

Chapter 2. Describing, Exploring, and Comparing Data

There are many tools used in Statistics to visualize, summarize, and describe data. This chapter will focus on using SPSS to create the tables, graphs, and charts, and calculate the descriptive statistics that are discussed in *Elementary Statistics*. See *Elementary Statistics* for a thorough discussion of the uses and misuses of these techniques. You should be familiar with Chapter 2 of *Elementary Statistics* prior to beginning this chapter.

Section 2-1. Frequency Distributions

For a first look at data, it is practical to begin by describing the distribution of values in a data file. Many statistical tools have been developed for this purpose. SPSS can simplify the creation of the tables, graphs, and charts used for exploring a data file. A common tool used for describing the distribution of values is the **frequency distribution**.

SPSS makes a **frequencies report**. A frequencies report is different from a frequency distribution. The frequencies report simply lists distinct data values with their frequencies; it does not list the frequency of data values in a list of categories. SPSS does not provide a way to customize the class limits to enable a frequency table to be created. It is not difficult to make a frequency distribution by hand from the frequencies report.

Frequencies Report for a Variable

The **Health Exam Results** data from Data Set 1 in Appendix B of *Elementary Statistics* (this data appears on disk as **Mhealth.sav**) has 13 different measurements on 40 men. The variables being measured are age, height, weight, waist circumference, pulse rate, systolic and diastolic blood pressure, cholesterol level, body mass index, upper leg length, elbow breadth, wrist breadth, and arm circumference. These data are from the U.S. Department of Health and Human Services, National Center for Health Statistics, Third National Health and Nutrition Survey.

Open the **Mhealth.sav** data file (see Section 0-3). To obtain a frequencies report for **age** and **ht**, choose **Analyze > Descriptive Statistics > Frequencies...** to open the Frequencies dialog box (Figure 2 - 1).

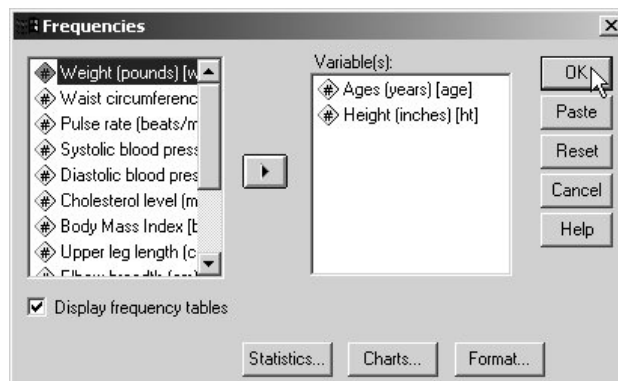



Figure 2 - 1

Highlight the variable, *age*, by clicking on its label “**Ages (years) [age]**” and then clicking the **Variable Paste**  button to copy the variable to the Variable(s) list. In a similar manner, copy the variable, *ht*, to the Variable(s) list.

If the variable has many distinct data values the resulting frequencies report can be very long. The frequencies report(s) can be suppressed by unchecking the checkbox for **Display frequency tables** in the Frequencies dialog box. If you uncheck this box, SPSS may give a warning that says, “You have turned off all output. Unless you request Display Frequency Tables, Statistics, or Charts, FREQUENCIES will generate no output.” Frequency reports are often the first tool used to describe a data file. SPSS provides buttons on the Frequencies dialog box for calculating some common statistics and charts that are useful for describing data files. In the following sections we will discuss many of these charts and statistics. The checkbox for **Display frequency tables** should be checked now since the goal here is to produce the frequencies table.

Click the **OK** button and SPSS will produce the frequencies report in a new window called the **Output Viewer Window**. The frequencies reports for both *age* and *ht* are rather long and so only the beginning and end of the frequencies report for *age* is shown below (Figure 2 - 2). The Output Viewer window displays the entire frequencies report for both variables.

Ages (years)

	Frequency	Percent	Valid Percent	Cumulative Percent
Valid 17	2	5.0	5.0	5.0
18	2	5.0	5.0	10.0
20	4	10.0	10.0	20.0
22	1	2.5	2.5	22.5
25	1	2.5	2.5	25.0
54	1	2.5	2.5	90.0
55	1	2.5	2.5	92.5
56	1	2.5	2.5	95.0
58	1	2.5	2.5	97.5
73	1	2.5	2.5	100.0
Total	40	100.0	100.0	

Figure 2 - 2

The frequencies report shows the frequencies associated with each distinct data value. Be careful when reading a frequencies report, for example, a common oversight is to not notice that there are no data values equal to 19, 21, 23-24, and so on in this data file. The frequencies report indicates that the minimum and maximum data values are 17 and 73, respectively. The report shows that the data value 17 occurs two times, and that the data value 20 occurs four times. The last line of the frequencies report indicates that there are 40 data values in the data file. A frequency distribution can easily be obtained from the frequencies report.

The column labeled, Percent, indicates the relative frequency with which each data value occurs. This column can be used to make a relative frequency distribution of the ages. The last column labeled, Cumulative Percent, gives the cumulative relative frequencies. This column can be used to make a cumulative frequency distribution for ages.

Output Viewer Window

This section provides an overview of the Output Viewer window. The Output Viewer window shows the output that results from a command procedure being run in the dialog window. If there is no Output Viewer window open, SPSS will open a new Output Viewer window. The Output

Viewer window can be used to hide, recenter, or delete the results of procedures. The frequencies procedure results are shown in the Output Viewer window (Figure 2 - 3).

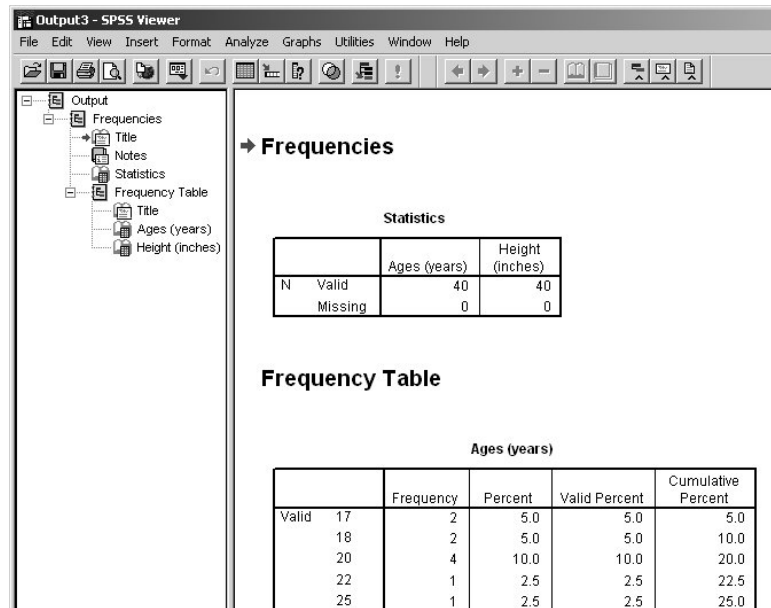





Figure 2 - 3

The SPSS Output Viewer window is divided into two independent windows. The left hand window (**Outline window**) shows the output in outline view. The right hand window (**Output window**) shows the output.

The Outline window is useful for choosing output to be displayed in the Output window. Individual items in the Outline window can be displayed or hidden. For example, **Notes**, which has information about the execution of a procedure, is hidden by default. Click on **Notes** (or the icon  in front of it) to select the item **Notes** and choose **View > Show** to cause the Notes to be displayed in the Output window. Choose **View > Hide** to hide a selected item in the Output window.

Clicking the minus sign,  in front of an item will hide all the output associated with the item. This is the same as selecting the item and choosing **View > Collapse**. Click on the **minus sign** in front of **Frequencies**; notice that the output window is now blank. Click on **Frequencies** to select it, and then choose **View > Expand** to make the output reappear. Alternately, clicking on the **plus sign**  in front of a collapsed item will also expand the output.

Clicking on (selecting) an item in the Outline window will cause the Output window to be re-centered on the item. For example, click on **Ages (years)** in the Outline window to display the frequencies report for the variable *age* in the Output window. Click on **Height (inches)** in the Outline window and the display in the Output window changes to the frequencies report for the variable *ht*.

If you want to delete an item from the Output window, then click on the item and choose **Edit > Delete**. For example, to delete the frequencies report for height, click on **Height (inches)** in the Outline window and choose **Edit > Delete** (or simply press the delete key). If you mistakenly delete some output you may choose **Edit > Undo** and SPSS will undo the previous action.

Frequencies Reports for subgroups of a Variable

Often the same variable is measured for several groups. In situations like this the data file will contain two variables. One variable will contain the values being measured and a second variable that indicates into which group the particular case belongs. An example of such a situation is the **Passive and Active Smoke** data set.

Consider the data from the Chapter Problem *Are nonsmoking people really affected by others who are smoking cigarettes, or is the effect of secondhand smoke really a myth?* in Chapter 2 of *Elementary Statistics*. Data Set 6 in Appendix B of *Elementary Statistics* includes some of the latest data from the National Institutes of Health (this data also appears on disk as **Cotinine.sav**). The data were obtained as part of the National Health and Nutrition Examination Survey. The data values consist of measured levels of serum cotinine (in ng/ml) in people selected as study subjects. (The data were rounded to the nearest whole number, so a value of zero does not necessarily indicate the total absence of cotinine. In fact, all of the original values were greater than zero.) Cotinine is a metabolite of nicotine, meaning that when the body absorbs nicotine, cotinine is produced. Because it is known that nicotine is absorbed through cigarette smoking, we have a way of measuring the effective presence of cigarette smoke indirectly through cotinine.

Open the **Cotinine** data (see Section 0-3). The data file contains two variables, *cotinine* and *group*. The variable, *cotinine*, is the measured cotinine levels for the 120 people. There are 40 people in each of three groups. The variable, *group*, indicates whether the person was a smoker-SMOKER, a nonsmoker who was exposed to secondhand smoke- ETS, or a nonsmoker who was not exposed to secondhand smoke- NOETS).

To obtain the frequencies reports of the cotinine levels for all three groups (Smoker, ETS, and NOETS), choose **Analyze > Descriptive Statistics > Crosstabs...** to open the Crosstabs dialog box (Figure 2 - 4). The Crosstabs procedure forms two-way (and even three-way) tables and like the Frequencies procedure creates a frequencies report and calculates a variety of statistics and graphical displays.

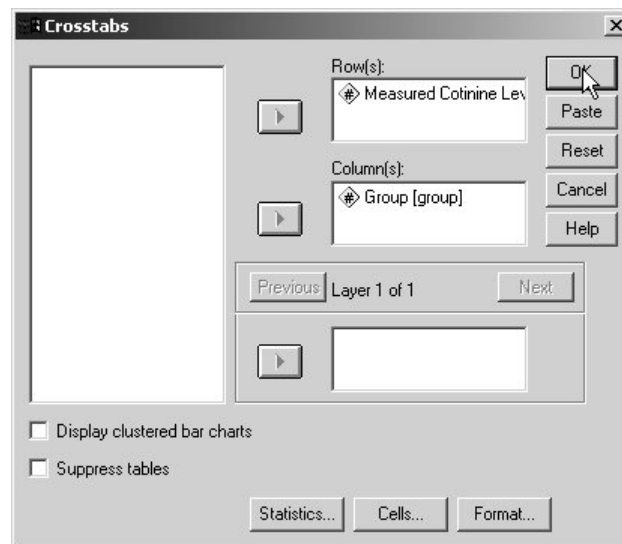


Figure 2 - 4

Copy the variable, *cotinine*, to the Row(s) list and the variable, *group*, to the Column(s) list by selecting the variable and clicking the **Variable paste** button. Click the **OK** button and the

frequencies report will appear in the Output Viewer window. The frequencies reports are very long; only the beginning and end of the frequencies reports are shown below (Figure 2 - 5).

	Count	Group			Total
		Smoker	ETS (Environmental Tobacco Smoke)	NOETS (No Environmental Tobacco Smoke)	
Measured Cotinine Levels	0	7	9	34	44
	1	2	11	2	15
	2		2		2
	3	1	3		4
	4		1		1
	477	1			1
	491	1			1
	543		1		1
	551		1		1
Total		40	40	40	120

Figure 2 - 5

Notice that the column labeled Smoker gives the same frequencies report obtained in the previous section. The next two columns give the frequencies reports for ETS and NOETS. Differences among the distributions of the three groups can be found by comparing the three frequency distributions. For example, it is easy to see that the cotinine levels tend to be largest for the smokers, less for the ETS group, and smallest for the NOETS group (not apparent from the shortened table above, but clear from looking at the whole table in the SPSS Output Viewer window). Strangely, the two largest cotinine values are in the ETS group.

Section 2-2. Visualizing Data

It is said that a picture is worth a thousand words. Describing the distribution of data can be accomplished much more succinctly with pictures than with a frequencies report or a frequency table. Histograms, pie charts, bar charts, stem-and-leaf plots, boxplots, and other graphs are pictures of the information in a frequency table. A **histogram** is a picture of the information in the frequency table that shows the shape, center, and spread of the distribution. When the data have nominal or ordinal levels of measurement **pie charts** and **bar charts** can be used to describe the data. **Stem-and-leaf plots** are useful when the number of data values is not too large (say less than 100). They provide a way to see the shape of the distribution and without losing the original data values. **Boxplots** show the shape of the distribution and also provide a way to check for outliers in the data file.

The Frequencies procedure can be used to create graphical displays and calculate descriptive statistics for a single factor. To create bar charts, pie charts, or histograms open the Frequencies dialog box (Figure 2 - 1) and click the **Charts...** button. We discuss using the Explore procedure to produce descriptive statistics and graphical displays.

Histograms

To make histograms of the cotinine levels of the Smokers, ETS, and NOETS groups, choose **Analyze > Descriptive Statistics > Explore...** to open the Explore dialog box (Figure 2 - 6).

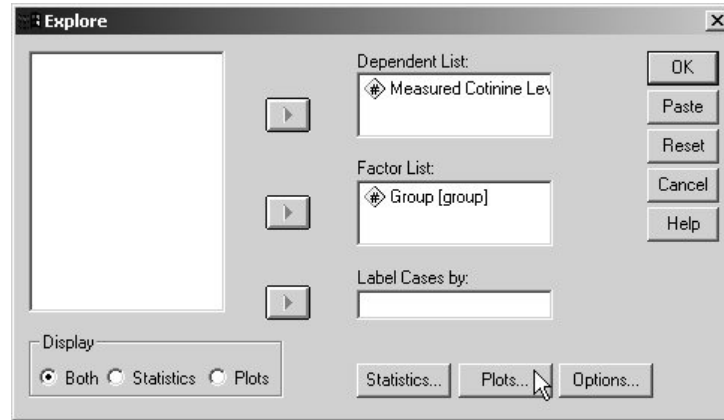


Figure 2 - 6

Copy the variable, *cotinine*, to the Dependent List and the variable, *group*, to the Factor List by selecting the variable and clicking the **Variable Paste** button. Next click the **Plots...** button to open the Explore: Plots dialog box (Figure 2 - 7).

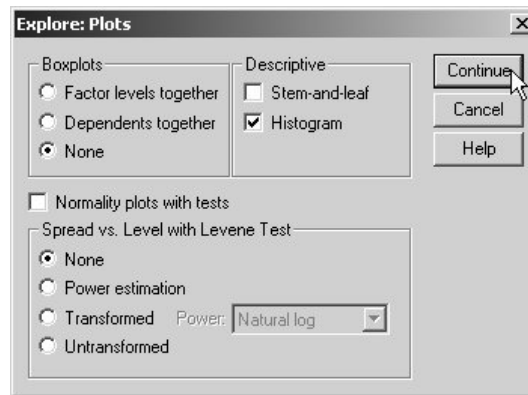


Figure 2 - 7

Notice that Figure 2 - 7 has both bullets and checkboxes. Checkboxes and bullets serve two different purposes. Within a grouping, only a single bullet may be chosen while multiple checkboxes can be selected. For example, if the checkboxes for **Histogram** and **Stem-and-leaf** are both chosen then both a histogram and stem-and-leaf plot will be produced.

Choose the checkbox for **Histograms** to make a Histogram for each combination of the factor levels (each group) of the variable in the Factor List and variable in the Dependent list. For example, in this problem histograms for Smoker, ETS, and NOETS of the cotinine variable will be produced. This same dialog box can be used to make **Stem-and-leaf Plots** and **Boxplots**, as well as some other plots that we will not discuss here. For now, choose the bullet for **None** under Boxplots. Click the **Continue** button to return to the Explore dialog box and then click the **OK** button to display the histograms in the Output Viewer window.

Scroll down the Output Viewer window to see the histograms for Smoker, ETS, and NOETS. The histogram for the Smoker group (Figure 2 - 8) shows the distribution of the data values to be fairly uniform between 0 and 325, except for two extremely large values. The two large values are separated from the rest of the data and are likely to be outliers. The histograms for ETS and NOETS are extremely right skewed with their values equal to zero.

Chart Editor

Histograms, in fact any chart created in SPSS can be customized. To customize the histogram for *smokers*, first select the histogram by clicking on **Group = Smoker** in the Outline window and then choose **Edit > SPSS Chart Object > Open** to open the chart editor (Figure 2 - 8). Alternately, you can double-click on the Histogram to open the Chart Editor. The **Chart Editor** window has its own menu (notice the menu along the top of the Chart Editor window has changed).

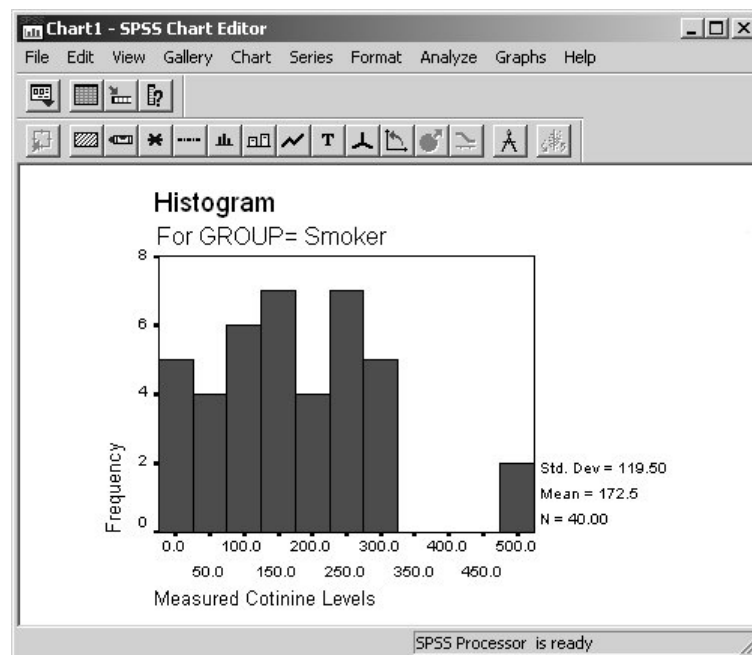


Figure 2 - 8

There are two axes, the **Interval Axis** (the axis from which the bars originate) and the **Scale Axis** (the axis which displays numerical values to scale). To modify the **Interval Axis**, choose **Chart > Axis...** and the Axis Selection dialog box will open. Choose the bullet for **Interval**, and then click the **OK** button. This will open the Interval Axis dialog box (Figure 2 - 9).

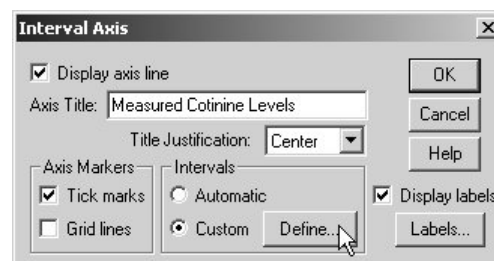


Figure 2 - 9

Replacing “**Measured Cotinine Levels**” in the Axis Title box with a new title will change the axis title. Choosing **Center** from the drop down list of choices under **Title Justification** will center the axis title under the graph. This dialog box can also be used to show or hide **Tick marks** and **Grid Lines**. These modifications are cosmetic and will have no affect on the shape of the histogram.

SPSS automatically determines the number of intervals and interval widths when making a histogram. Changing the number of intervals or the interval width will have an affect on the shape of the histogram. To change the number of class intervals and the interval widths in the histogram, choose the bullet for **Custom** and then click the **Define...** button to open the Interval Axis: Define Custom Interval dialog box (Figure 2 - 10).

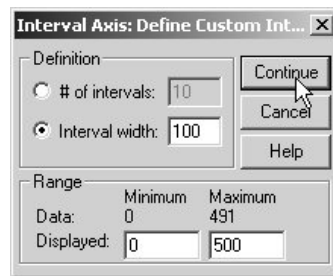


Figure 2 - 10

Click the button for **Interval width** and type in *100* and then type *0* in the box for Minimum and *500* in the box for Maximum. This will create classes of *0-100*, *100-200*, *200-300*, *300-400*, and *400-500*. This notation might be confusing because it is not be clear into which class a data value of *100* would be included. SPSS understands the interval *0-100* to be $0 \leq x < 100$, therefore *100* goes into the interval labeled *100-200*.

After making these changes, click the **Continue** button and then click the **OK** button in the Interval Axis dialog box. The Histogram has now been updated; you can now make more changes or close the Chart Editor. Choose **File > Close** from the menu to close the Chart Editor.

The shape of the histogram for smokers (Figure 2 - 11) indicates that the cotinine levels are fairly evenly distributed between *0* and *300*. Further analysis will be necessary to determine if the three data values larger than *300* are outliers.

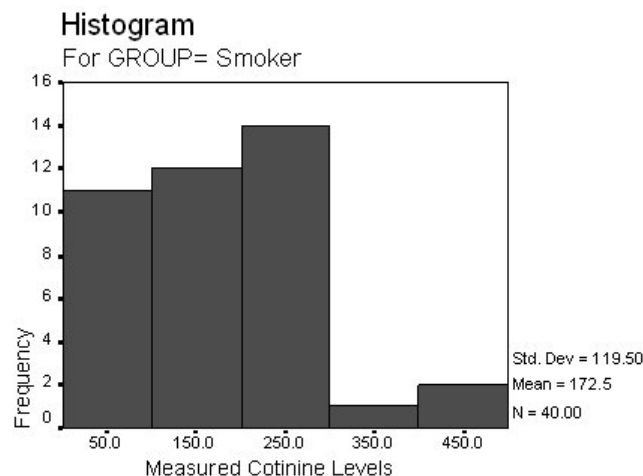


Figure 2 - 11

Stem-and-leaf plots

To make **stem-and-leaf plots** for the measured cotinine levels for the three groups: Smoker, ETS, and NOETS, choose **Analyze > Descriptive Statistics > Explore...** to open the Explore dialog box (Figure 2 - 6). Click the **Plots...** button to open the Explore: Plots dialog box (Figure 2 - 7). Choose the checkbox for **Stem-and-leaf** and click the **Continue** button to close the dialog box. Click the **OK** button in the Explore dialog box and the stem-and-leaf plots will appear in the Output Viewer window (Figure 2 - 12).

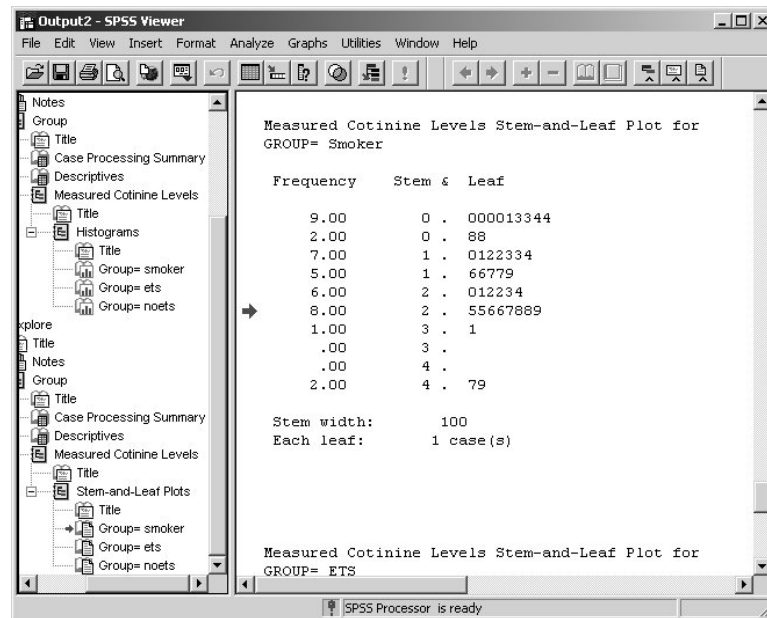


Figure 2 - 12

SPSS only shows the leading digit and the next digit in the number in the stem-and-leaf plot; that is, it truncates the data values to only two digits. To know the size of the data values you must read the Stem width. In the plot above, the Stem width is 100, which means that the first two stems in the plot display data values between 00-49 and 50-99, respectively. From the stem-and-leaf plot it is impossible to know what the exact value of the 1 on the 3-stem represents. We only know it is a value between 310 and 319. Looking back at the frequencies table we see this value is 313.

The stem-and-leaf plot for the Smoker group shows the similar information to that of the histogram. The data values are fairly evenly divided between 0 and 300 with 3 values larger than 300. There is more information here though since the actual data values are available. For example, it can be seen that one of the three data values that are larger than 300 is only about 310.

Boxplots

To make boxplots for the cotinine levels for the three groups: Smoker, ETS, and NOETS, choose **Analyze > Descriptive Statistics > Explore...** to open the Explore dialog box (Figure 2 - 6). Click the **Plots...** button to open the Explore: Plots dialog box (Figure 2 - 7). Choose the bullet for **Factor levels together** to make a separate Boxplot for each of the variables in the Dependent List in the Explore dialog box. If there are several variables in the Dependent List, choose the bullet for **Dependents together** to obtain side-by-side Boxplots. In this case, it does not matter since there is only one variable, *cotinine*, in the Dependent List. Click the **Continue** button, and then click the

OK button in the Explore dialog box, and the boxplots will appear side-by-side in the Output Viewer window (Figure 2 - 13).

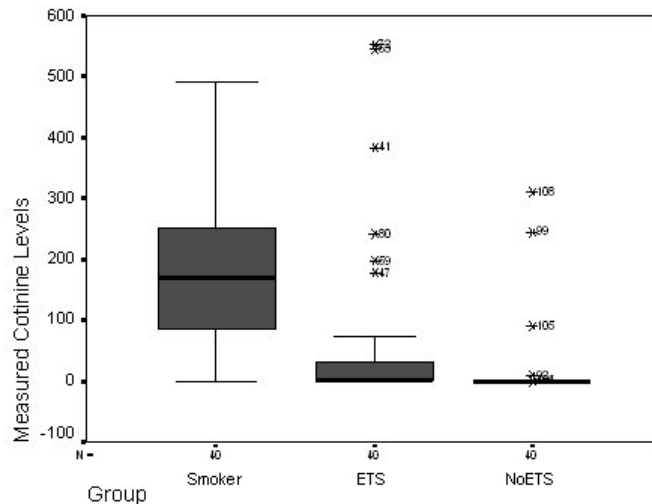


Figure 2 - 13

The boxplots clearly show that the cotinine levels for smokers are much larger than the cotinine levels of the ETS and NOETS groups. The two (or three) suspiciously large data values identified in the histogram and stem-and-leaf plots for Smokers apparently are not outliers. The ETS group has larger cotinine levels than the NOETS group. There are many outliers in the ETS and NOETS groups and the distributions are very right skewed.

Section 2-3. Scatter Diagram

A **Scatter Diagram** (also known as a **Scatterplot**) is a plot of paired (x, y) data with a horizontal x-axis and a vertical y-axis. The y-axis variable determines the vertical position of the point and the x-axis variable determines the horizontal position of the point. Scatter diagrams are useful for exploring the relationship between two variables. Exploring the relationship between two variables will be discussed in more detail in Chapter 9.

The **Health Exam Results** data from Data Set 1 in Appendix B of *Elementary Statistics* (this data appears on disk as **Mhealth.sav**) includes the weight (in pounds) and the waist circumference (in cm) for 40 males. Make a scatter diagram to determine if there is a relationship between the weight and waist circumference measurements.

Open the **Mhealth** (see Section 0-3) data file. To make a **Scatter diagram** of weight versus waist circumference choose **Graphs > Scatter...** to open the Scatterplot dialog box (Figure 2 - 14).

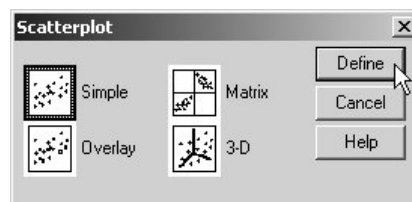


Figure 2 - 14

Choose the icon for **Simple** and then click the **Define** button to open the Simple Scatterplot dialog box (Figure 2 - 15).

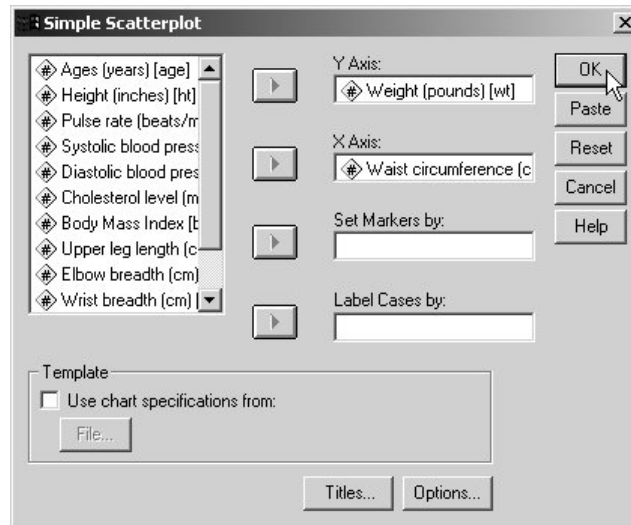


Figure 2 - 15

Select the **Weight (pounds) [wt]** for the Y Axis and **Waist circumference (cm) [waist]** for the X Axis by clicking on the variable label and then clicking the **Variable Paste** button. Click the **Titles...** button if you want to give your Scatter diagram a title. Click the **OK** button and the Scatterplot (Figure 2 - 16) will appear in the Output Viewer window.

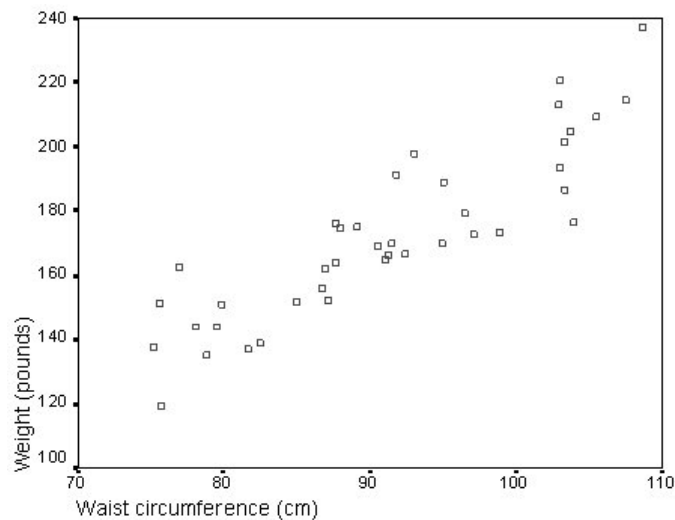


Figure 2 - 16

The scatter diagram indicates that as the waist circumference increases the weight of the person tends to increase as well. There appears to be an approximately linear relationship between the weight and the waist measurements.

Section 2-4. Descriptive Statistics

Often descriptive statistics are used to describe characteristics of variables in a data file. The **arithmetic mean** (or **average**), **median**, and **midrange** are measures of the center of a distribution. The **standard deviation**, **variance**, and **range** are measures of the spread of a distribution. **Quartiles** and **percentiles** are measures of position within in a distribution.

Descriptive Statistics for a Variable

Open the **Mhealth** data file (see Section 0-3). The Frequencies procedure can be used since there are no subgroups in this data file. Choose **Analyze > Descriptive Statistics > Frequencies...** to open the Frequencies: dialog box (see Figure 2 - 1). Copy the variables *age*, *weight*, and *pulse* to the Variable(s) list and click the **Statistics...** button to open the Frequencies: Statistics dialog box (Figure 2 - 17).

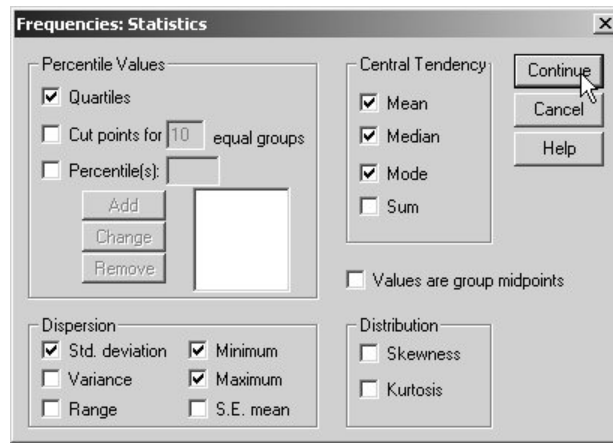


Figure 2 - 17

Check the checkboxes for any descriptive statistics to be included in the analysis. Click the **Continue** button to return to the Frequencies dialog box. The frequencies reports will be rather long and so it is probably best to uncheck the checkbox for **Display frequency tables**. Click the **OK** button and the descriptive statistics will appear in the Output Viewer Window (Figure 2 - 18).

Statistics

		Ages (years)	Waist circumference (cm)	Pulse rate (beats/min)
N	Valid	40	40	40
	Missing	0	0	0
Mean		35.48	91.285	69.40
Median		32.50	91.200	66.00
Mode		20	87.7 ^a	60 ^a
Std. Deviation		13.927	9.8619	11.297
Minimum		17	75.2	56
Maximum		73	108.7	96
Percentiles	25	25.25	83.125	60.00
	50	32.50	91.200	66.00
	75	45.50	101.900	76.00

a. Multiple modes exist. The smallest value is shown

Figure 2 - 18

The descriptive statistics for each variable are placed into one table. This makes it easy to compare the various statistics. SPSS appends a note to the table to indicate that the mode is not unique for waist circumference and pulse rate.

Descriptive Statistics for subgroups of a Variable

Open the **Cotinine** data file. Since there are three subgroups in this data file (indicated by the variable **group**) the Explore procedure will be used. Choose **Analyze > Descriptive Statistics > Explore...** to open the Explore dialog box (Figure 2 - 6). Copy **cotinine** into the Dependent List box and **group** into the Factor List box. To display descriptive statistics only and suppress the creation of plots choose the bullet for **Statistics**. Click the button for **Statistics...** to open the Explore: Statistics dialog box (Figure 2 - 19).

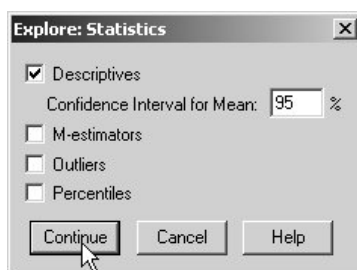


Figure 2 - 19

Choose the checkbox for **Descriptives** and click the **Continue** button. Click the **OK** button and the descriptive statistics will appear in the Output Viewer window. Only a portion of the Descriptives table is shown (Figure 2 - 20) since the output is quite long.

Descriptives						
Group				Statistic	Std. Error	
Measured Cotinine Levels	Smoker	Mean		172.48	18.894	
		95% Confidence Interval for Mean	Lower Bound	134.26		
			Upper Bound	210.69		
		5% Trimmed Mean		164.72		
		Median		170.00		
		Variance		14279.846		
		Std. Deviation		119.498		
		Minimum		0		
		Maximum		491		
		Range		491		
		Interquartile Range		166.00		
		Skewness		.588		.374
		Kurtosis		.520		.733
	ETS	Smoker	Mean			60.58
95% Confidence Interval for Mean			Lower Bound	16.41		
			Upper Bound	104.74		
		5% Trimmed Mean		36.92		
		Median		1.50		
		Variance		19067.174		
		Std. Deviation		138.084		

Figure 2 - 20

The Explore procedure calculates the several descriptive statistics (mean, median, variance, standard deviation, minimum, maximum, range and interquartile range) for each variable. There is other information in the Descriptives table, which we will put off describing it until we discuss inferential statistics.

Section 2-5. Exercises

1. Consider problem 20 in Chapter 2-2 of *Elementary Statistics. Regular Coke and Diet Coke*. Refer to Data Set 17 in Appendix B (this data appears on disk as **Cola.sav**). Construct a relative frequency distribution for the weights of regular coke by starting the first class at 0.7900 lb. and use a class width of 0.0050 lb. Then construct another relative frequency distribution for weights of diet Coke by starting the first class at 0.0075 lb and use a class width of 0.0050 lb. Then compare the results and determine whether there appears to be a significant difference. If so, provide a possible explanation for the difference.
2. Consider problem 9 in Chapter 2-3 of *Elementary Statistics. Bears*. Refer to Data Set 9 in Appendix B (this data appears on disk as **Bears.sav**). Construct a histogram for the measured weight of the bears with 11 classes with a lower class limit of 0 and a class width of 50 lbs.
3. *Regular Coke and Diet Coke*. Refer to Data Set 17 in Appendix B (this data appears on disk as **Cola.sav**). The data set contains the weight and volume measurements on 36 cans of Regular Coke, Diet Coke, Regular Pepsi, and Diet Pepsi.
 - a. Determine the mean and standard deviation of the weights of Regular Coke, Diet Coke, Regular Pepsi, and Diet Pepsi. What can you conclude about the weights of Colas?
 - b. Find the 5-number summary for the weights of Regular Coke, Diet Coke, Regular Pepsi, and Diet Pepsi.
 - c. Make side-by-side boxplots of the weights of Regular Coke, Diet Coke, Regular Pepsi, and Diet Pepsi. What can you conclude about the distributions of the Colas?
4. *Age of Presidents*. A Senator is considering running for the U.S. presidency, but she is only 35 years of age, which is the minimum required age. While investigating this issue, she finds the ages of the past presidents when they were inaugurated, and those ages are listed in Table 2 - 1.

57	61	57	57	58	57	61	54	68	51
49	64	50	48	65	52	56	46	54	49
51	47	55	55	54	42	51	56	55	51
54	51	60	52	43	55	56	61	52	69
64	46	54							

Table 2 - 1

Using the listed ages, find the

- a. Mean,
- b. Median,
- c. Mode,
- d. Midrange,
- e. Range,
- f. Standard deviation,
- g. Variance,
- h. Q_1 and Q_3 ,
- i. and P_{10} .