Chapter 5

Summarizing Multivariate Data: Association Between Numerical Variables

In this chapter, we illustrate methods for summarizing the relationship or association between two or more variables measured on numerical scales. (Chapter 6 discusses association among categorical variables.) This association can be expressed either graphically or numerically. The graphical technique is the scatter plot, and the numerical index is the correlation coefficient. In the sample, the correlation coefficient is represented by *r*.

These methods can be used to answer questions about the relationship between variables such as:

- Is there a relationship between scores on the language and nonlanguage portions of an IQ test? That is, do students with high language scores also have high nonlanguage scores? (If so, this would be a positive association.)
- Is there a relationship between the number of days students are absent from class and their score on the final exam? That is, do students with a large number of absences have lower grades, whereas those with few absences have higher grades?
- Is the degree of exposure to radioactive materials associated with the rate of cancer mortality?

- Is there a relationship between fat content and number of calories in different breakfast cereals?

This chapter describes how to obtain scatter plots and correlation coefficients between numerical variables using SPSS for Windows.

5.1 ASSOCIATION OF TWO NUMERICAL VARIABLES

Scatter Plots

A scatter plot is a graphical technique used to illustrate the association of two numerical variables. Data are represented visually by making a graph with two axes: horizontal (x axis) and vertical (y axis). Each point in the plot represents one observation. When all observations are plotted, the diagram conveys information about the direction and magnitude of the association of the two variables (x and y).

To illustrate the use of SPSS for scatter plots, let us examine the relationship between language and nonlanguage IQ for 23 second-grade children. The data are contained in the file "IQ.sav."

To obtain a scatter plot, do the following:

1. Click on **Graphs** from the menu bar.
2. Click on **Scatter/Dot** from the pull-down menu.
3. Click on **Simple Scatter** and then on **Define** to open the Simple Scatterplot dialog box (Fig. 5.1).
4. Click on the "nonlanguage IQ" variable and move it to Y Axis box.
5. Click on the "language IQ" variable and move it to X Axis box.
6. Click on **OK** to close this dialog box and create the scatter plot.

The SPSS Viewer contains the scatter plot like that in Figure 5.2. Each point on the plot represents one observation. For example, one person had a language IQ score of 84 and a nonlanguage IQ score of 30, as marked in the graph. Another individual had scores of 109 and 74, respectively. Can you find it on the graph?

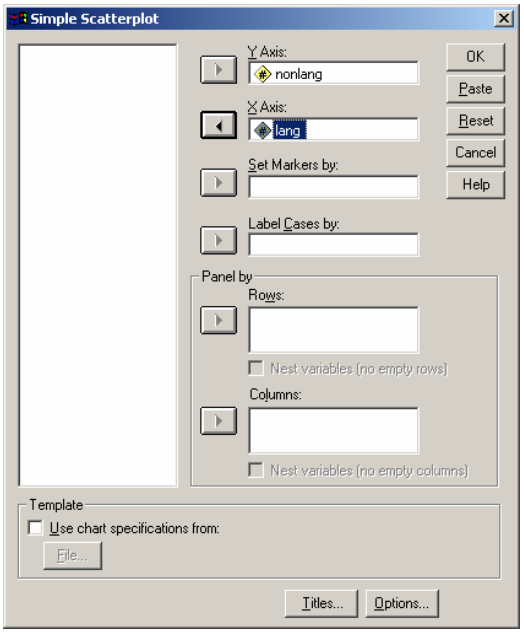

Figure 5.1 Simple Scatter Plot Dialog Box

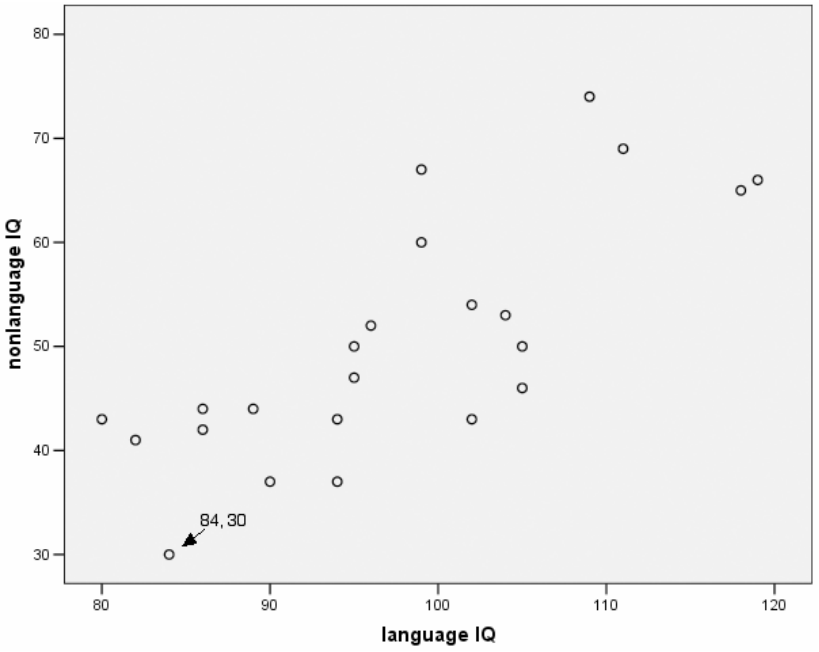

Figure 5.2 Scatter Plot Showing Positive Association

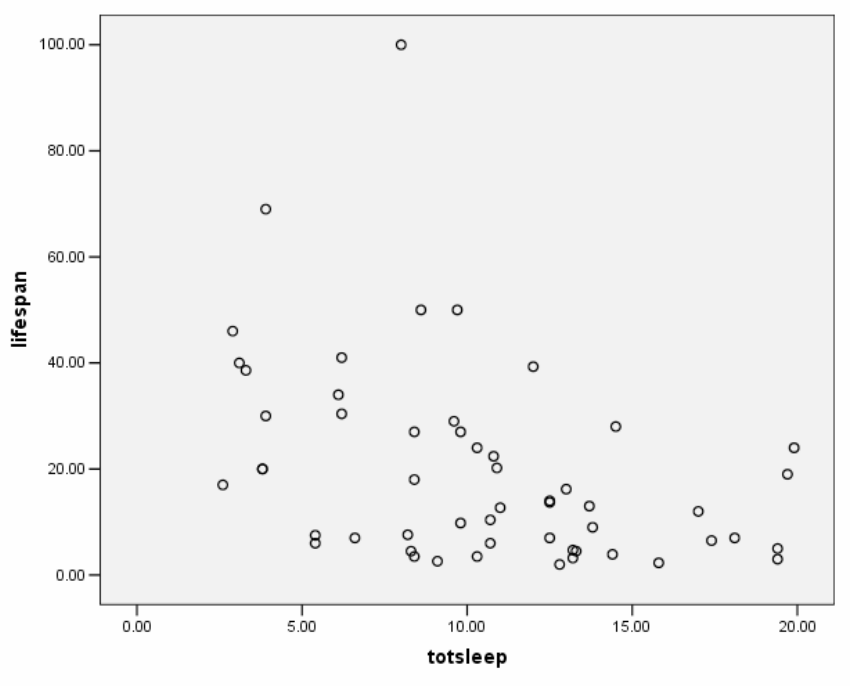

Figure 5.3 Scatter Plot Showing Negative Association

There is a positive association between language IQ and nonlanguage IQ. That is, individuals with high language IQ scores also tend to have high nonlanguage IQ scores, and those with low scores on one measure tend to have low scores on the other.

We will create another scatter plot using the data in "sleep.sav." In Chapter 4, we summarized the "lifespan" and "totsleep" variables separately. Now let us examine the relationship between the two by creating a scatter plot with total hours of sleep per day on the x-axis and lifespan on the y-axis. Following steps 1–6 will produce the results shown in Figure 5.3. In this plot, there appears to be a negative relationship between the two variables. That is, animal species that sleep many hours per day tend to have a lower life span than those who sleep fewer hours per day.

Changing the Scales of the Axes

SPSS chooses the scales for the *x* and *y* axes that best fit the range of the data, but you may manually adjust the scales if you wish. Below are steps to edit the *x*-axis scale of any scatter plot:

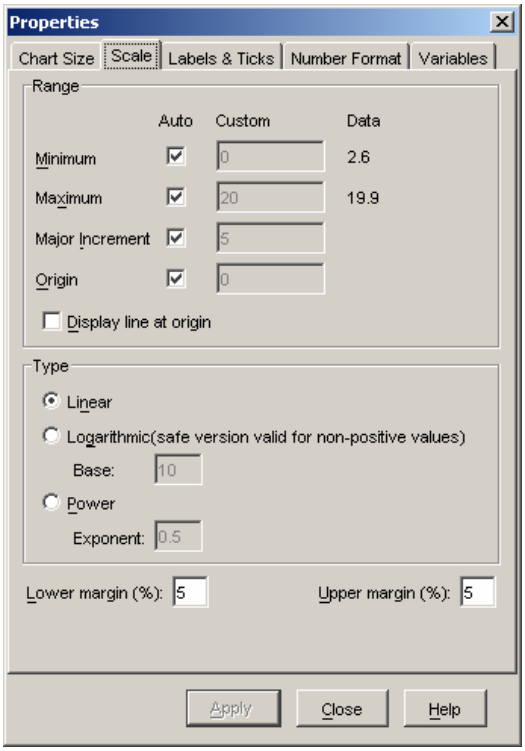

Figure 5.4 Scale Tab of Properties Dialog Box for X Axis of the Total Sleep by Life Span Scatter plot

1. Double click on the chart in the SPSS Viewer. This opens the Chart Editor.
2. Select **Edit** from the menu bar.
3. Click on **Select X Axis** from the pull-down menu to open the Properties Dialog box.
4. Select the **Scale** tab (Fig. 5.4).
5. Edit the Minimum and Maximum points of the range by clicking off the **Auto** selections and entering the desired numbers in the Custom sections.
6. Click **Apply** to redraw the scatter plot and then **Close** to apply the changes and close the Properties dialog box.
7. Click the **X** in the top right corner to close the Chart Edit Window.

Other Information Revealed by Scatter Plots

Examining a scatter plot can reveal important information. As we have discussed, it is possible to discern positive and negative associations. Scatter plots

lso allow one to determine, for instance, when a relationship between two variables is nonlinear and/or when bivariate outliers exist. We will illustrate the latter of these cases.

The data file "IQ.sav" contains information regarding language and nonlanguage IQ scores for 23 students. The range of the language scores is 80--119, and the range of the nonlanguage scores is 30--74. As we saw in Figure 5.2, the scatter plot of these two variables shows a positive association.

Now, open the data file "IQ2.sav." This file contains the same scores in the original "IQ.sav" file, plus one additional data point, a student with a language score of 80 and a nonlanguage score of 72. Considering each of these scores alone, neither is an outlier; each is within the range of the original scores for its variable.

Create the scatter plot for the variables with language on the *x*-axis and nonlanguage on the *y*-axis (see Fig. 5.5). Notice the outlier; we have labeled the point (80,72). This point represents an individual with a very low language IQ and a very high nonlanguage IQ. There are no other data points in its vicinity, and it is clearly an outlier. The data analyst should attempt to understand why this unusual pairing of values occurred.

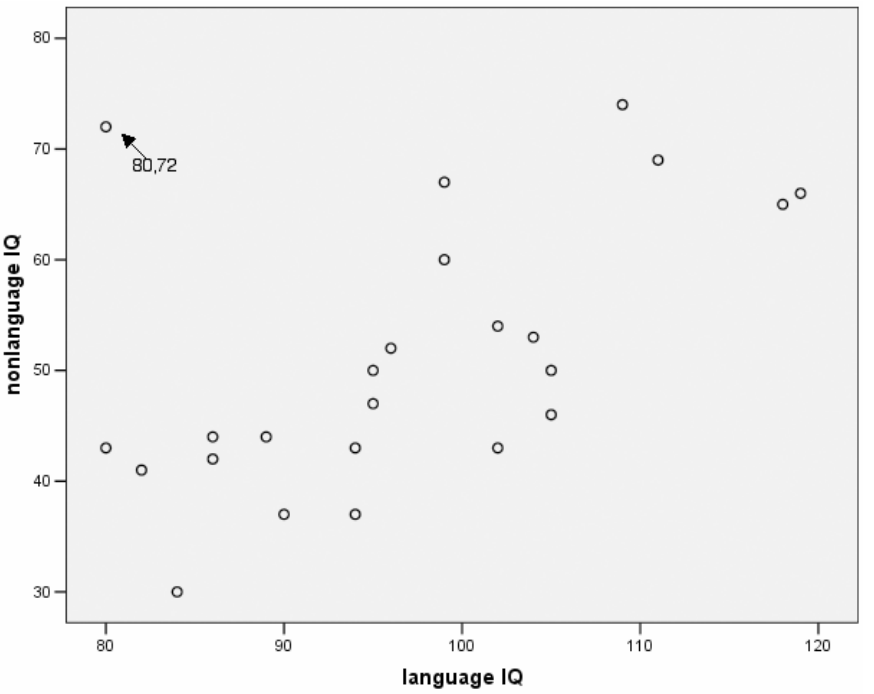

Figure 5.5 Scatter Plot Showing a Bivariate Outlier

The Correlation Coefficient

The Pearson correlation coefficient indicates the degree of linear association of two numerical variables. The correlation coefficient ranges from –1 to +1. A positive value (e.g., .10, .40, .80) reflects a direct associations between the two variables and a negative value (e.g., –.20, –.40, –.80) reflects a negative or inverse relationship. The strength of association is indicated by the absolute value of the correlations; for example, the values –.80 and .80 represent equally strong relationships. Zero is the weakest correlation, and 1 (or –1) the strongest. As a rule of thumb, correlations between 0 and .30 (absolute value) are considered weak; those between .31 and .60 (absolute value) are considered moderate, and those greater than .60 (absolute value) are considered strong.

We will illustrate the procedure for calculating the correlation coefficient using the "cereal.sav" data file. This file contains nutritional information on 77 brands of cereal. We will examine the relationship between sugar content and calories. You may begin by opening the file. To compute the correlation coefficient for these two measured variables:

1. Click on **Analyze** from the menu bar.
2. Click on **Correlate** from the pull-down menu.
3. Click on **Bivariate** from the pull-down menu. This opens the Bivariate Correlations dialog box (see Fig. 5.6).
4. Click on the "sugar" and "calories" variables and move them to the Variables box by clicking on the **right arrow button**.
5. Click on **OK** to run the procedure.

The output should look like that in Figure 5.7.

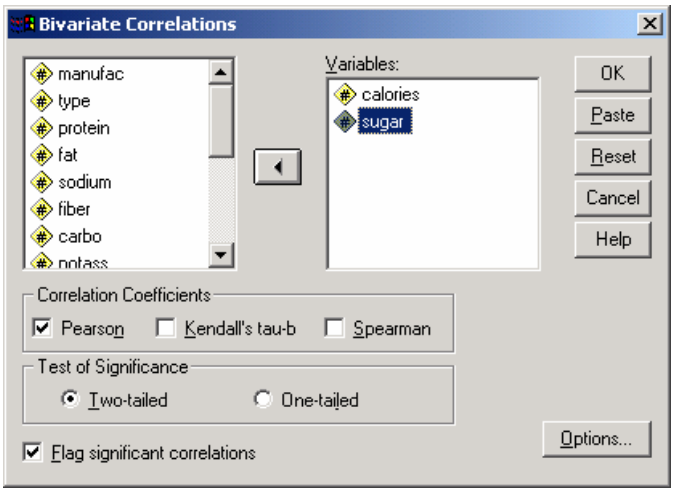

Figure 5.6 Bivariate Correlations Dialog Box

Correlations

		calories	sugar
calories	Pearson Correlation	1	.562**
	Sig. (2-tailed)		.000
	N	77	77
sugar	Pearson Correlation	.562**	1
	Sig. (2-tailed)	.000	
	N	77	77

**. Correlation is significant at the 0.01 level

Figure 5.7 Correlation of Sugar Content and Calories

SPSS lists the correlation coefficients it calculates in a correlation matrix. The values on the diagonal are all 1 because they represent the correlations of each variable with itself. The values above and below the diagonal are identical. In the example, the correlation between amount of sugar and calories is .562. This is a moderate, positive association; cereals with higher sugar content tend to have more calories.

Let us repeat this procedure two more times, first with the "IQ.sav" data file, and then with the "IQ2.sav" data file. For the original data file, you should determine that the correlation is .769. Simply by adding one outlying observation ("IQ2.sav"), the correlation coefficient decreases to .561. This is a dramatic change, and underscores the need to examine the scatter plot of two variables in addition to calculating the coefficient.

Rank Correlation

A (Spearman) rank correlation is the correlation coefficient computed from two variables that are measured on an ordinal rather than a numerical scale. Because the coefficient is calculated in a similar manner to the Pearson coefficient, it is discussed in this chapter. Some variables may be recorded as ranks in the first place, for example, fastest runner (1), second fastest (2), ..., slowest runner (n). Data on numerical scales can be re-expressed as ranks. For example, the student scoring 100 on a test may be given rank 1, the student with 96 given rank 2, and so on. Let us repeat the calculation of the correlation of sugar content and calories, this time using the Spearman procedure. Because these data are on a refined numerical scale, we shall rank them in order to illustrate the rank correlation procedure. (Note: we do this only for illustrative purposes; generally, when two variables are measured on a numerical scale, the Pearson correlation is a more precise measure of linear association.)

To rank these two variables:

1. Click on **Transform** from the menu bar.
2. Click on **Rank Cases** from the pull-down menu. This opens the Rank Cases dialog box (see Fig. 5.8).
3. Move the "sugar" and "calories" variables to the Variable(s) box by clicking on the name of each variable and then on the **top right arrow button**.
4. Click on **OK**.

SPSS creates two new variables, "rcalorie" and "rsugar," which consist of ranked data. (SPSS automatically creates the new variable names by adding an r-prefix to the original variable names.) The procedure to compute the Spearman correlation for these two variables is similar to the steps outlined for calculating the Pearson coefficient. The only difference is that you must click off the Pearson option (the default) and click on the Spearman option in the Correlation Coefficients box (see Fig. 5.6).

Your output should look like Figure 5.9.

The correlation of mortality and exposure, when ranked, is .596. This is slightly larger, but consistent with, the Pearson correlation coefficient, .562.

5.2 MORE THAN TWO VARIABLES

Correlation Matrix

In some instances, you may be interested in examining the correlation between many pairs of variables. SPSS can calculate many pairs of correlations at one time. The correlations for each pair of measures are computed and arranged in a matrix that has one row for each variable and one column for each variable.

The procedure is the same as that detailed in Section 5.1, but you need to include the names of all of the variables for which you desire correlations in the Variables box of the Bivariate Correlations dialog box.

To illustrate, open the "fire.sav" data file and create a correlation matrix for the following variables: "stair," 'body," "obstacle," "agility," "written," and "composit" by including the names of all of them in the Variables box. Your output should look like Figure 5.10.

To find the correlation between body drag test time and obstacle course time, for instance, find the intersection of the column labeled body and the row labeled obstacle in Figure 5.10. The correlation coefficient is .759. (Note: this is the same number that is found at the intersection of the row labeled body and the column labeled obstacle. In other words, the correlation matrix is symmetrical.)

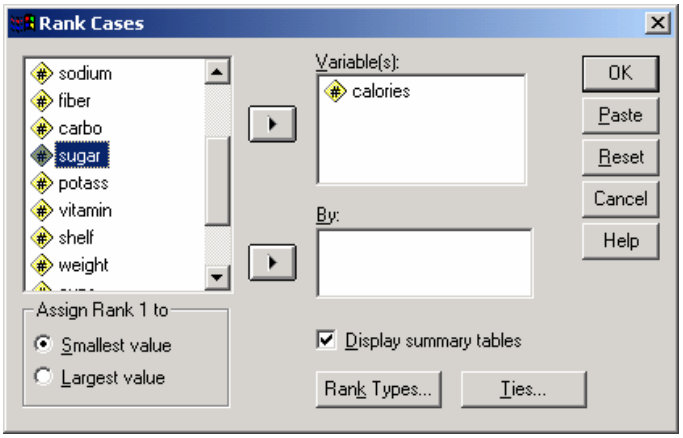

Figure 5.8 Rank Cases Dialog Box

Correlations

			RANK of CALORIES	RANK of SUGAR
Spearman's rho	RANK of CALORIES	Correlation Coefficient	1.000	.596**
		Sig. (2-tailed)	.	.000
		N	77	77
	RANK of SUGAR	Correlation Coefficient	.596**	1.000
		Sig. (2-tailed)	.000	.
		N	77	77

**. Correlation is significant at the .01 level (2-tailed).

Figure 5.9 Spearman Correlation Output

It is also possible to discern patterns among correlations in the matrix. For example, the correlations among body drag, stair climb, and obstacle course times are all positive and moderately strong to strong (ranging from .734 for obstacle course with stair climb, to .906 for stair climb with body drag). Because all are measures of athletic behavior, the positive association is to be expected. The written test score is negatively correlated with these tasks. Thus, more agile applicants (those with lower time scores) tend to have higher scores on the written test, and vice versa.

Correlations

		stair	body	obstacle	agility	written	composit
stair	Pearson Correlation	1	.906**	.734**	.954**	-.337	-.898**
	Sig. (2-tailed)		.000	.000	.000	.079	.000
	N	28	28	28	28	28	28
body	Pearson Correlation	.906**	1	.759**	.962**	-.466*	-.940**
	Sig. (2-tailed)	.000		.000	.000	.012	.000
	N	28	28	28	28	28	28
obstacle	Pearson Correlation	.734**	.759**	1	.875**	-.495**	-.875**
	Sig. (2-tailed)	.000	.000		.000	.007	.000
	N	28	28	28	28	28	28
agility	Pearson Correlation	.954**	.962**	.875**	1	-.458*	-.970**
	Sig. (2-tailed)	.000	.000	.000		.014	.000
	N	28	28	28	28	28	28
written	Pearson Correlation	-.337	-.466*	-.495**	-.458*	1	.660**
	Sig. (2-tailed)	.079	.012	.007	.014		.000
	N	28	28	28	28	28	28
composit	Pearson Correlation	-.898**	-.940**	-.875**	-.970**	.660**	1
	Sig. (2-tailed)	.000	.000	.000	.000	.000	
	N	28	28	28	28	28	28

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Figure 5.10 Correlation Matrix

Missing Values

Only those cases with values for both variables can be used in computing a correlation coefficient. There are two ways to cause SPSS to eliminate cases with missing values: "listwise deletion" and "pairwise deletion." (See Chapter 1 for a discussion of missing values.) As an example, suppose that the third case in our data file were missing a score on the written test. Because the third case does not have complete information on all of the variables, listwise deletion eliminates the third case from the computation of all correlation coefficients. All correlations are calculated from the remaining 27 cases. Pairwise deletion eliminates the third case when computing only those correlations that involve the written test; thus some coefficients would be based on 27 observations and others would be based on all 28.

The default option in SPSS is pairwise deletion, which uses the maximum number of cases for each coefficient. You may request listwise deletion be clicking on the **Options button** of the Bivariate Correlations dialog box (Fig. 5.6) and choosing **Exclude cases listwise**.

Chapter Exercises

5.1 In Section 5.1 of this chapter, we created the scatter plot of total hours of sleep with lifespan for the 62 mammals in the "sleep.sav" data file.

a. Refer to the graph (Fig. 5.3) and estimate the value of the correlation coefficient. Check your estimate by calculating the Pearson correlation using SPSS.

b. Eliminate any outliers and recalculate the correlation. How did this affect your results?

5.2 The "cancer.sav" data file contains data on cancer mortality and index of exposure to nuclear materials for residents in counties near a nuclear power plant. Using this data file:

a. Compute the Pearson correlation coefficient between "expose" and "mortalit" and comment on the strength and direction.

b. Rank the two variables using the Transform procedure, and compute the Spearman correlation coefficient of the ranked variables.

c. How do the two coefficients compare?

5.3 The data file "enroll.sav" contains information from a random sample of 26 school districts. Information was obtained on the following variables:

(1) district enrollment;

(2) the percentage of students in the district who are African-American;

(3) the percentage of students who pay full price for lunches;

(4) an index of racial disproportion in classes for emotionally disturbed children (which is positive if the proportion of African-American students is greater than the proportion of white students).

Using this data file, compute a correlation matrix for all four variables and use it to answer the following questions:

a. Which correlation is largest (in magnitude)?

b. Explain what is meant by the negative correlation between enrollment and percent of African-Americans.

c. Racial disproportion is most highly associated with which other variable? What is the magnitude and direction of the association?

5.4 It has been shown that the relationship between amount of stress and work productivity is curvilinear. In other words, extremely low and extremely

high amounts of stress are related to low work productivity, but moderate amounts of stress are associated with the maximum amount of productivity.

a. Create a hypothetical data set with two variables — amount of stress and work productivity — that you think will illustrate this relationship. Your data file should have a minimum of 20 observations.

b. Use SPSS to produce a scatter plot of your data. Did you succeed in simulating a curvilinear relationship?

c. Name another instance in which you might find a curvilinear relationship between two variables.