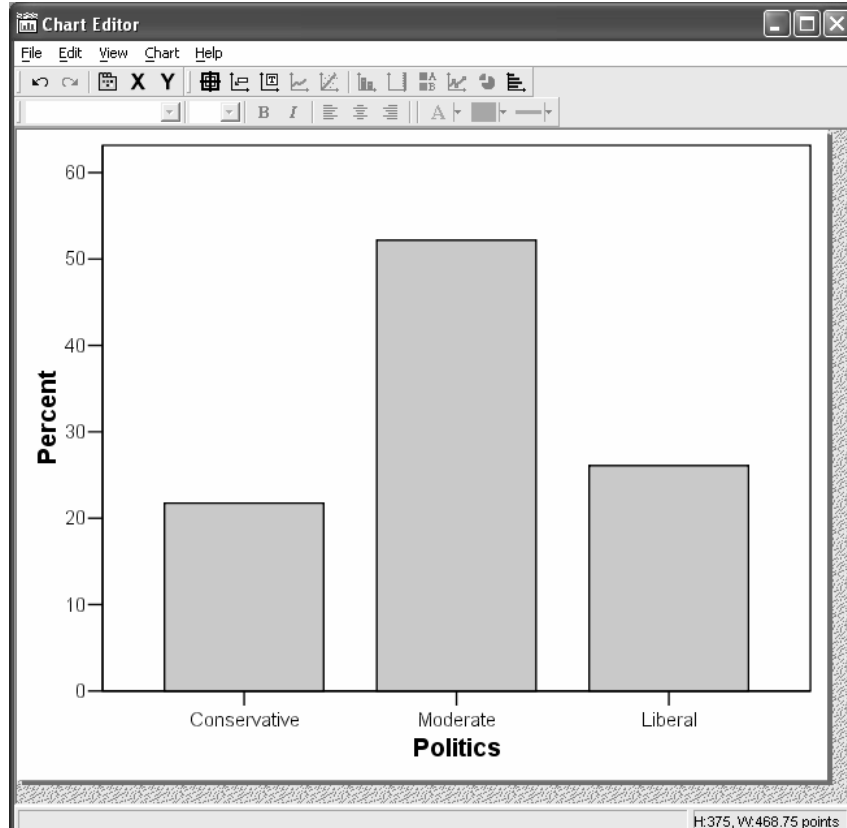



Figure 2-7

Many menu options and icons appear at the top of the Chart Editor. For now, we will illustrate how to change the color of the bars. Follow these steps:

1. Click on the area within the chart for which a color change is needed. For example, click inside any bar to select all the bars.
2. Click the color box  on the toolbar to bring down the menu shown in Figure 2-8.
3. Click on the desired color for the selected bars. Generally, light gray or white are good choices even if you have a color printer. Using a bright color serves no purpose here except to distract from the data because you do not need colors to display additional information.
4. If you are finished editing the chart, click **File** and then click **Close** to return to the “Output” window.

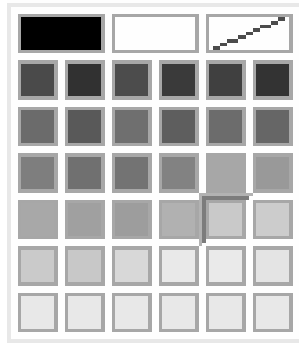


Figure 2-8

For SPSS versions 11.5 or earlier: The “Chart Editor” window in Figure 2-9 appears.

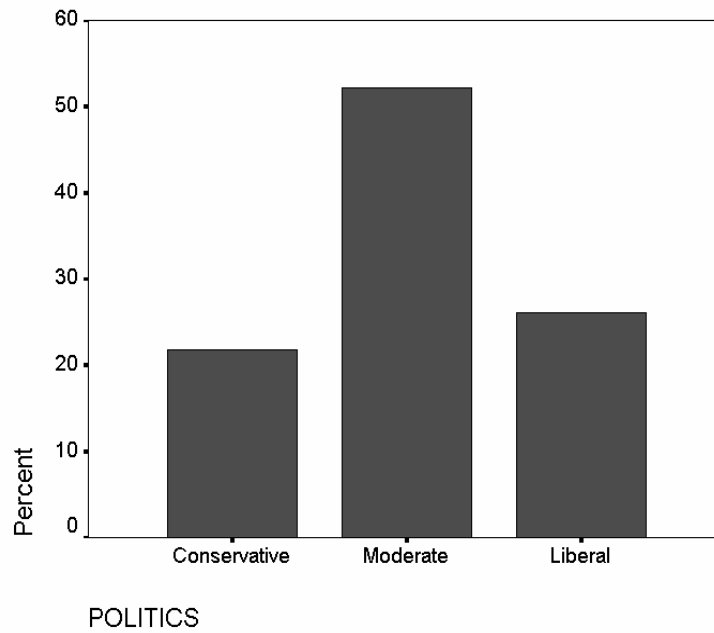



Figure 2-9

Many menu options and icons appear at the top of the Chart Editor. For now, all we might possibly need is to change the color of the bars. Follow these steps.

5. Click on the area within the chart for which a color change is needed. For example, click inside any bar to select all the bars.
6. Click  on the toolbar. The SPSS window in Figure 2-10 appears.
7. Make sure that “Fill” rather than “Border” is selected within the “Color” box. Click on the desired color for the fill. Light gray or white are good choices even if you have a color printer. Using a bright color serves no purpose here except to distract from the data because you do not need colors to display additional information.
8. Click **Apply**.
9. Click **Close**.
10. If you are finished editing the chart, click **File** and then click **Close** to return to the “Output” window.

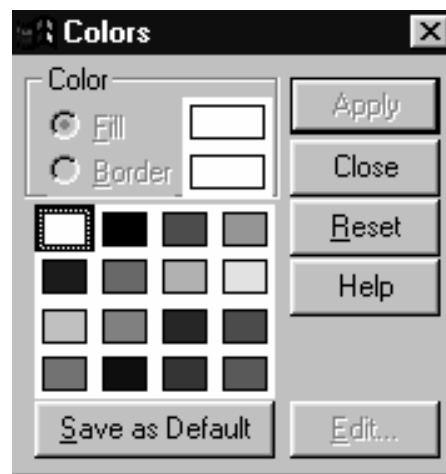


Figure 2-10

Pie charts

A pie chart is an alternative to the bar graph for displaying a categorical variable. To create a pie chart in SPSS, click **Graphs, Legacy Dialogs**, and click **Pie**. The procedure is then almost exactly the same as for creating bar graphs (page 18 of this manual).

2.2 Methods for quantitative variables


For quantitative variables, numerical summaries include the mean, standard deviation, and the five-number summary. Histograms and time plots (for data measured over time) are the primary graphical displays. Boxplots, described on page 37 of this manual, are another graphical display for quantitative variables which are most useful for comparing distributions.

Descriptive statistics for quantitative variables

There are several ways to get descriptive statistics, such as the mean, median, standard deviation, and five-number summary for quantitative variables. A method using the **Frequencies** procedure is described here. Another method using **Explore** is discussed in this section under the heading **Comparing distributions (Numerical comparisons)** on page 35.

Example 1-1: Student Measurements (page 5), continued: obtain descriptive statistics for students' heights.

First method for obtaining descriptive statistics for a quantitative variable:

1. Click **Analyze**, then click **Descriptive Statistics**, then click **Frequencies**. The SPSS window in Figure 2-11 appears.
2. Click *height* and click  to move *height* into the "Variable(s)" box. You can obtain summary statistics for several variables at the same time by moving additional variable names to the "Variable(s)" box.
3. Click the **Statistics** button in the "Frequencies" window. The SPSS window in Figure 2-12 appears.
4. Click **Quartiles** (this gives the 25th, 50th, and 75th percentiles), **Std. Deviation**, **Minimum**, **Maximum**, **Mean**, and **Median** so that a check appears in each selected box (the boxes may already have checks if you have already used this command during the current session).
5. Click **Continue**.
6. In the "Frequencies" window (Figure 2-11), click on **Display frequency tables** so that this box no longer has a check. If you forget to turn off this option, the output will include a table of every distinct value in the data set. This can be quite large for a large data set.
7. Click **OK**.

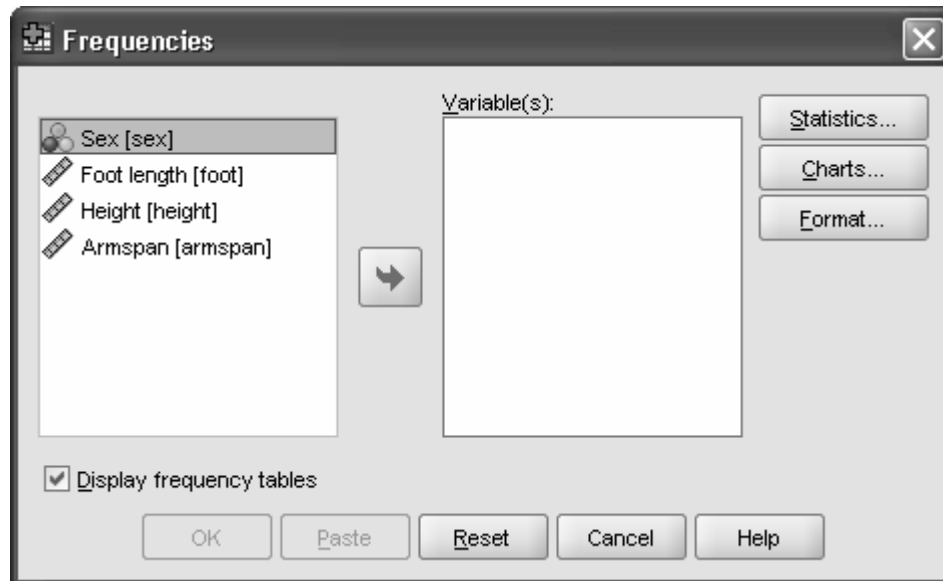


Figure 2-11

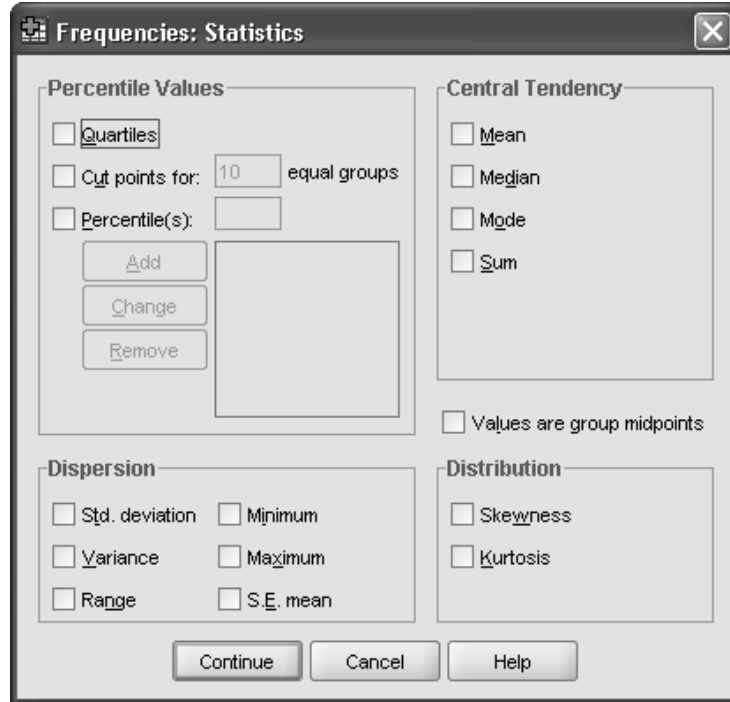


Figure 2-12

The SPSS output is shown in Table 2-2. The mean height of the students is 164.84 cm. and the standard deviation is 7.61 cm. The median height is 165 cm. The first quartile (denoted Q1 in the text) is 159. The third quartile (denoted Q3 in the text) is 170.5, giving an IQR of 11.5 cm. The five-number summary is (152, 159, 165, 170.5, 180).

Statistics

HEIGHT		
N	Valid	25
	Missing	0
Mean		164.84
Median		165.00
Std. Deviation		7.61
Minimum		152
Maximum		180
Percentiles	25	159.00
	50	165.00
	75	170.50


Table 2-2

Histograms

A histogram breaks the range of values of a quantitative variable into equal-width intervals and displays the count or percent in each interval.

Example 1-1: Student Measurements (page 5), continued: obtain a histogram of students' heights.

To create a histogram of *height*, follow these steps.

1. Click **Graphs, Legacy Dialogs,** and then click **Histogram.** The SPSS window in Figure 2-13 appears.
2. Click *height*, then click  to move *height* to the “Variable” box.
3. Click **OK.**

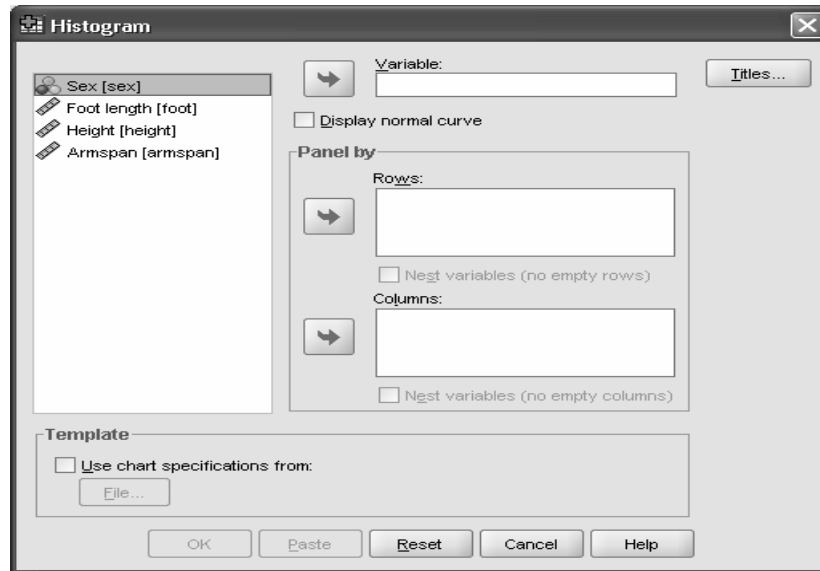


Figure 2-13

Figure 2-14 is the histogram created by SPSS.

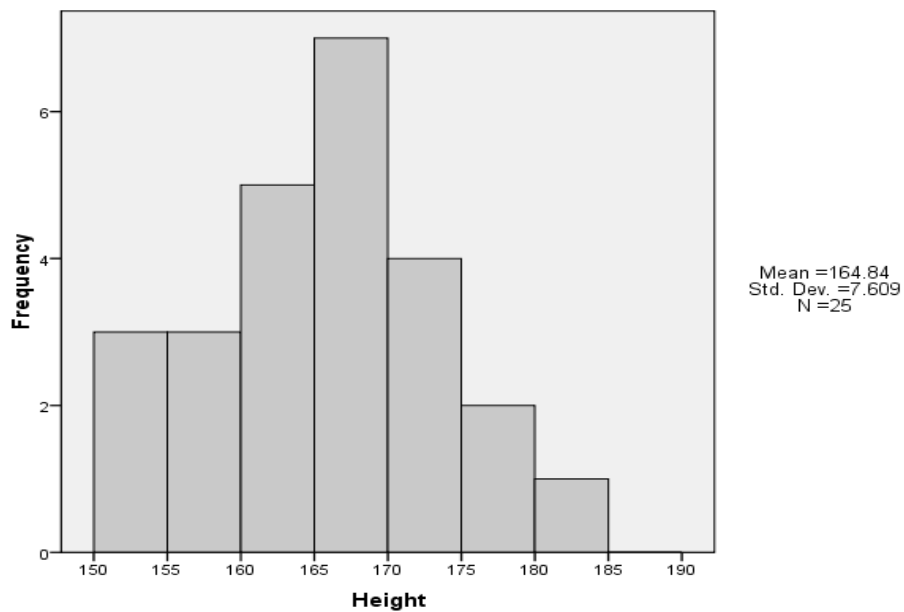


Figure 2-14

Editing histograms

You can change many aspects of a histogram by using the SPSS Chart Editor. To start the editor, double-click on the histogram in the “Output1- SPSS Viewer” window (the Output window). Many menu options and icons appear at the top of the Chart Editor (see Figure 2-7, page 20 or Figure 2-9, page 21). A few of the changes you can make are described here; you can easily discover more on your own.

A shortcut to steps 1 and 2 in the procedures below is to double-click on the part of the histogram you want to change – for example, double-click on the numbers on the x-axis to change the histogram intervals.

The major difference between versions 12.0 – 16.0 of SPSS and earlier versions of SPSS is how chart editing is handled through the SPSS Chart Editor. For this reason, use of the Chart Editor for editing a histogram will first be illustrated for version 16.0 (same for 12.0 – 15.0) and then separately for earlier versions.

For SPSS versions 12.0 - 16.0: To make changes to the *x* axis (such as changing the axis label and the number of bars), follow these steps.

1. Click **Edit** and then click **Select X Axis**, or click the large bold **X** in the icon toolbar. The SPSS window in Figure 2-15 appears.

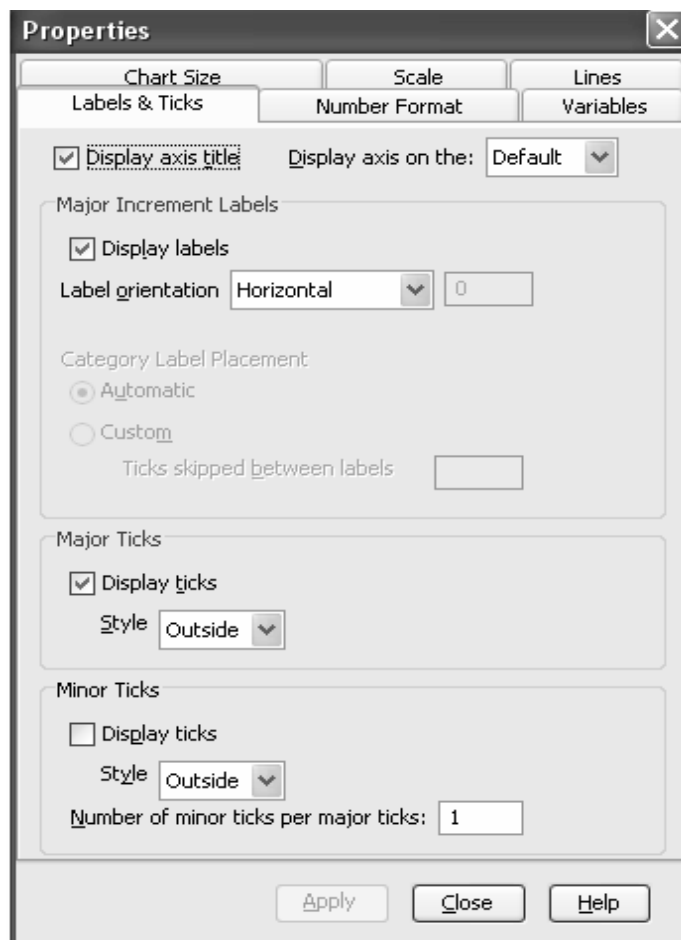


Figure 2-15

2. To change the range of values on the x-axis, click **Scale** at the top of the window. The SPSS window in Figure 2-16 appears. To have a histogram ranging from 150 to 182, for example, change the range maximum from 180 to 182.
3. For SPSS Version 16.0, to change the interval width of the bars, select **Elements** and **Show Data Labels** from the Chart Editor menu. This will add two tabs to the top of the Properties window. First, click **Count** in the “Displayed” window, click the red “X” and click **Apply** to remove the bar counts. Now, select the **Binning** tab at the top of the Properties window. To change the interval width to 4, for example, click **Custom**, then **Interval width**, and change the width to 4.
4. For earlier SPSS versions, to change the interval width of the bars, click **Histogram Options** at the top of the window. The SPSS window in Figure 2-17 appears. To change the interval width to 4, for example, click **Custom**, then **Interval width**, and change the width to 4.
5. Click **Apply**.
6. Click **Close**.

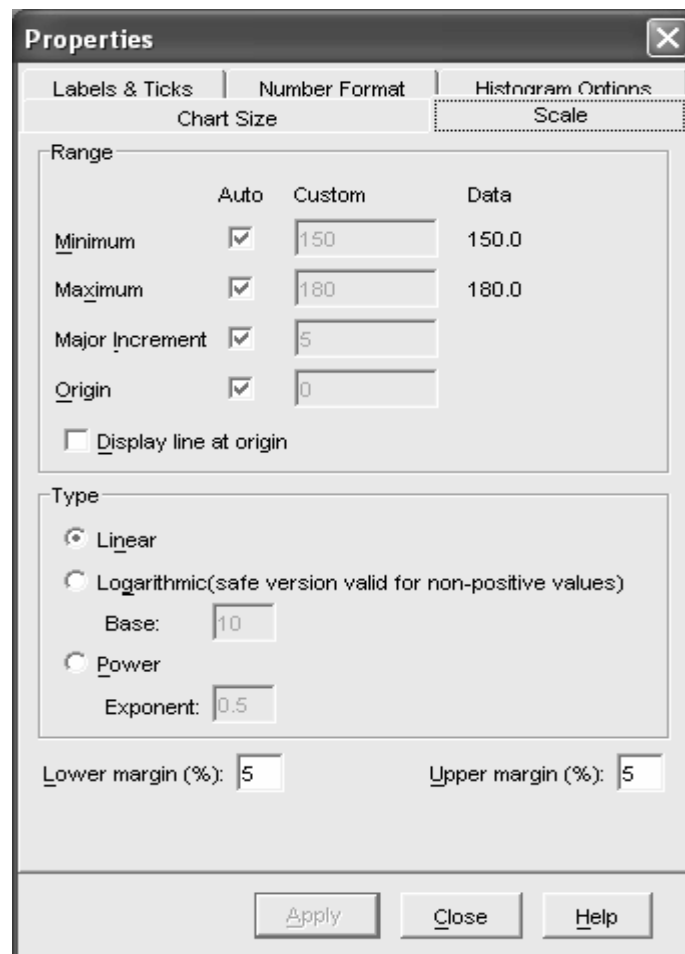


Figure 2-16

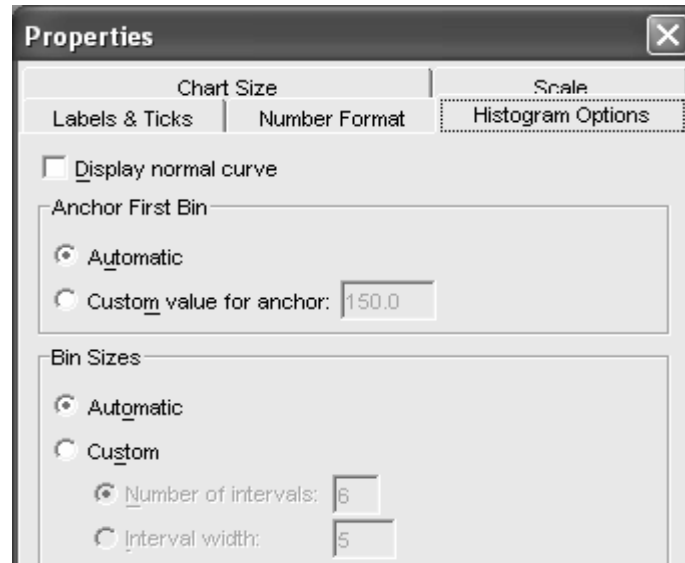


Figure 2-17

To change the title of the x-axis, follow these steps:

1. Click on the current axis title to highlight it.
2. Click on the title a second time and type any changes desired. For example, you might change the title from **Height** to **Height (cm)**. If you have defined a variable label (Section 1.8, page 9) for this variable, then that is what will be used as the axis title.
3. Hit the **Enter** key.

To make changes to the y-axis, such as changing the spacing of the tick marks, follow these steps:

1. Click **Edit** and then click **Select Y Axis**, or click the large bold **Y** in the icon toolbar. The SPSS window in Figure 2-15 appears.

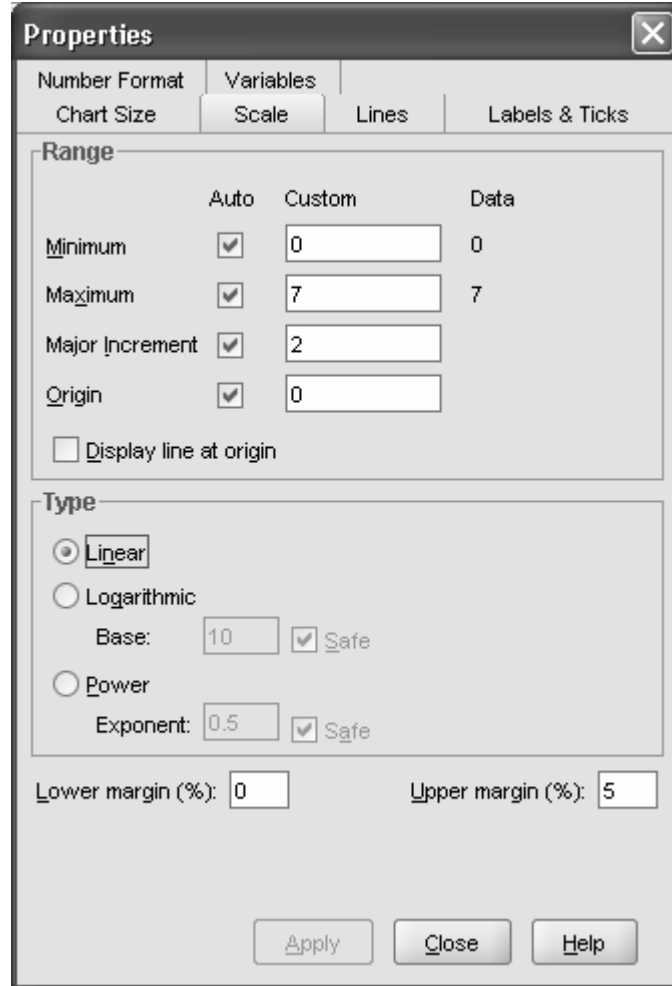


Figure 2-18

2. Click **Scale** at the top of the window if it is not already selected. The SPSS window in Figure 2-18 appears.
3. To change the spacing of the tick marks on the y-axis from 1 to 2, uncheck the **Major** increment box and replace the **1** in the adjacent column with a **2**.
4. Click **Apply**.
5. Click **Close**.

To eliminate the box containing the mean, standard deviation, and sample size, left-click on the text box to highlight it, then right-click to bring down the pop-down menu. Highlight the menu option **Delete** to eliminate the text box. To change the color of any part of the histogram, such as the bars, follow the steps given in Section 2.1, Editing bar graphs (page 20). When you are finished editing the chart, click **File** and then click **Close** to return to the “Output1 – SPSS Viewer” window.

Figure 2-19 is the final result after making all the changes above.

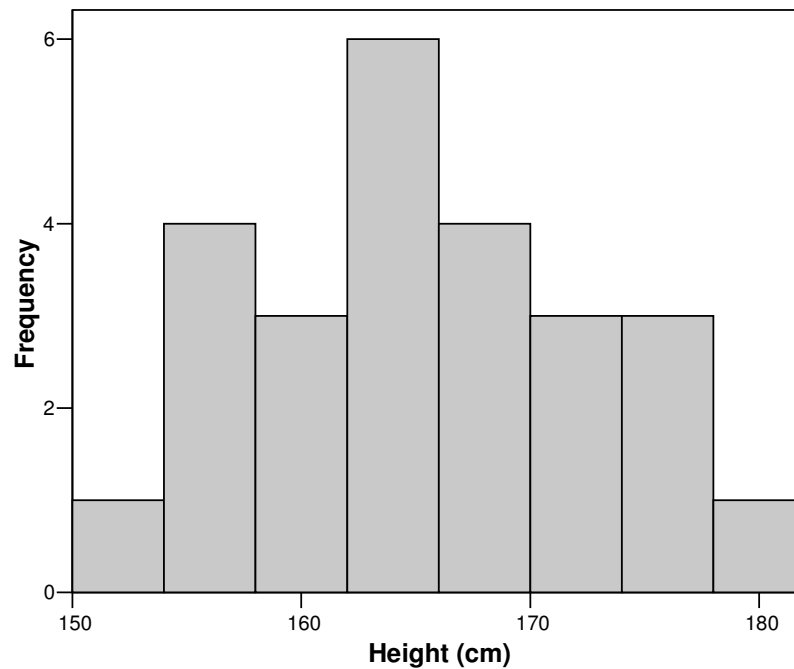


Figure 2-19

For SPSS versions 11.5 or earlier: To make changes to the x-axis (such as changing the axis label and the number of bars), follow these steps.

1. Click **Chart** and then click **Axis**. The SPSS window in Figure 2-20 appears.
2. Click **Interval** and then click **OK**. The SPSS window in Figure 2-21 appears.
3. To change the title of the axis, replace Height with the desired title, for example, **Height (cm)**. If you have defined a variable label (Section 0, page 9) for this variable, then that is what will be used as the axis title.
4. To center the axis title, click ▼ in the “Title Justification” box and click **Center**.
5. To change the range of values on the x axis and/or the interval width of the bars, click **Custom** and then click **Define** located within the “Intervals” box. The SPSS window in Figure 2-22 appears.
6. To change the interval width to 5, for example, click **Interval width** and then type **5** in the “Interval width” box.
7. To change the range of the x axis to 150 to 185, replace 151.25 with **150** in the “Minimum Displayed” box and replace 181.25 with **185** in the “Maximum Displayed” box.
8. Click **Continue**.
9. Click **OK**.

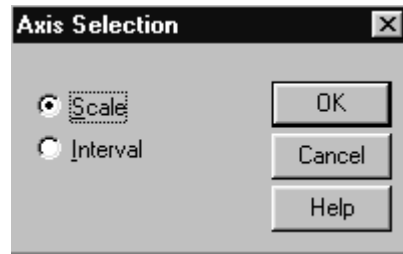


Figure 2-20

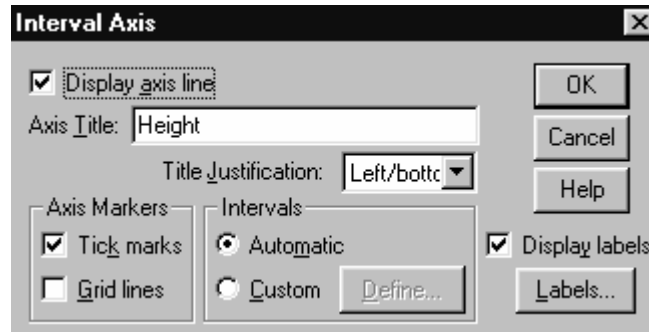


Figure 2-21

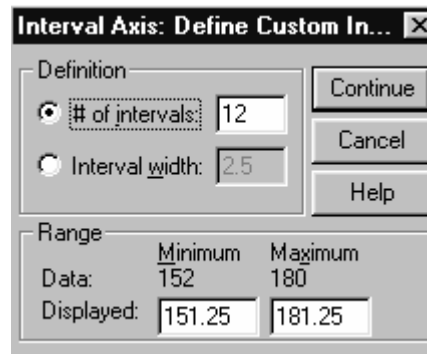


Figure 2-22

To make changes to the y axis, such as adding an axis label and changing the spacing of the tick marks, follow these steps.

1. Click **Chart** and then click **Axis**. The “Axis Selection” window shown in Figure 2-20 appears.
2. Click **Scale** and then click **OK**. The SPSS window in Figure 2-23 appears.
3. To label the y axis, type the desired label in the “Axis Title” box, such as **Frequency**.
4. To center the axis title, click ▼ in the “Title Justification” box and click **Center**.
5. To change the spacing of the tick marks on the y axis from 2 to 1, replace 2 with **1** in the “Major Divisions Increment” box.
6. Click **OK**.

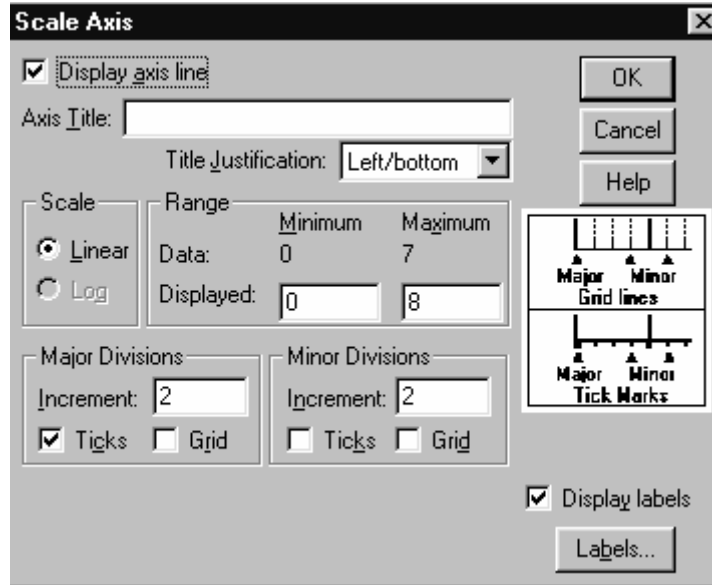


Figure 2-23

To eliminate the box containing the mean, standard deviation and sample size, click **Chart**, click **Legend**, click “Display legend” to unselect it, then click **OK**. To change the color of any part of the histogram, such as the bars, follow the steps given in Section 2.1, Editing bar graphs (page 20). When you are finished editing the chart, click **File** and then click **Close** to return to the “Output1 - SPSS Viewer” window.

Figure 2-24 is the final result after making all the changes above (including changing the bar fill color to light gray).

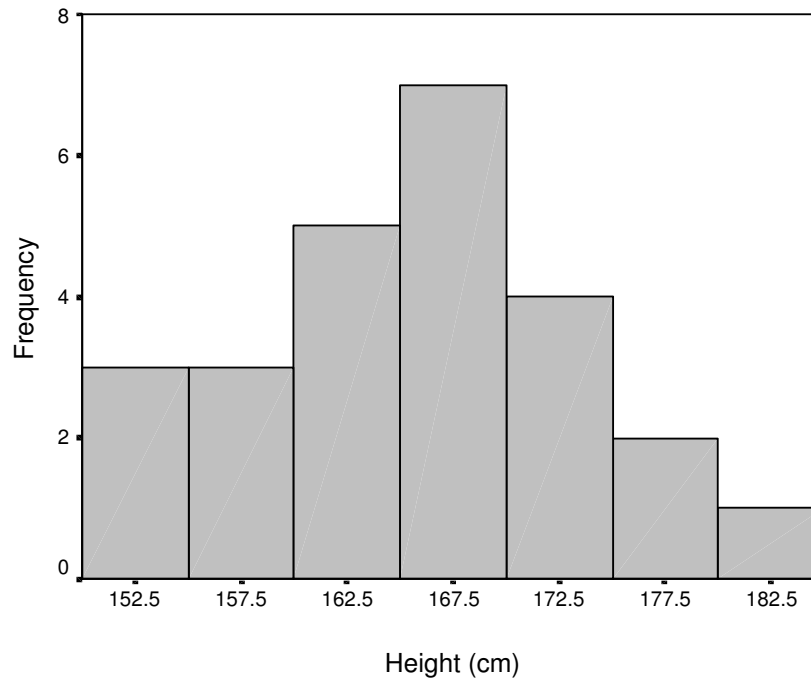


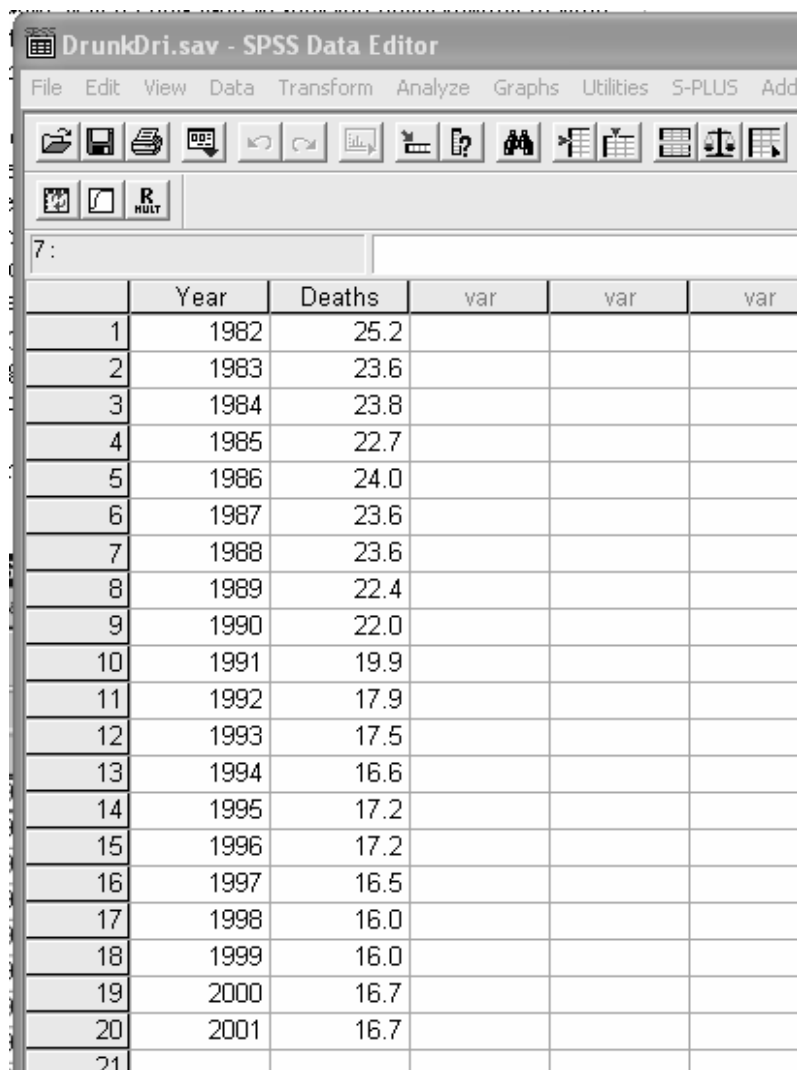
Figure 2-24

Time plots

Many data sets consist of a variable measured at regular intervals over time. With such data, it is a good idea to plot the observations in time order. A time plot puts time on the horizontal (x) axis and the quantitative variable on the vertical (y) axis. It also connects the observations with line segments. This is what makes it different from the scatterplot discussed in Chapter 3.

For time plots, there are usually two variables: the quantitative variable of interest and the time variable which records the time at which each observation occurred. The time variable is optional; if it is not given, then SPSS will plot the observations in the sequence in which they are entered in the data sheet.



Example 2-2: Drunk driving deaths: Data were collected on the annual number of drunk driving deaths (in thousands) in the United States from 1982 through 2001. The data were entered into SPSS using two variables: *year* and *deaths*. The data are shown in Figure 2-25.



	Year	Deaths	var	var	var
1	1982	25.2			
2	1983	23.6			
3	1984	23.8			
4	1985	22.7			
5	1986	24.0			
6	1987	23.6			
7	1988	23.6			
8	1989	22.4			
9	1990	22.0			
10	1991	19.9			
11	1992	17.9			
12	1993	17.5			
13	1994	16.6			
14	1995	17.2			
15	1996	17.2			
16	1997	16.5			
17	1998	16.0			
18	1999	16.0			
19	2000	16.7			
20	2001	16.7			
21					

Figure 2-25

To obtain a time plot of *deaths* against *year*, follow these steps:

1. Click **Analyze**, **Time Series**, and then **Sequence Charts**. The SPSS window shown in Figure 2-26 appears.
2. Click *deaths*, then click  to move *deaths* to the “Variables” box.
3. Click *year*, then click  to move *year* to the “Time Axis Labels” box.
4. Click **OK**.

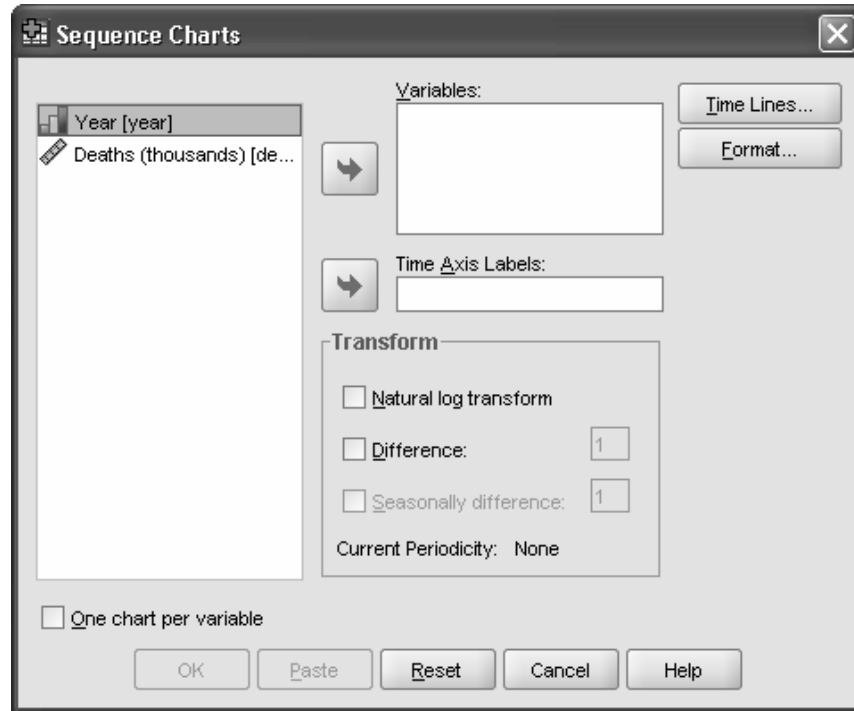


Figure 2-26

Figure 2-27 shows the resulting SPSS output. To edit this plot, use the SPSS Chart Editor as described in the section **Editing histograms** on page 26.

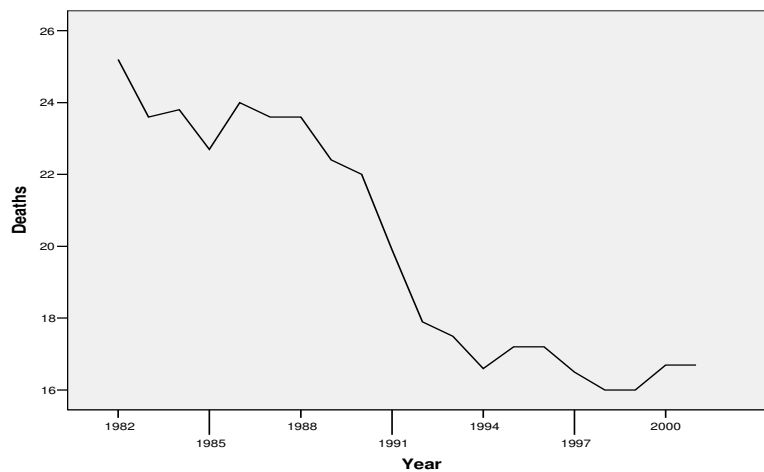




Figure 2-27

Comparing distributions

You can compare groups within a data set on a quantitative variable. There must be a categorical variable (such as sex or political orientation) which defines the groups. Numerical comparisons are accomplished by obtaining descriptive statistics for each group separately. Graphical comparisons are best accomplished through side-by-side boxplots.

Numerical comparisons

Example 1-1: Student Measurements (page 5), continued: compute descriptive statistics for *height* broken down by *sex*. To compute descriptive statistics for *height* broken down by *sex*, follow these steps.

1. Click **Analyze**, then click **Descriptive Statistics** then click **Explore**. The SPSS window in Figure 2-28 appears.
2. Click *height*, then click  to move *height* into the “Dependent List” box.
3. Click *sex*, then click  to move *sex* into the “Factor List” box.
4. By default, the “Display” box in the lower left has **Both** selected. Click **Statistics**.
5. Click the **Statistics** button located in the bottom center of the window. The SPSS window in Figure 2-29 appears.
6. Click **Percentiles** so that it is checked. Be sure that “Descriptives” is also checked.
7. Click **Continue**.
8. Click **OK**.

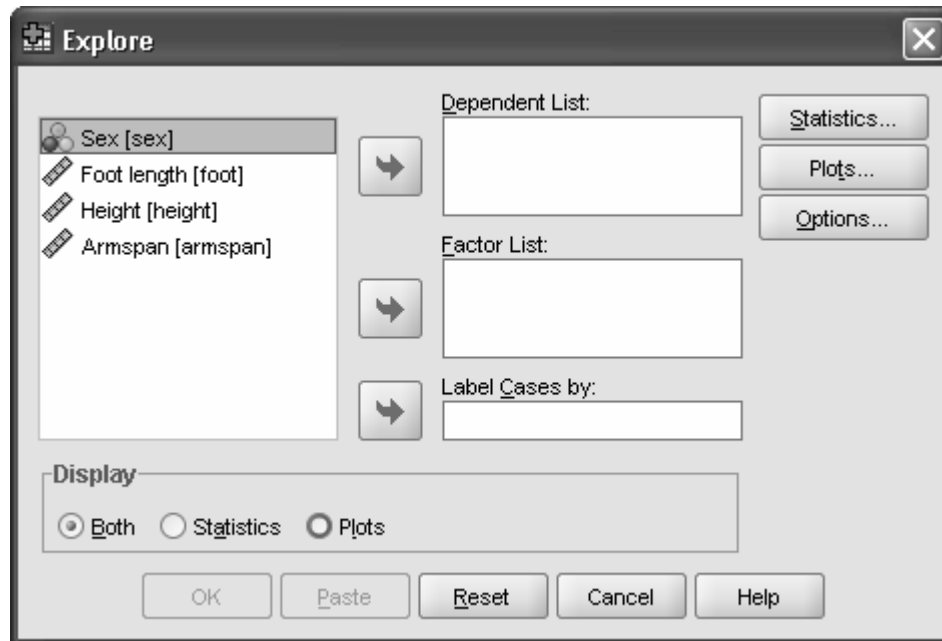


Figure 2-28

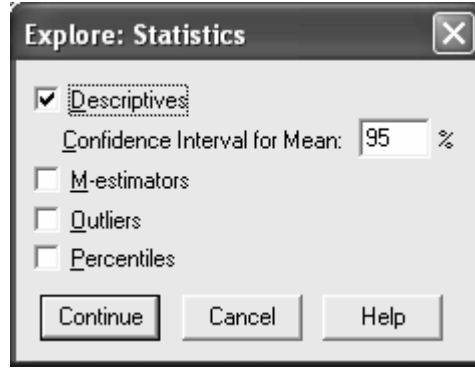


Figure 2-29

Table 2-3 and Table 2-4 are part of the SPSS output. Since we assigned value labels (page 9) to the variable *sex*, these labels appear in the output. We can see that the mean height is 161.83 cm. for females and 172.57 cm. for males; the standard deviation is 6.03 cm. for females and 5.62 cm. for males, etc. The five-number summary is (152, 156.5, 162.5, 166, 174) for females and (165, 166, 173, 177, 180) for males. In Table 2-4, use the lines labeled “Weighted Average” and ignore the lines labeled “Tukey’s Hinges”. Most of the other output in Table 2-3 and Table 2-4 is extraneous and we will not use it.

Descriptives

Sex		Statistic	Std. Error	
Height	Female	Mean	161.83	
		95% Confidence Interval for Mean	1.42	
		Lower Bound	158.83	
		Upper Bound	164.83	
		5% Trimmed Mean	161.70	
		Median	162.50	
		Variance	36.382	
		Std. Deviation	6.03	
		Minimum	152	
		Maximum	174	
		Range	22	
		Interquartile Range	9.50	
		Skewness	.116	.536
		Kurtosis	-.544	1.038
	Male	Mean	172.57	
		95% Confidence Interval for Mean	2.13	
		Lower Bound	167.37	
		Upper Bound	177.77	
		5% Trimmed Mean	172.58	
		Median	173.00	
		Variance	31.619	
		Std. Deviation	5.62	
		Minimum	165	
		Maximum	180	
		Range	15	
		Interquartile Range	11.00	
		Skewness	-.242	.794
		Kurtosis	-1.327	1.587

Table 2-3



Percentiles

			Percentiles						
SEX			5	10	25	50	75	90	95
Weighted Average(Definition 1)	HEIGHT	F	152.00	153.80	156.50	162.50	166.00	170.40	.
		M	165.00	165.00	166.00	173.00	177.00	.	.
Tukey's Hinges	HEIGHT	F			157.00	162.50	166.00		
		M			168.50	173.00	176.50		

Table 2-4

Side-by-side boxplots

Side-by-side boxplots (page 78 of the text) are used to display the distribution of a quantitative variable broken down by a categorical variable. In **Example 1-1: Student Measurements.** (page 5), follow these steps to create side-by-side boxplots of *height* by *sex*.

1. Click **Graphs, Legacy Dialogs,** and then **Boxplot.** The “Boxplot” window shown in Figure 2-30 appears. Note that the “Simple” boxplot is highlighted by default. Click on it if it is not.
2. Click **Define.** The SPSS window in Figure 2-31 appears.
3. Click *height*, then click  to move *height* into the “Variable” box.
4. Click *sex*, then click  to move *sex* into the “Category Axis” box.
5. Click **OK.**

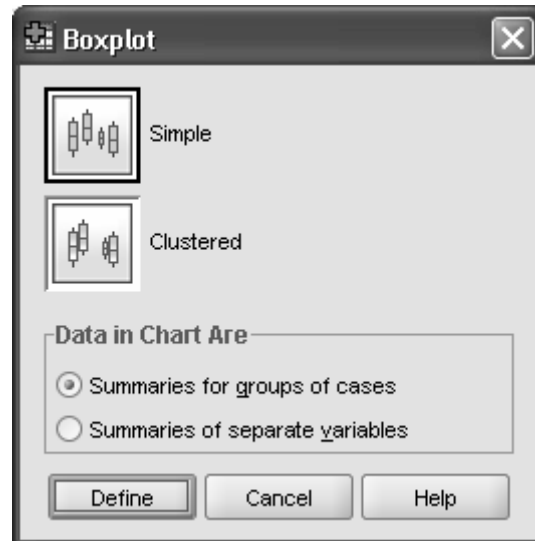


Figure 2-30

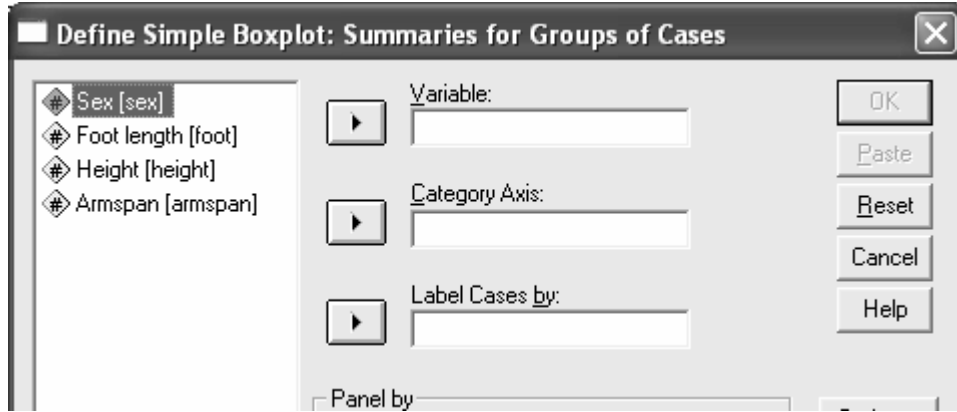


Figure 2-31

Figure 2-32 is the resulting SPSS output (except for the color scheme and axis labels). If there are outliers according to the $1.5 \times \text{IQR}$ criterion (see the roller coaster example on page 188 of the text), SPSS will plot them individually on the boxplot with either an asterisk or a small circle.

To change the color of the boxes, the axis titles, or other aspects of the boxplot, follow the directions given in Editing bar graphs (page 20).

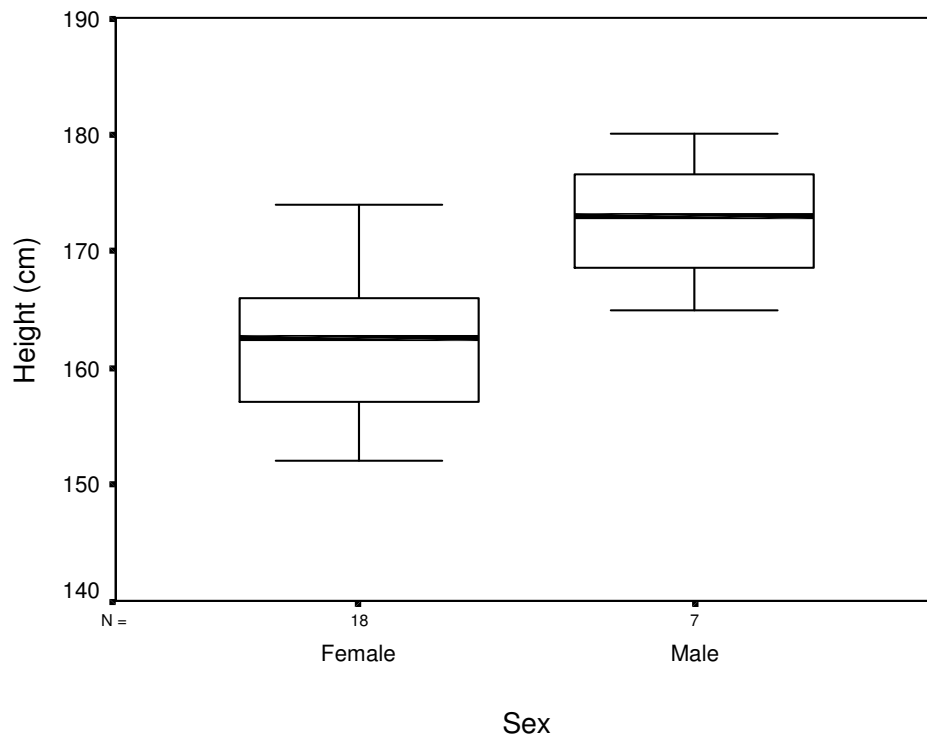


Figure 2-32