Chapter 2

Summarizing and Graphing Data

2-1 Overview
2-2 Frequency Distributions
2-3 Histograms
2-4 Statistical Graphics
2-5 Critical Thinking: Bad Graphs

2-1 Overview

This chapter discusses how SAS Learning Edition (SASLE) may be used to derive tables, and to construct graphs and plots. The objective is to see how these tools may be used to reveal the important characteristics of a set of data. You should be familiar with Chapter 2 of *Elementary Statistics* prior to beginning this chapter.

2-2 Frequency Distributions

Let us say you would like to construct a frequency distribution table for a set of numeric data values. For example, consider the data presented in Table 2-1 of *Elementary Statistics*. The pulse rates are reproduced below.

76, 72, 88, 60, 72, 68, 80, 64, 68, 68, 80, 76, 68, 72, 96, 72, 68, 72, 64, 80, 64, 80, 76, 76, 76, 80, 104, 88, 60, 76, 72, 72, 88, 80, 60, 72, 88, 88, 124, 64

Figure 2-1: Pulse Rates (beats per minute) – Females

These values are used in Section 2-1 of *Elementary Statistics* to illustrate the manual procedure used to construct Table 2-2. There are several ways of constructing such a table with SASLE. One way, is to use a two-step process that involves the use of the SASLE **Create Format** and **One-Way Frequencies** tasks. This approach automates the error prone tallying step of the manual procedure but does not determine class limits. We assume that you have launched SASLE and opened a project. We also assume that the pulse rate data is available as a dataset and has been added to your project. If necessary, see Section 1-2 to see how a dataset may be created.

Step 1: Choose the **Data > Create Format** menu option. Alternatively, you may double-click on the **Create Format** task in the **Task List** window. The following dialog box will appear.

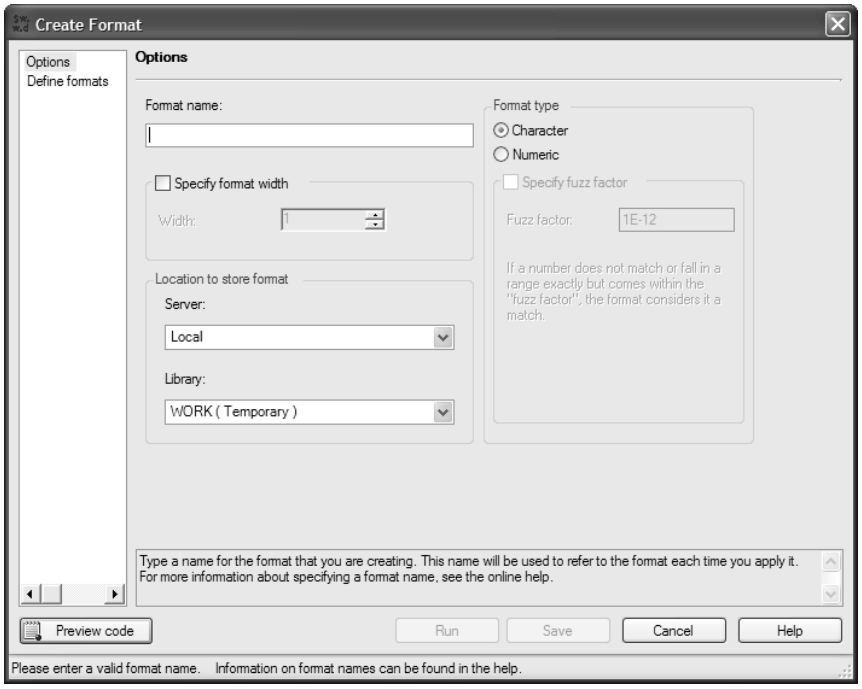

Figure 2-2

Enter a **Format name:** (e.g. CLASSES) and then select the **Numeric** option in the **Format type** box. Select the **Define formats** option from the selection pane. The following dialog box, which allows you to define class limits, will appear:

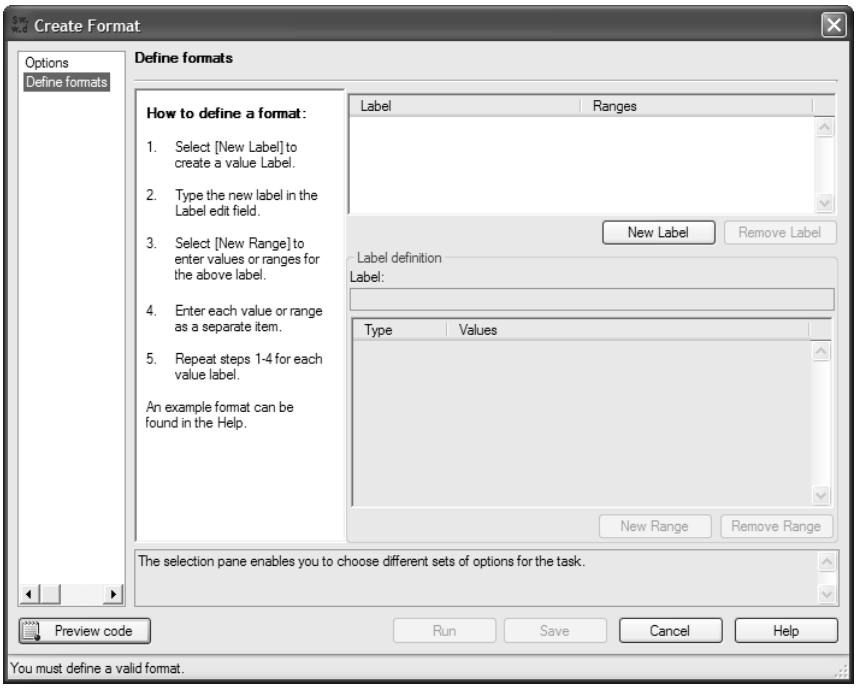

Figure 2-3

Notice the **How to define a format:** window on the left. To create class limits you must follow the five-step procedure outlined in this window for each class limit. For example, to create the 60-69 class limit you must first click on the **New Label** button. Note that you are initially unable to click the **New Range** button. Clicking the **New Label** button will allow you to enter a name (i.e. label) for the 60-69 class limit. You may use any name but, for the sake of convenience, it is probably a good idea to use the text "60-69" as the label. The **New Range** button will now be enabled. Clicking the **New Range** button will allow you to define the range of values for this class. That is, between 60 and 69 inclusive.

The dialog box below shows several class limits already defined.

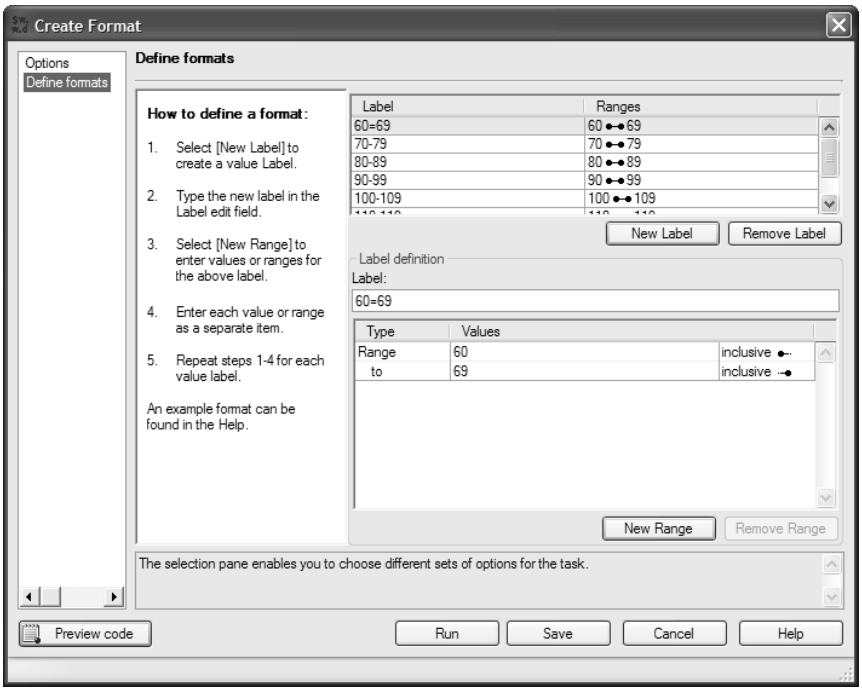

Figure 2-4

Clicking on the **Run** button will create the format. You will notice that the **Workspace** window does not change but the **Project Explorer** window will be updated to reflect the result of this task.

Step 2: Choose the **Describe** > **One-Way Frequencies** menu option. Alternatively, you may double-click on the **One-Way Frequencies** task in the **Task List** window. The following dialog box will appear.

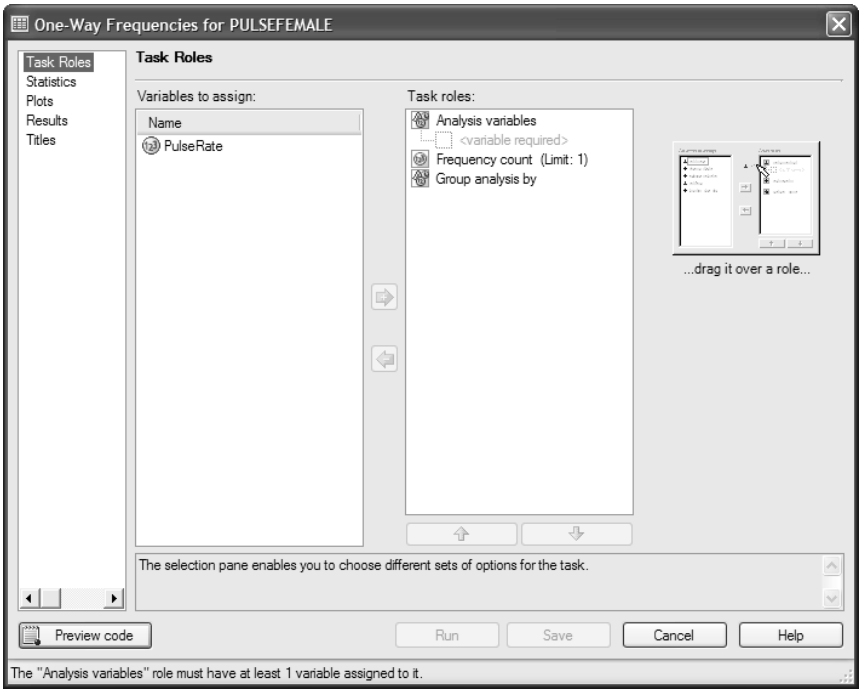

Figure 2-5

Select the **PulseRate** variable, click on the button, and then choose the **Analysis variables** menu option. The **PulseRate** variable now appears in both the **Variables to assign** and **Task roles** windows. If you right click on the **PulseRate** variable (in either window), the following menu appears.

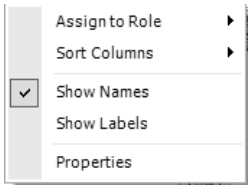

Figure 2-6

Choose **Properties** and then click on the **Change** button adjacent to the **Format** field in the Properties dialog box. The following **Formats** dialog box will then appear.

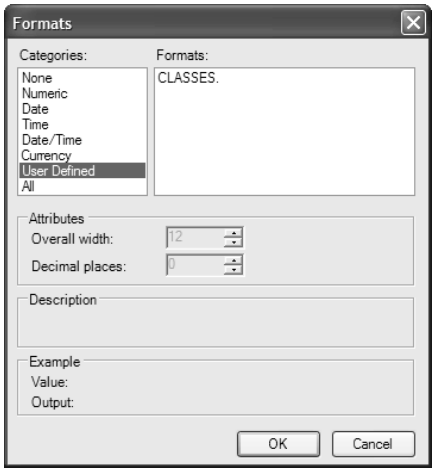

Figure 2-7

Select the **User Defined** option in the **Categories** window. The CLASSES format that was defined in Step 1 above will appear in the **Formats** window. Select it and click the **OK** button. Continue clicking **OK** buttons until the **One-Way Frequencies** dialog box reappears. The following frequency distribution report will be created when you click the **Run** button.

PulseRate	Frequency	Percent	Cumulative Frequency	Cumulative Percent
60=69	12	30.00	12	30.00
70-79	14	35.00	26	65.00
80-89	11	27.50	37	92.50
90-99	1	2.50	38	95.00
100-109	1	2.50	39	97.50
120-129	1	2.50	40	100.00

Figure 2-8

Notice that, in addition to the Frequency column, the table also contains a Percent column and a Cumulative Frequency column. These columns correspond to the Relative Frequency Distribution (Table 2-3) and the Cumulative Frequency Distribution (Table 2-4) tables in *Elementary Statistics*. However, you may customize what appears in the table. By selecting the **Statistics** option from the **One-Way Frequencies** dialog box (see Figure 2-5 above) before clicking the **Run** button you have the option of selecting either **Frequencies only**, **Frequencies and percentages**, **Frequencies and cumulative frequencies**, or **Frequencies and percentages with cumulatives** (i.e. the default option).

The **One-Way Frequencies** task provides several additional options. For example, Table 2-8 in *Elementary Statistics* is a frequency distribution table organized by group. The **Group analysis by** role (see Figure 2-5) is intended for variables that may be used to classify the values of the **Analysis variable** into groups. If this option is used then a separate frequency distribution table will be created for each group.

The **One-Way Frequencies** task may also be used to create frequency distribution tables for categorical data values. The **Create Format** step is not needed in such cases.

2-3 Histograms

In Section 2-3 of *Elementary Statistics* a histogram is defined as a bar graph where the horizontal scale represents classes of data values, the vertical scale represents frequencies, and bars are adjacent to each other. The SASLE **Bar Chart** task is a flexible graphing task that may be used to create a variety of bar graphs (see Figure 2-9 below) including histograms. It is one of several SASLE tasks that may be used to create histograms and is the one that will be discussed here.

Note that you do not need to construct a frequency distribution table before constructing the histogram. Also, you do not need to define class limits as an initial step. The SASLE **Bar Chart** task can automatically determine class limits from the data values. However, if class limits have been determined and are available as a format (see the discussion on the **Create Format** task in Section 2-2) then the format may be used in the **Bar Chart** task to create the histogram. Alternatively, the task allows you to specify either the number of class levels or the class limits themselves.

Again, we will use the pulse rate data to illustrate and assume that SASLE has been launched and a project opened. Choose the **Graph > Bar Chart** menu option. Alternatively, you may double-click on the **Bar Chart** task in the **Task List** window.

The following dialog box will appear.

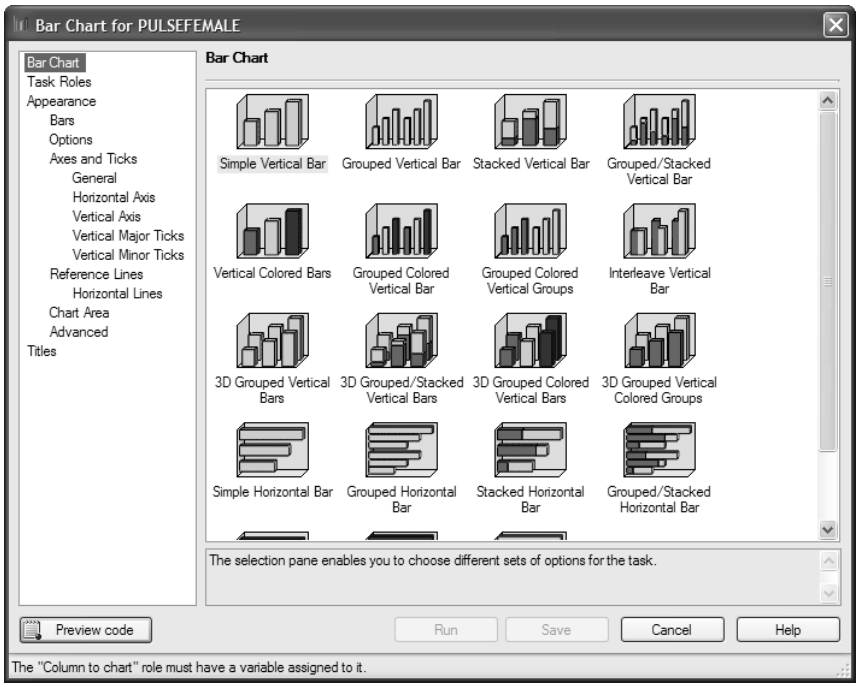

Figure 2-9

Select **Simple Vertical Bar** and then choose the **Task Roles** option. A dialog box similar to Figure 2-5 will appear. Select the **PulseRate** variable from the **Columns to assign** window, click on the  button, and then choose the **Column to chart** menu option. The **PulseRate** variable now appears in both the **Columns to assign** and **Task roles** windows. If class limits are available as a format then they should be associated with the **Column to chart** variable at this point (see Figure 2-6, Figure 2-7, and the corresponding discussion above). If you choose the **Bars** option, the following dialog box appears.

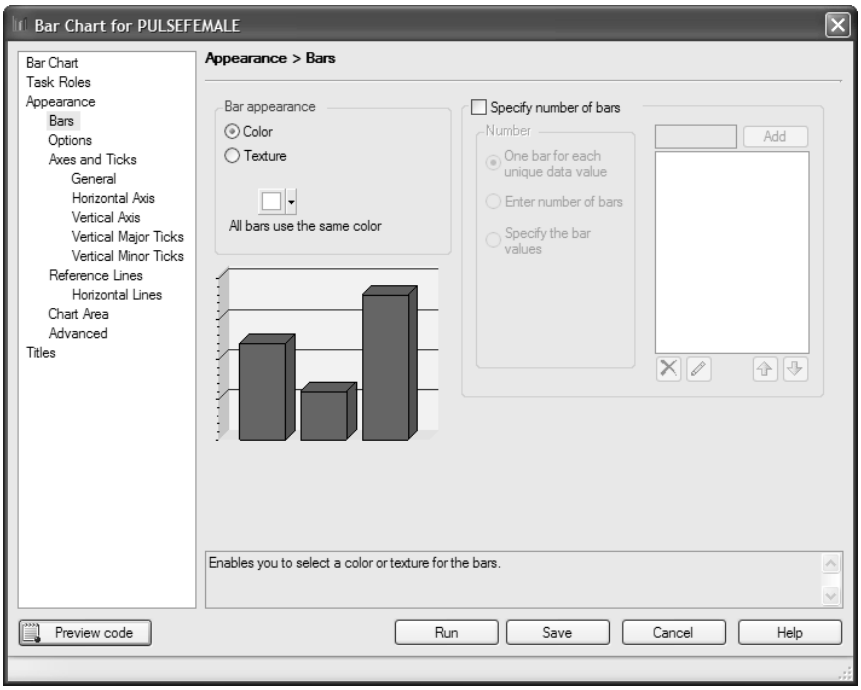

Figure 2-10

Notice the **Specify number of bars** checkbox. This checkbox allows you to specify class limits. If you wish to use class limits previously defined as a format then choose the **One bar for each unique data value** option. Also, you may specify the number of classes, that is the **Enter number of bars** option, or you may specify the actual limits by selecting the **Specify the bar values** option. Selecting this option enables the field next to the **Add** button. You may use this field to enter class limits. This may be done in several ways. For example, for the pulse rate data you may enter the midpoints of the class limits specified in Section 2-2 as **64.5 to 124.5 by 10**. You must then click the **Add** button to complete the entry. The **Specify the bar values** option was used for this illustration.

The **Options** option in the selection pane allows you to customize the appearance of the resulting graph. Check the **2D** checkbox, and then select the **Set spacing** option from the **Bar Size** drop down list box. The **Set spacing** option allows you to set the spacing to 0 between bars as indicated below.

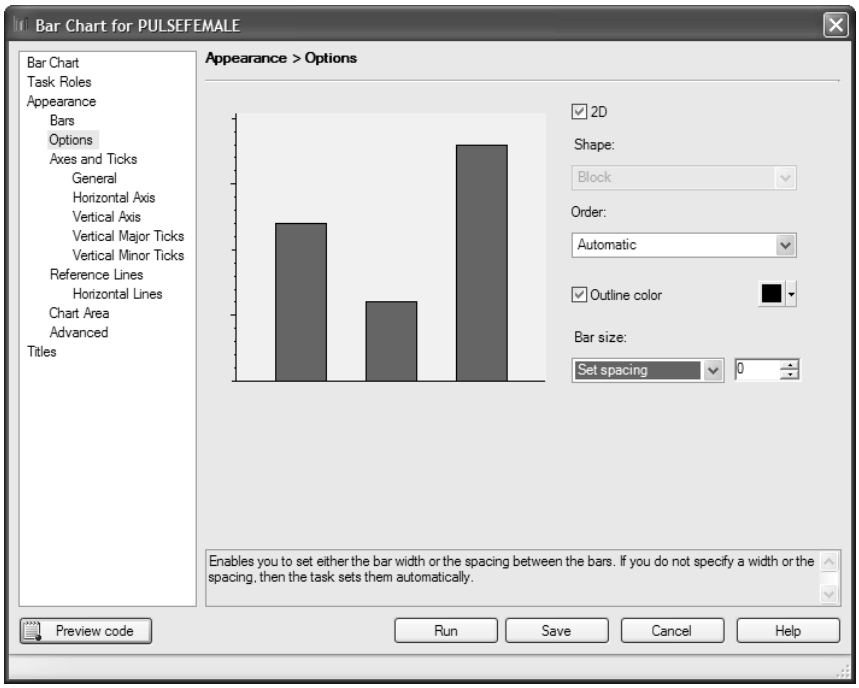

Figure 2-11

The following histogram will appear when you click the **Run** button.

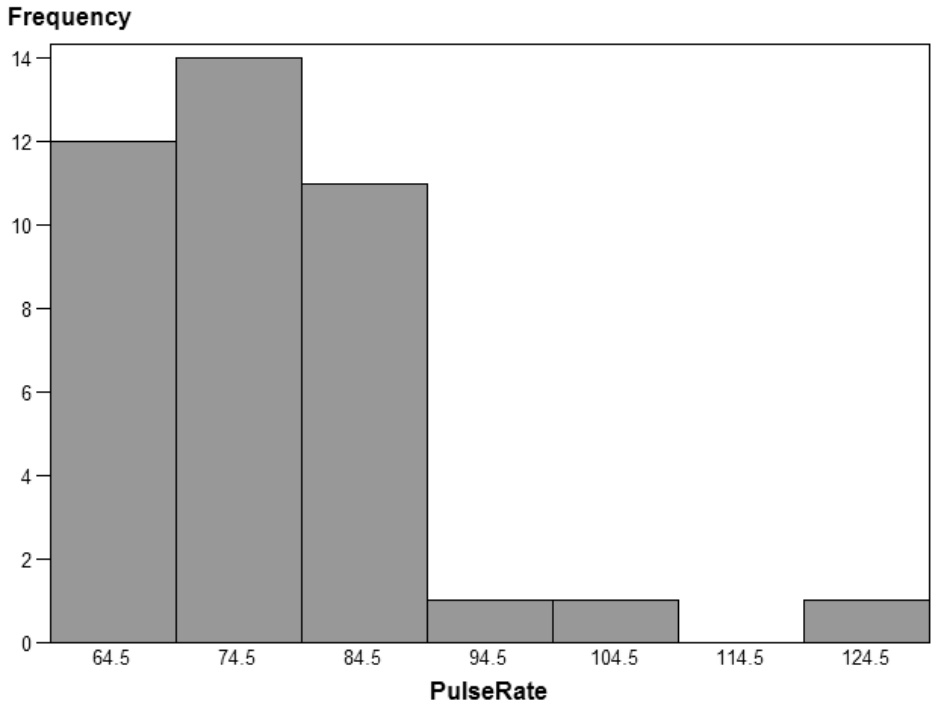

Figure 2-12

The **Bar Chart** task provides several additional options. For example, a **relative frequency histogram** (see *Elementary Statistics*, Figure 2-3) may be created by choosing the **Percentage** item, instead of the default **Frequency** item, from the **Advanced** option (see Figure 2-10).

2-4 Statistical Graphics

In the previous section we saw that a frequency histogram may be used to visualize data. In Section 2-4 of *Elementary Statistics* the author discusses several other graphical tools that may also be used. As we saw in the previous section, SASLE provides many options for producing graphs. Most of these are available from the **Graph** menu, or as a task in the **Graph** category from the **Task List** window. However, as we will see in subsequent sections and chapters, SASLE provides some graphing capabilities from other menus or tasks. In addition, for those cases where SASLE does not directly provide a particular graphing option, it is likely that you may be able to write a very simple SAS program to produce the desired graph. For example, SASLE does not provide direct support for producing frequency polygons, ogives, or dotplots. However, as you will see from Appendix A, SASLE makes it very easy to write and debug a SAS program.

Pareto Charts

A **Pareto chart** is a bar graph, where the bars are arranged from left to right in order of decreasing frequency. Consider the actress data. To use SASLE to construct a **Pareto chart**, select the **Analyze** > **Pareto Chart** menu option or double-click on the **Pareto Chart** task in the **Task List** window.

The following dialog box will appear.

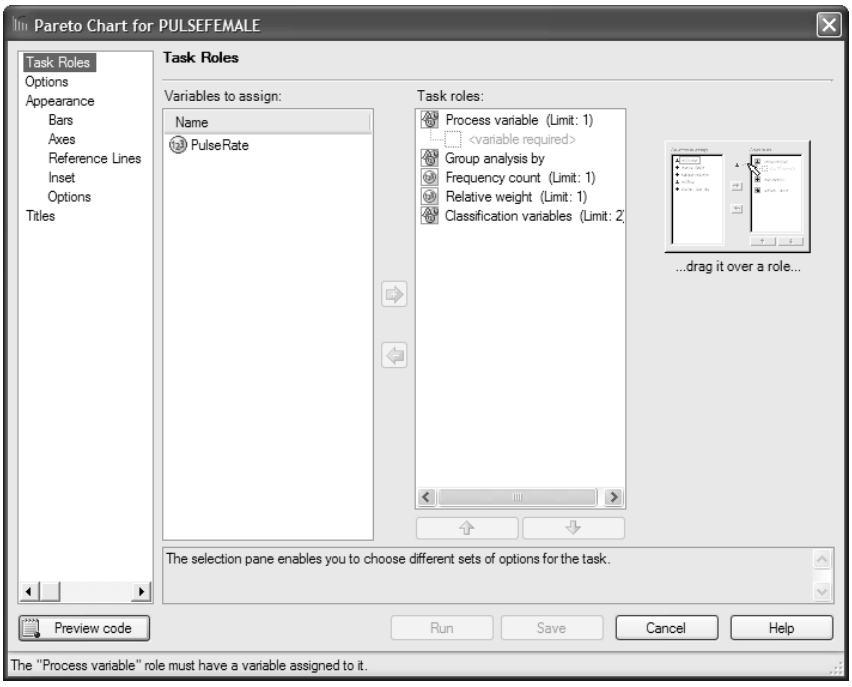

Figure 2-13

Pareto charts are typically used for qualitative data but may be used for quantitative data if the data is organized into discrete classes. The CLASSES format will accomplish this (see Figure 2-6, Figure 2-7, and the accompanying discussion). Assign the **PulseRate** variable to the **Process Variable** role and then click the **Run** button. The Pareto chart below will appear.

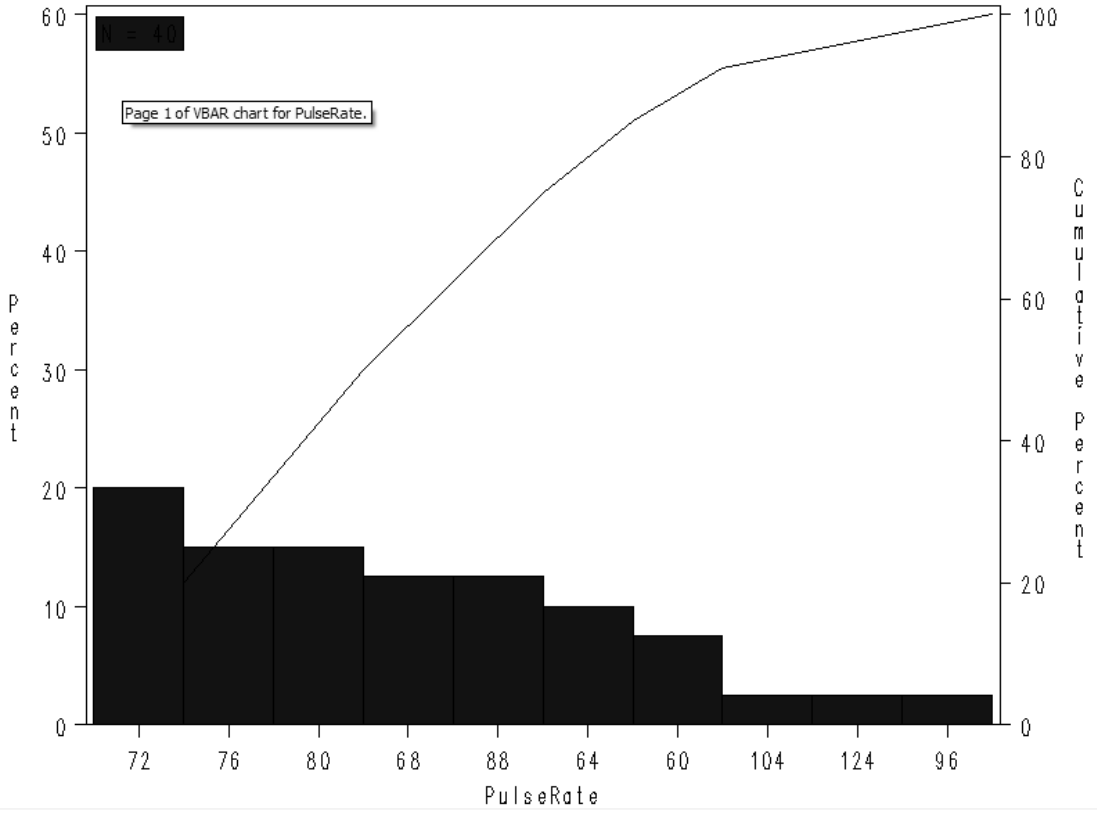

Figure 2-14

Pie Charts

SASLE provides a **Pie** task for producing **pie charts**. This is a very flexible graphing task that allows a variety of pie charts to be created. You may choose the **Graph > Pie Chart** menu option or double-click on the **Pie Chart** task in the **Task List** window. The dialog box that appears is similar in operation to the dialog box for creating **Bar Charts** (see Figure 2-9 above).

Scatterplots

A scatter plot (also known as a scatter diagram) may be used to pictorially summarize paired data. This type of plot is particularly useful if you have reason to believe that the values for each pair are related in some way. For example, consider the health exam data (see dataset 1, Appendix B). This dataset consists of several data values for each male in a U.S. Department of Health and Human Services survey. Waist size, in centimeters, and weight, in pounds, are two of these data values. The (*waist*, *weight*) pair for each of the forty males in the survey is reproduced below.

(90.6, 169.1), (78.1, 144.2), (96.5, 179.3), (87.7, 175.8), (87.1, 152.6), (92.4, 166.8), (78.8, 135), (103.3, 201.5), (89.1, 175.2), (82.5, 139), (86.7, 156.3), (103.3, 186.6), (91.6, 191.1), (75.6, 151.3), (105.5, 209.4), (108.7, 237.1), (104, 176.7), (103, 220.6), (91.3, 166.1), (75.2, 137.4), (87.7, 164.2), (77, 162.4), (85, 151.8), (79.6, 144.1), (103.8, 204.6), (103, 193.8), (97.1, 172.9), (86.9, 161.9), (88, 174.8), (91.5, 169.8), (102.9, 213.3), (93.1, 198), (98.9, 173.3), (107.5, 214.5), (81.6, 137.1), (75.7, 119.5), (95, 189.1), (91.1, 164.7), (94.9, 170.1), (79.9, 151)

Figure 2-15: Waist size (cm) & Weight (lbs)

We may produce a scatter diagram of these paired values by selecting waist to be on the x-axis and weight to be on the y-axis. Each (*waist*, *weight*) pair then becomes a point in the x-y plane.

We will use these paired data values to illustrate how SASLE may be used to create scatter diagrams. As usual, we assume that SASLE has been launched and a project opened. Also, we assume that the health dataset is available and has been added to your project. If necessary, see Section 1-2 to see how a dataset may be created. SASLE provides a **Scatter Plot** task for producing scatter diagrams. This task allows a variety of scatter diagrams to be created. You may choose the **Graph > Scatter Plot** menu option or double-click on the **Scatter Plot** task in the **Task List** window.

The following dialog box will appear.

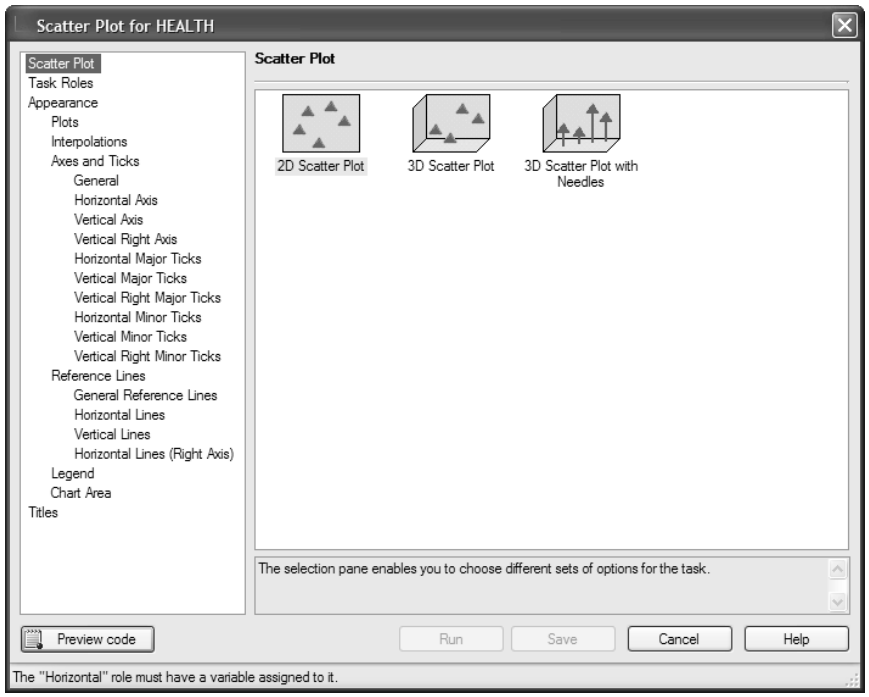

Figure 2-16

Select **2D Scatter Plot** and select the **Task Roles** option from the selection pane. A dialog box similar to Figure 2-13 will appear. Select the **waist** variable from the **Columns to assign** window, click the button, and then choose the **Horizontal** menu option (i.e. x-axis). Now, select the **weight** variable, click the button, and choose the **Vertical** menu option (i.e. y-axis). You could click the **Run** button at this point to obtain the scatter diagram. However, the other options in the selection pane may be used to customize the appearance of the scatter diagram. For example, the **Titles** option may be used to specify a title for the scatter plot and the **Vertical Axis**, plus **Horizontal Axis**, options to define descriptive labels for the axes. Note that SASLE does not provide a default title. Also, the default labels used for the axes are the variable names associated with the data values, that is, weight and waist.

Following is the scatter diagram obtained after clicking the **Run** button.

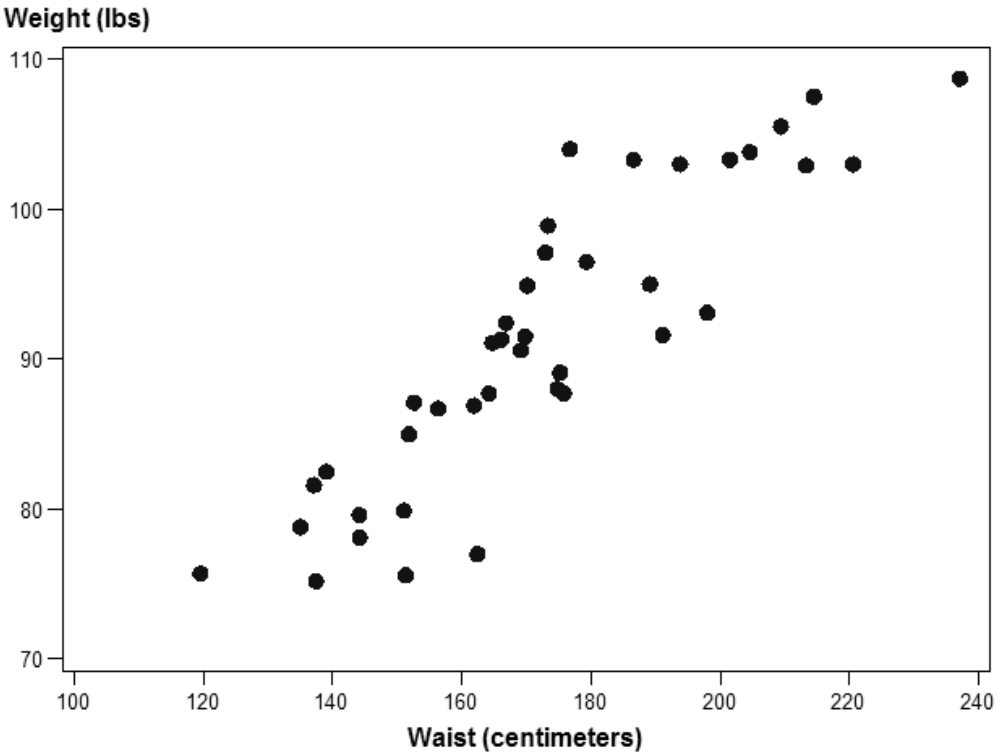

Figure 2-18

Notice the descriptive labels **Weight (lbs)** and **Waist (centimeters)**. As indicated above, the **Vertical Axis**, plus **Horizontal Axis**, options were used to specify these labels. This scatter diagram suggests a relationship between weight and waist. The relationship seems to be a linear relationship with an upward sloping trend. That is, as waist size increases, weight tends to increase also.

Time-Series Graph

Time-series data are data that have been collected over time and are indexed by some time measure. A time-series graph may be used to pictorially summarize such data. To illustrate, consider the following data. Each (*djiahigh*, *year*) pair represents the Dow Jones Industrial Average (DJIA) high value for each year between 1980 and 2000.

(1000, 1980), (1024, 1981), (1071, 1982), (1287, 1983), (1287, 1984), (1553, 1985), (1956, 1986), (2722, 1987), (2184, 1988), (2791, 1989), (3000, 1990), (3169, 1991), (3413, 1992), (3794, 1993), (3978, 1994), (5216, 1995), (6561, 1996), (8259, 1997), (9374, 1998), (11568, 1999), (11401, 2000)

Figure 2-19: DJIA Annual High Values

We will use these paired data values to illustrate how SASLE may be used to create time-series graphs. As usual, we assume that SASLE has been launched and a project opened. Also, we assume that the **djia** dataset is available and has been added to your project. If necessary, see section 1-2 to see how a dataset may be created.

SASLE provides a **Basic Forecasting** task that may be used for producing time-series graphs. This task is intended for time series forecasting but allows for a variety of time-series graphs to be created. You may choose the **Analyze > Time Series > Basic Forecasting** menu option or double-click on the **Basic Forecasting** task in the **Task List** window. The following dialog box will appear.

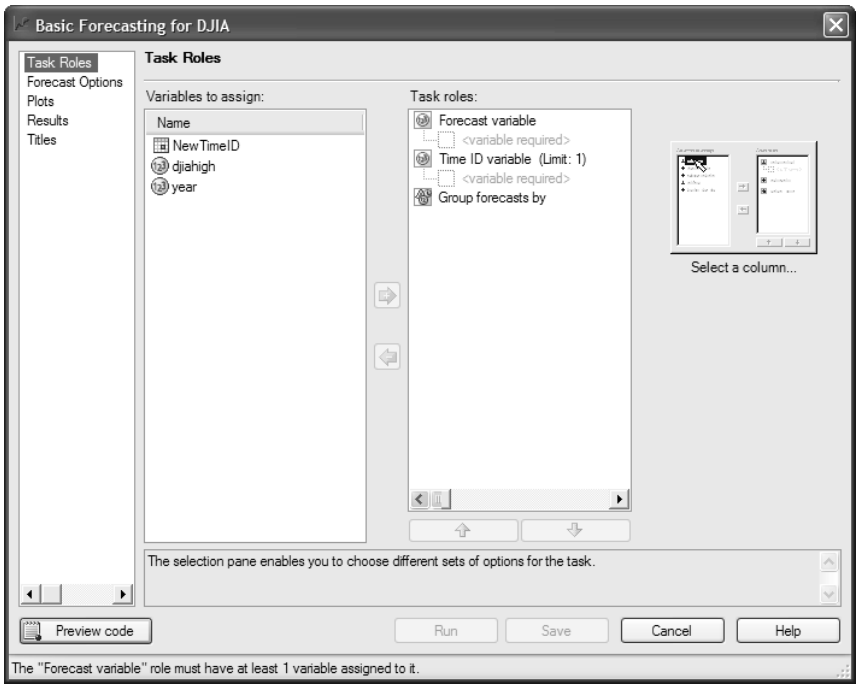

Figure 2-20

Assign the **year** variable to the **Time ID variable** role and the **djiahigh** variable to the **Forecast variable** role. If you select the **Plots** option at this point the following dialog box appears.

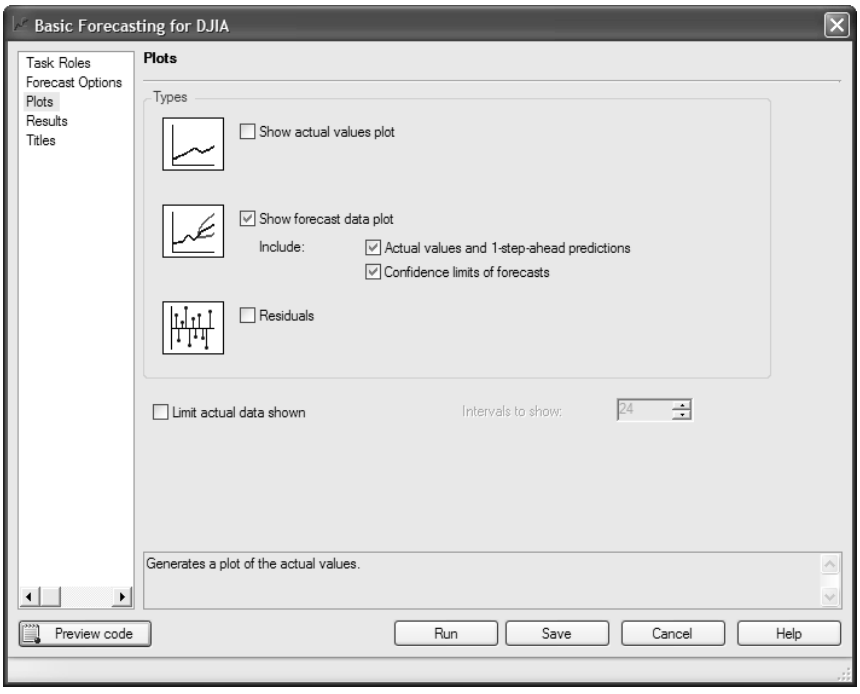

Figure 2-21

Since we want a basic time-series graph, select the **Show actual values plot** checkbox in the **Types** box and deselect the **Show forecast data plot** checkbox. If you click the **Run** button at this point, the following time-series graph will be produced.

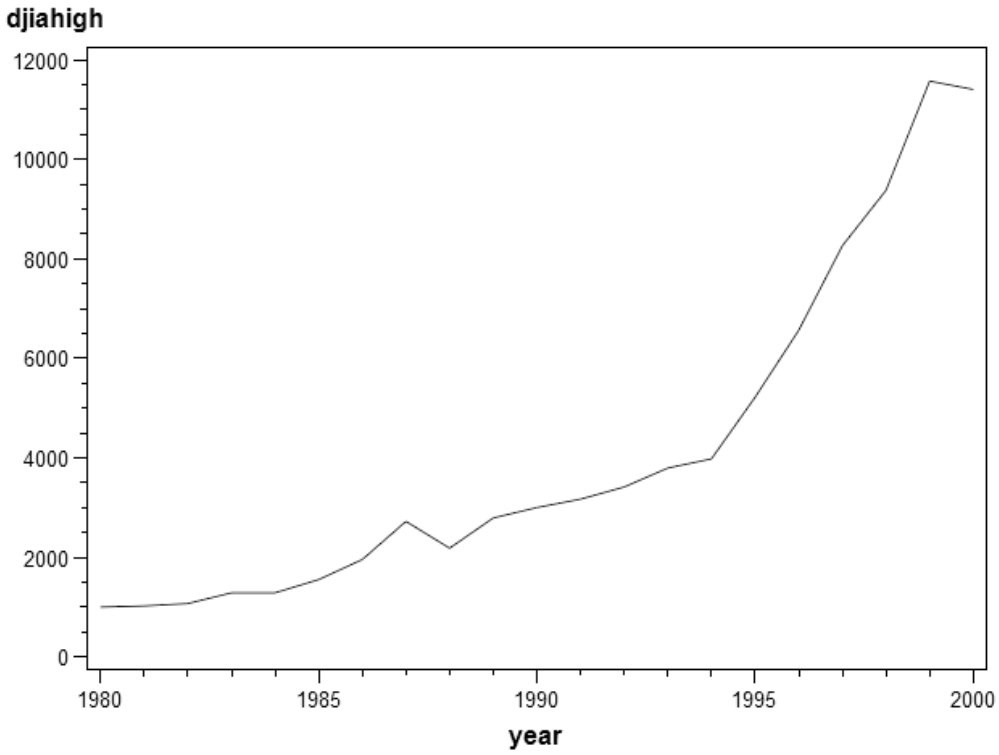

Figure 2-22

Notice that the pattern is a trend of increasing values with dips occurring in 1987 and 2000.

2-5 Critical Thinking: Bad Graphs

SASLE does not provide any support for identifying bad graphs.

Exercises

Refer to Chapter 2 of *Elementary Statistics* for details of the following exercises. Try to use the appropriate SASLE tasks where possible. A detailed description of the data sets mentioned, including the names of the associated data files on the supplied CD-ROM, is available in Appendix B of *Elementary Statistics*. For a few problems, the data values needed are not available as files. Also, it is possible that mentioned files might be missing. In such cases, review the appropriate sections from Chapter 1 of this manual on creating datasets.

1. Work problems 13 to 25 from Section 2-2. Read the problems carefully. Some problems require frequency distributions and others require relative frequency distributions. Use SASLE to construct the required distribution in each case.

2. Work problems 9 to 17 from Section 2-3. Use SASLE to construct the required histogram in each case.

3. Work problems 6 and 9 from Section 2-4.
   Hint: You may use the Distribution Analysis task to construct stem-and-leaf plots.

4. Work problems 13 to 24 from Section 2-4. Use SASLE to construct the required charts in each case. Datasets are available for problems 21 and 22.

5. Work problems 1, 2, and 4 to 6 from the Review Exercises section. Use SASLE to construct the plots in each case.
   Hint: You may use the Distribution Analysis task to construct stem-and-leaf plots.