

Chapter 2

This Chapter addresses how to use SAS Learning Edition (SASLE) to derive tables, graphs, plots, and summary statistics. You should be familiar with Chapter 2 of *Elementary Statistics* prior to beginning this Chapter.

2-1 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 serum cotinine problem presented in *Elementary Statistics* (see Table 2-1). The cotinine level for each of the 40 smokers in the smoker group is reproduced below.

1, 0, 131, 173, 265, 210, 44, 277, 32, 3, 35, 112, 477, 289, 227, 103, 222, 149, 313, 491, 130, 234, 164, 198, 17, 253, 87, 121, 266, 290, 123, 167, 250, 245, 48, 86, 284, 1, 208, 173

Figure 2-1: Smokers - Cotinine Level

These data values are used in Section 2-2 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 the class limits.

The SASLE two-step process below assumes that you have launched SASLE and have opened a project (see Chapter 0). It also assumes that the cotinine data is available as a SAS dataset. If the cotinine data is unavailable then it may be easily created (see Chapter 0 for details).

Step 1: Choose the **Tools > 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.

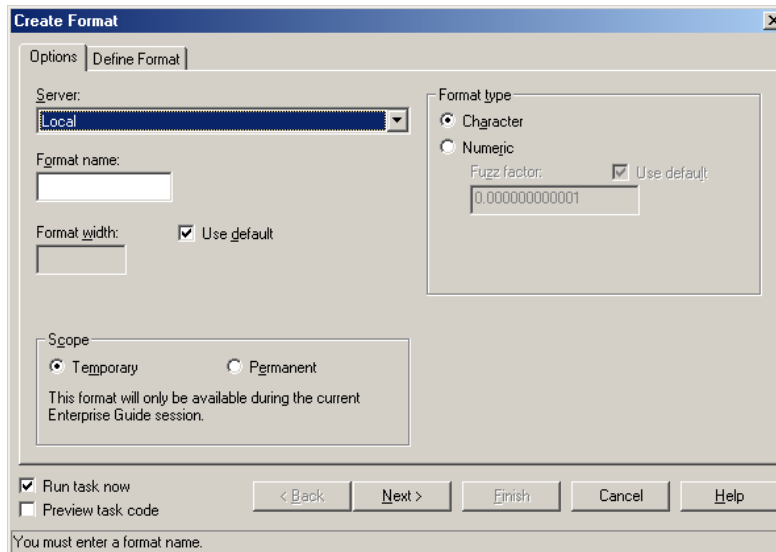


Figure 2-2

Enter a **Format name** (e.g. CLASSES) and then select the **Numeric** Format type. Click the **Next** button. The following dialog box, which allows you to define class limits, will then appear:

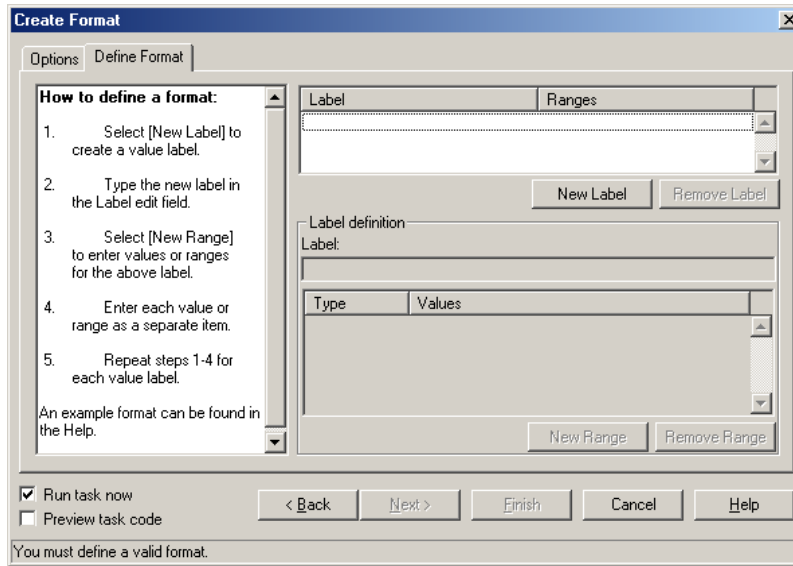


Figure 2-3

Notice the **How to define a format** window on the left. To create the class limits you must follow the five-step procedure outlined in this window for each class limit. For example, to create the 0-99 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 0-99 class limit. You may use any name but, for the sake of convenience, it is probably a good idea to use the text “0-99” 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 0 and 99 inclusive. The dialog box below shows two class limits already defined.

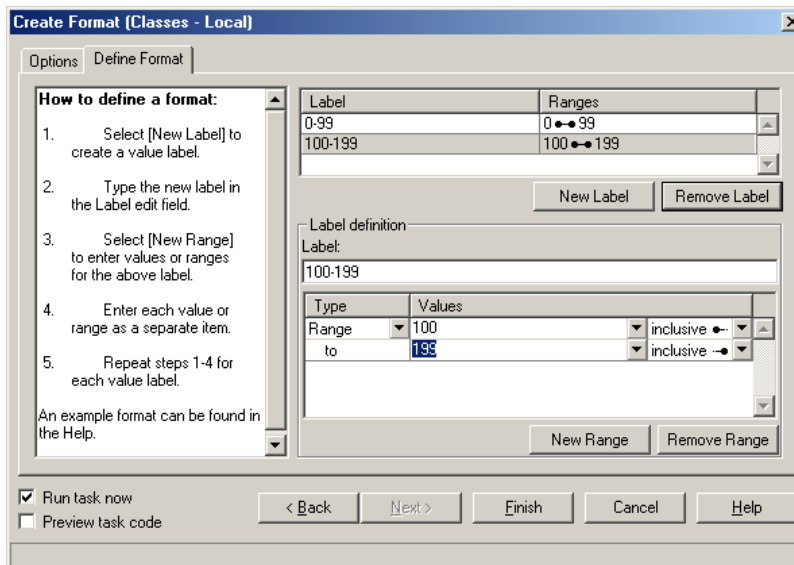


Figure 2-4

Step 2: Choose the **Analysis > Descriptive > 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.

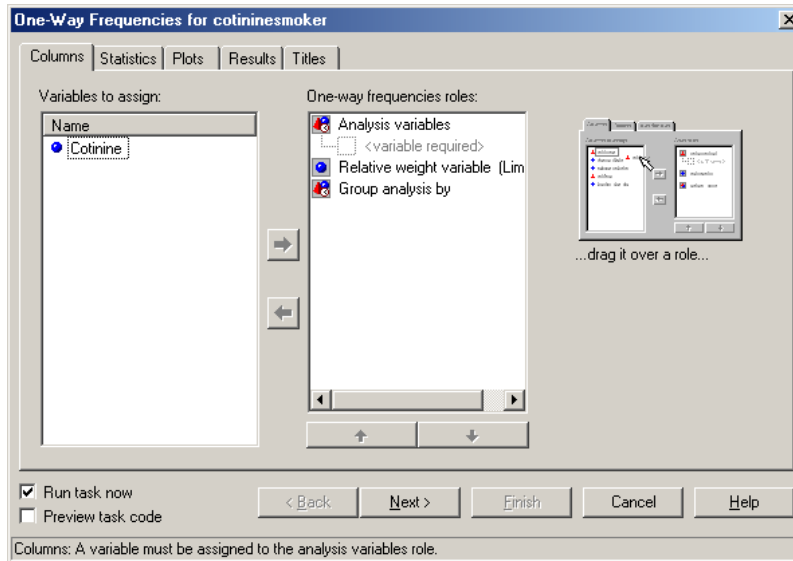



Figure 2-5

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

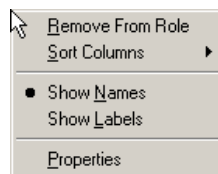



Figure 2-6

Choose **Properties**, then click on the  button, adjacent to the **Format** field, in the Properties dialog box. The following **Formats** dialog box will then appear. Select the **User Defined** option in the **Categories** window as shown below.

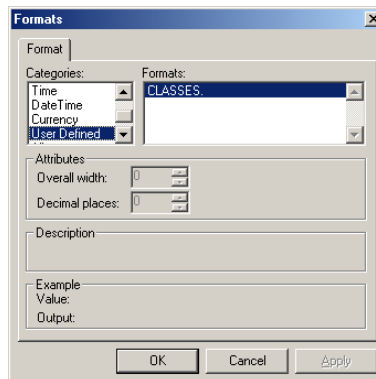


Figure 2-7

The CLASSES format that appears in the **Formats** window is the format that was defined in Step 1 above. Select it and click the **OK** button. Continue clicking **OK** buttons until the **One-Way Frequencies** dialog box reappears. The following frequency distribution table will appear when you click the **Finish** button.

The screenshot shows the SAS Enterprise Guide interface. At the top, it says "Enterprise GUIDE" and "sas. The Power to Know.". Below that, it says "One-Way Frequencies Results" and "The FREQ Procedure". The main content is a table with the following data:

Cotinine	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0-99	11	27.50	11	27.50
100-199	12	30.00	23	57.50
200-299	14	35.00	37	92.50
300-399	1	2.50	38	95.00
400-499	2	5.00	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** tab of the **One-Way Frequencies** dialog box (see Figure 2-5 above) before clicking on **Finish**, 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 2-7 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 would 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.

2-2 Visualizing Data

In the previous section we saw that a frequency distribution table may be used to summarize data. There are several pictorial tools that may also be used. These include bar charts/histograms, pie charts, box plots, stem-and-leaf plots, and others. Since these are pictorial tools they provide a way to visualize the data and so, when used appropriately, they can be very effective tools for summarizing data.

Histograms

In *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** 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** 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 2-1) then the format may be used in the **Bar** 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 cotinine data to illustrate and assume that SASLE has been launched and a project opened. Choose the **Graph > Bar** menu option. Alternatively, you may double-click on the **Bar** task in the **Task List** window. The following dialog box will appear.

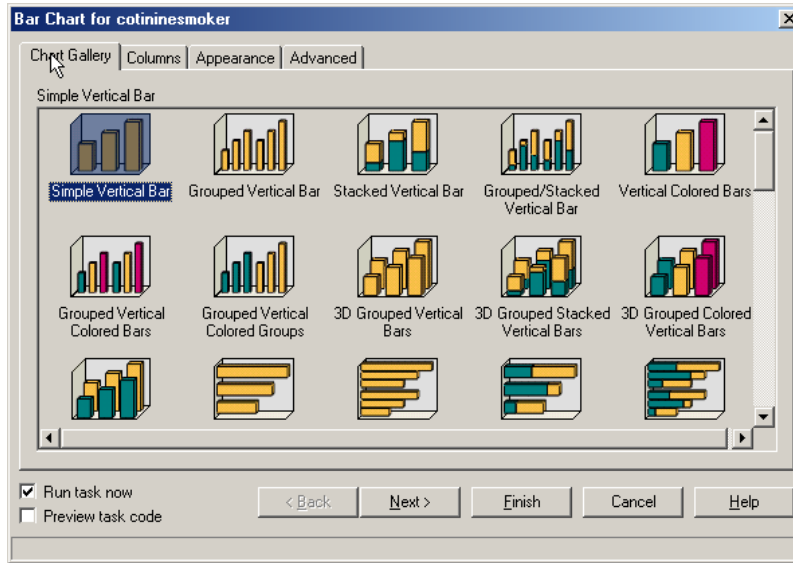


Figure 2-9

Select **Simple Vertical Bar** and click on the **Next** button. A dialog box similar to Figure 2-5 will appear. Select the Cotinine variable from the **Columns to assign** window, click on the **+** button, and then choose the **Column to chart** menu option. The Cotinine variable now appears in both the **Columns to assign** and **Simple 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 now click on the **Next** button, the following dialog box appears.

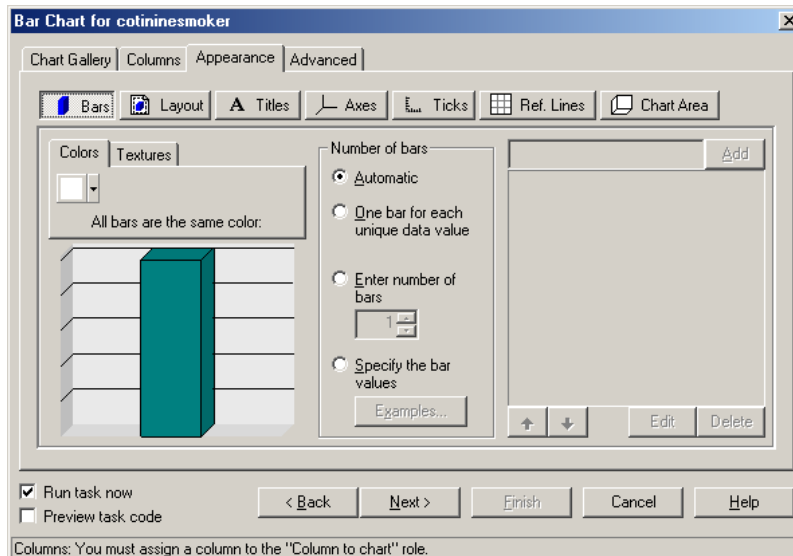


Figure 2-10

Notice that you may specify **Number of bars**. This is where you choose how to specify class limits. The default choice is **Automatic**. 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 **Add** button, which allows you to enter class limits. This may be done in several ways. For example, for the cotinine data, you may enter the midpoints of the class limits specified in section 2-1 as **50 to 450 by 100**. You must then click the **Add** button to complete the entry. The **Specify the bar values** option was used for this illustration.

The button bar shown in Figure 2-10 and 2-11 allows you to customize the appearance of the resulting graph. For example, SASLE does not provide a default title and so the **Titles** button could be used to specify one. We will not specify a title for this illustration. However, to ensure the proper layout of the histogram you will need to click on the **Layout** button. The resulting dialog box appears below. Check the **2D** checkbox, and then select the **Set spacing** option. The **Set spacing** option allows you to use the slider bar to specify 0% spacing between bars as indicated below.

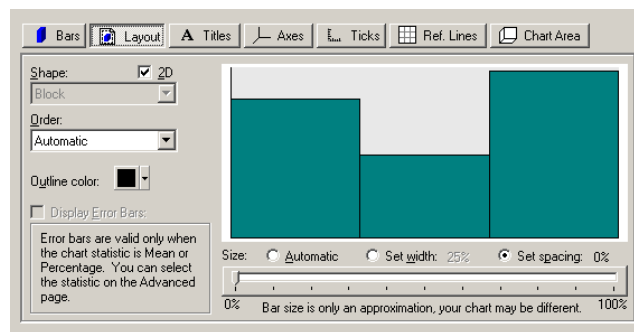


Figure 2-11

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

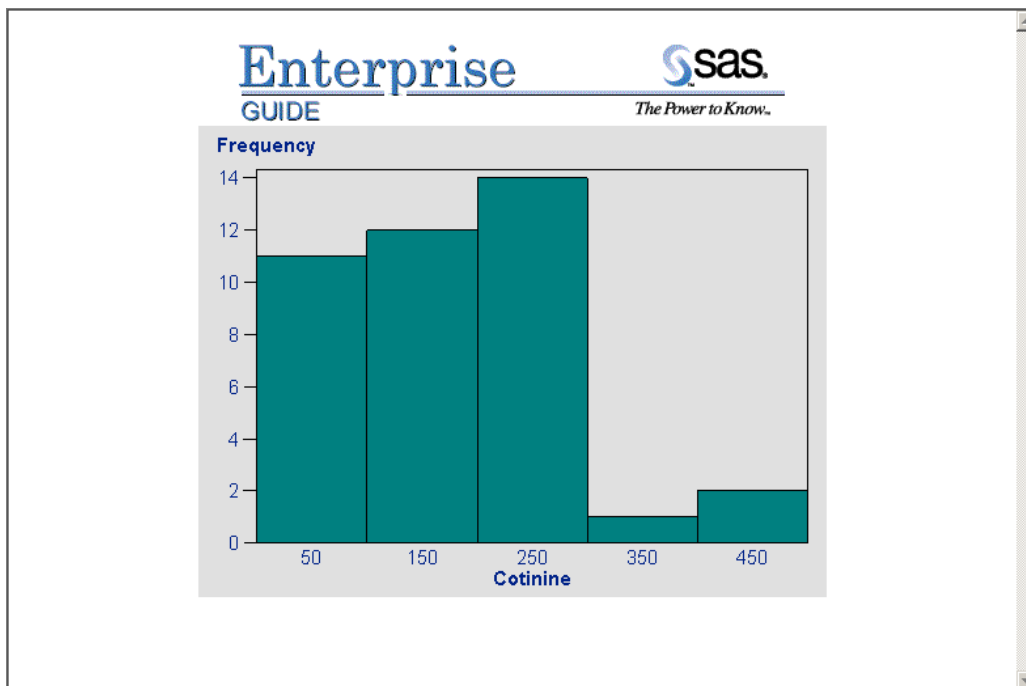


Figure 2-12

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

Stem-and-Leaf Plots

Unfortunately, SASLE does not provide a task that produces stem-and-leaf plots. However, it is possible to construct such a plot by using the **Create Code** task to write a very simple SAS program. An example of such a program is presented in Appendix A. Appendix A also provides an introduction to programming using the SAS programming language.

Pareto Charts

A **Pareto chart** is a bar graph, where the bars are arranged from left to right in order of decreasing frequency. You may use SASLE to construct a **Pareto chart** by choosing the **Analysis > Pareto Chart** menu option or you may double-click on the **Pareto Chart** task in the **Task List** window. The following dialog box will appear.

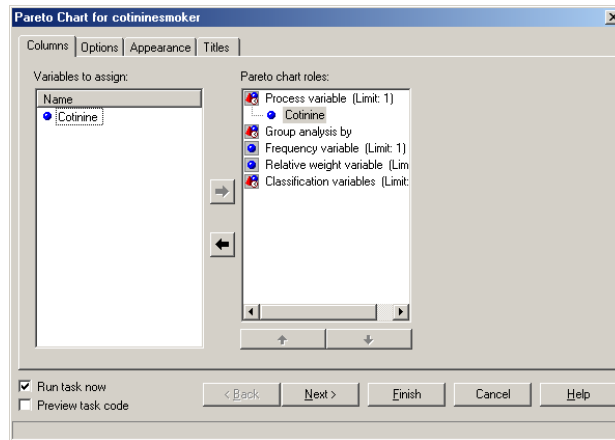


Figure 2-13

Notice that the cotinine data is being used. Pareto charts are typically used for qualitative data but may be used for quantitative data if the data is organized into discrete classes. In this case, the **CLASSES** format accomplishes this and so, by specifying the **CLASSES** format (see Figure 2-6, Figure 2-7, and the corresponding discussion above), we may create a Pareto chart. The Pareto chart below, will appear if the **Finish** button is clicked at this point.

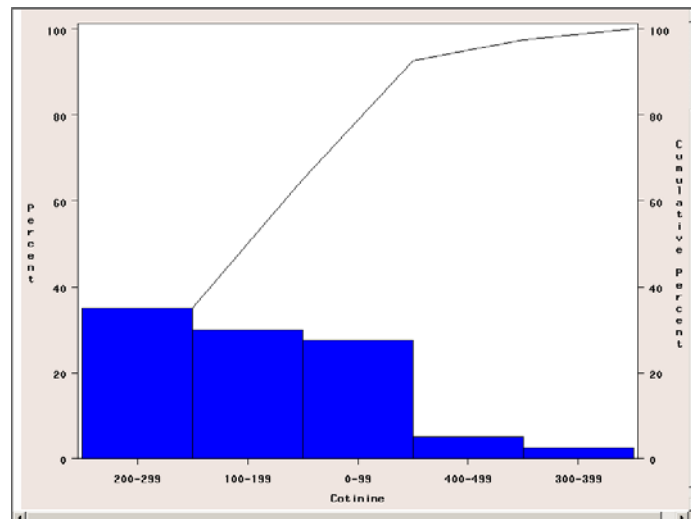


Figure 2-14

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** menu option or double-click on the **Pie** 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).

Scatter Diagrams

A **scatter diagram** (also known as a **scatter plot**) 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 discussed in *Elementary Statistics* (see Data Set 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 (see Chapter 0).

SASLE provides a **Scatter** task for producing **scatter diagrams**. This task allows a variety of scatter diagrams to be created. You may choose the **Graph > Scatter** menu option or double-click on the **Scatter** task in the **Task List** window. The following dialog box will appear.

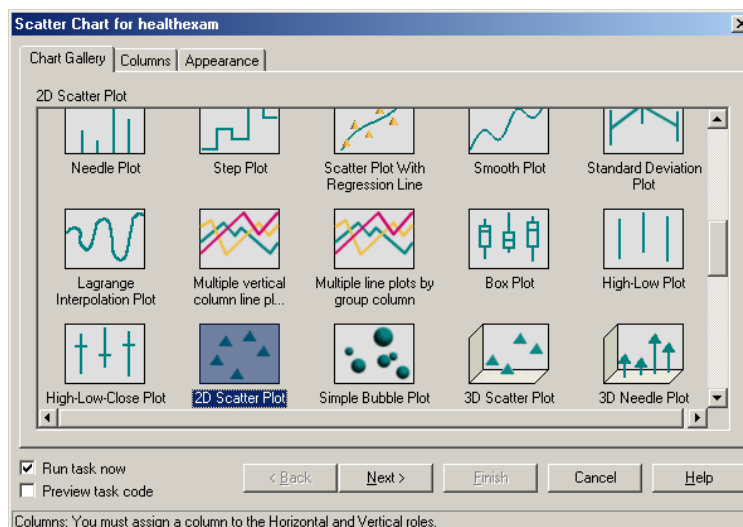


Figure 2-16

Select **2D Scatter Plot** and click on the **Next** button. A dialog box similar to Figure 2-5 will appear. Select the waist variable from the **Columns to assign** window, click on the **→** button, and then choose the **Horizontal** menu option (i.e. x-axis). Now, select the weight variable, click on the **→** button, and choose the **Vertical** menu option (i.e. y-axis). You could click on the **Finish** button at this point to obtain the scatter diagram. However, clicking on the **Next** button results in a dialog box with the following button bar. These buttons may be used to customize the appearance of the scatter diagram.



Figure 2-17

You may use the **Titles** and **Axes** buttons to specify a title for the scatter plot and 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 **Finish** button.

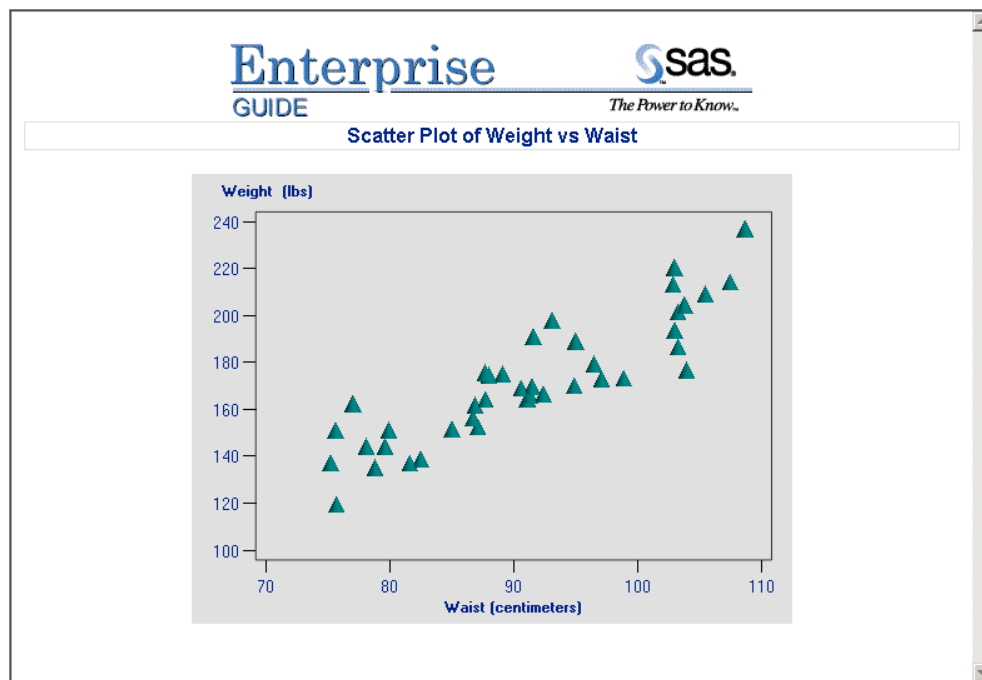


Figure 2-18

Notice that a title, **Scatter Plot of Weight vs Waist**, has been specified, as well as descriptive labels, **Weight (lbs)** and **Waist (centimeters)**.

As pointed out in *Elementary Statistics*, 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.

2-3 Descriptive Statistics

In the previous sections we saw how we may use tables and pictorial tools to summarize data values. We may also compute a variety of *measures* that may be used to summarize and describe a set of data values. Such measures are known as descriptive statistics and are discussed in Section 2-4 through Section 2-6 of *Elementary Statistics*. Section 2-4 discusses measures that define the center of a set of values. That is, measures such as the **arithmetic mean**, **median**, **mode**, and others. Section 2-5 discusses measures that define the variability of a set of values. That is,

measures such as the **range**, **standard deviation**, **variance**, and others. Section 2-6 discusses measures of relative standing for a set of values. That is, measures such as **quartiles**, **percentiles**, and others.

SASLE provides several tasks that may be used to compute descriptive statistics. The **Summary Statistics** task and the **Distribution Analysis** task are two such tasks. We will use the **Distribution Analysis** task to illustrate, since it provides more options than the Summary Statistics task. The cotinine dataset will be used. As usual, we assume that SASLE has been launched and a project opened (see Chapter 0).

To invoke the **Distribution Analysis** task, choose the **Analysis > Descriptive > Distribution Analysis** menu option or double-click on the **Distribution Analysis** task in the **Task List** window. The following dialog box will appear.

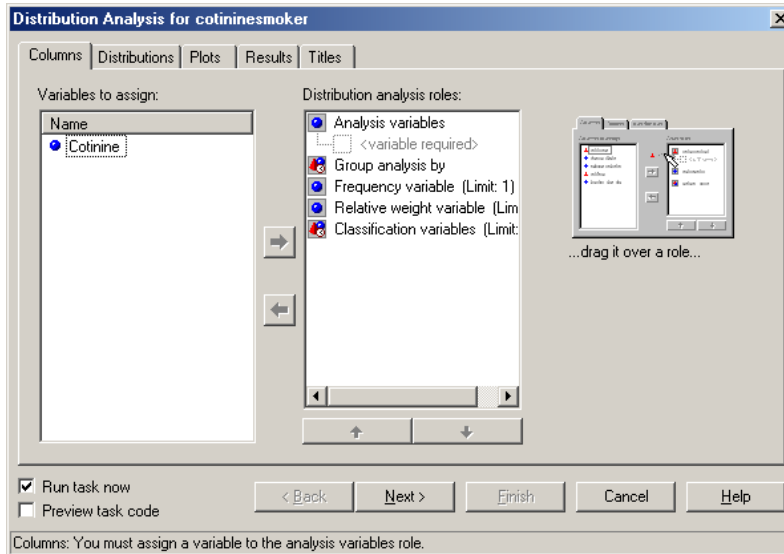


Figure 2-19

Select the Cotinine variable from the **Columns to assign** window, click on the **►** button, and then choose the **Analysis Variables** menu option. Now, click on the **Results** tab. This tab allows you to customize the report that will be produced. See the **Tables to include in report** window below for the default table selection.

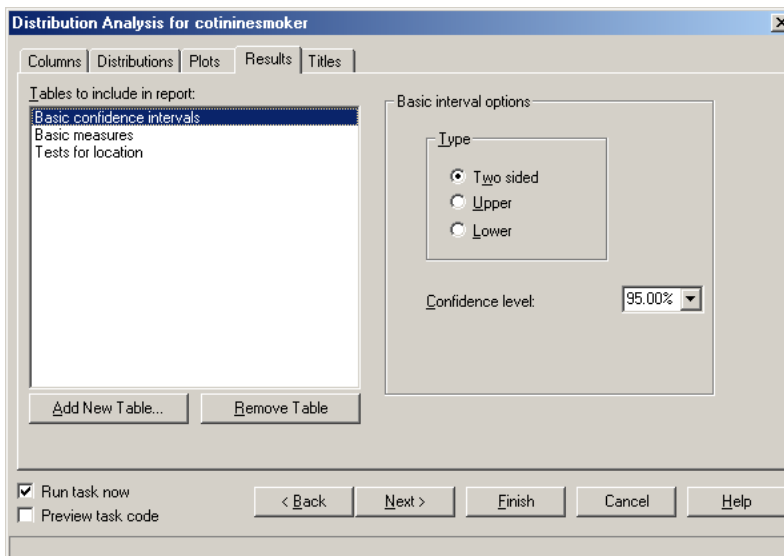


Figure 2-20

We are only interested in the **Basic Measures** table at this point. It contains the mean, mode, standard deviation, variance, range, and interquartile range, but does not include quartiles, and percentiles. Also, this basic table only reports one mode, and so it is not sufficient in those cases where there are multiple modes. We may use the **Remove Table** button to remove the **Test for location** and the **Basic confidence intervals** tables and the **Add Table** button to include the needed tables. If you click on the **Add Table** button, you will obtain the following dialog box. Notice that the **Tables** window contains a list of available tables that may be included in the report and the **Tables to include** window contains the tables selected. In this case, the **Modes** and **Quantiles** tables have been added and the **Test for location** and **Basic confidence intervals** tables removed.

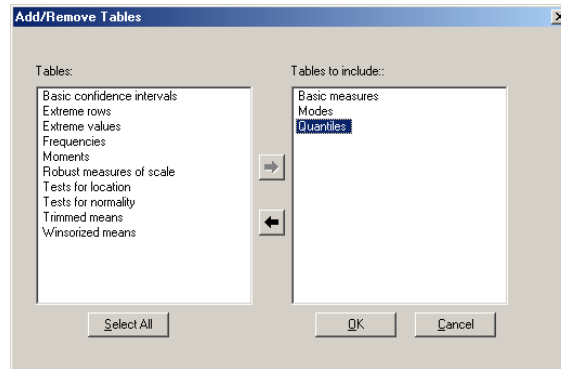


Figure 2-21

If you click on **OK** and then **Finish** the following reports (presented as Figure 2-22 – Figure 2-24) will be obtained.

The UNIVARIATE Procedure
Variable: Cotinine

Basic Statistical Measures			
Location		Variability	
Mean	172.4750	Std Deviation	119.49831
Median	170.0000	Variance	14280
Mode	1.0000	Range	491.00000
		Interquartile Range	165.00000

NOTE: The mode displayed is the smallest of 2 modes with a count of 2.

Figure 2-22

Figure 2-22 is the **Basic Measures** table mentioned above.

Modes	
Mode	Count
1	2
173	2

Figure 2-23

Figure 2-23 is the **Modes** table mentioned above. Notice that, in this case, we have two modes.

Quantiles (Definition 5)	
Quantile	Estimate
100% Max	491.0
99%	491.0
95%	395.0
90%	289.5
75% Q3	251.5
50% Median	170.0
25% Q1	86.5
10%	10.0
5%	1.0
1%	0.0
0% Min	0.0

Figure 2-24

Figure 2-24 is the **Quantiles** table, and so contains the median, quartiles, and percentiles.

The **Distribution Analysis** task provides several additional options. For example, measures such as **coefficient of variation** are available from the **Moments** table. Also, the **Distribution Analysis** task is one of several SASLE tasks that may be used to create **boxplots** (see Section 2-7 of *Elementary Statistics* for a discussion of boxplots). In addition to boxplots, the **Plots** tab (see Figure 2-17) also allows the creation of histograms, probability plots, and quantile-quantile plots.

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 SAS 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 SAS data files. Also, it is possible that mentioned SAS data files may be missing. In such cases, review the appropriate sections from Chapter 0 of this manual on creating SAS data files.

1. Work problems 14 to 20 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 7 to 10 from Section 2-3. Read the problems carefully. Some problems require frequency histograms and others require relative frequency histograms. Use SASLE to construct the required histogram in each case. Note that SAS data files are only available for problems 9 and 10.
3. Work problem 19 from Section 2-3. Remember that SASLE does not provide a task to construct stem-and-leaf plots. You will need to refer to Appendix A of this manual for details on writing a SAS program.
4. Work problems 21 to 24 from Section 2-3. Use SASLE to construct the required charts in each case.
5. Work problems 25 and 26 from Section 2-3. Use SASLE to construct the scatter diagrams in each case.
6. Work problems 13 to 16 from Section 2-4. Use SASLE to compute the required statistics in each case. Note that these problems may be worked in conjunction with problem 8 below.

7. Work problems 1 to 8 from Section 2-5. Use SASLE to compute the required statistics in each case. Note that SAS data files are only available for problems 1 to 4.
8. Work problems 13 to 16 from Section 2-5. Use SASLE to compute the standard deviation in each case. Note that these problems may be worked in conjunction with problem 6 above.
9. Work problem 41 from Section 2-6. Use SASLE to generate a Quantile report and then answer parts a – d.
10. Work problems 1 to 12 from Section 2-7. Use SASLE to construct the box-plots in each case. Note that you do not need to find the 5-number summary as an initial step. SASLE will automatically determine the 5-number summary and construct the box-plot from the supplied data values. Furthermore, SASLE will construct Modified box-plots. Also, note that there is no SAS data file for problem 9.