CHAPTER 4

Data with Two Variables

In Chapter 2, we discussed methods of displaying data in tables and charts. Chapter 3 dealt with methods for describing data with numbers. So far, however, the focus has been on **univariate** data sets, those with only one variable. This chapter will focus on methods for displaying and describing **bivariate** data sets, specifically those with two quantitative variables.

Why deal with two variables at once? Why not just display the data on each variable separately, using the methods we already know? The reason is that variables are sometimes related to each other: as one variable increases, the other tends to change in a particular direction, either increasing or decreasing. This type of relationship between variables is called **correlation**. By dealing with each variable separately, we would not see the correlation between the variables.

Strictly speaking, the relationships between variables that we will consider in this chapter are **linear relationships**. More complex relationships, involving curves rather than lines, are certainly possible, but are beyond the scope of this book.

4.1 Scatter Plots and Correlation

Data on two quantitative variables is typically displayed in a **scatter plot**. A scatter plot is essentially a line plot with two different number lines, one vertical and one horizontal. Anyone who has graphed equations in an algebra class is familiar with graphing points in a coordinate plane. That's all we're talking about. The difference between statistics and algebra is that the points we get from data should not be expected to lie on a nice straight line or a simple curve such as a parabola. Rethinking some ideas from high school algebra will be helpful in understanding the idea of correlation. One of the ideas we need to revisit is the idea of positive and negative slope.

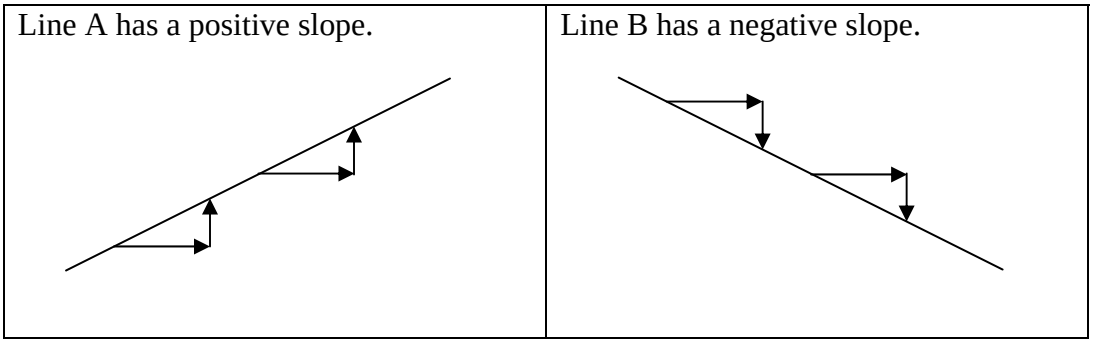


Figure 4.1.1 *Lines with positive and negative slope*

Exploration 4.1

Refer to the lines in Figure 4.1.1 above.

1. As you move from left to right along either Line A or Line B, are the ***x*-values** increasing or decreasing? (Hint: Your answer should be the same for both lines.)

2. Line A has a **positive** slope. As you move from left to right along this line, are the ***y*-values** increasing or decreasing?

3. Line B has a **negative** slope. As you move from left to right along this line, are the ***y*-values** increasing or decreasing? (Hint: Your answer should not be the same for both lines.)

4. As *x* increases along a line with positive slope, *y* _____.

5. As *x* increases along a line with negative slope, *y* _____.

Now look at the scatter plots in the figure below. They are taken from *MathScape: Looking Behind the Numbers* (p. 21).

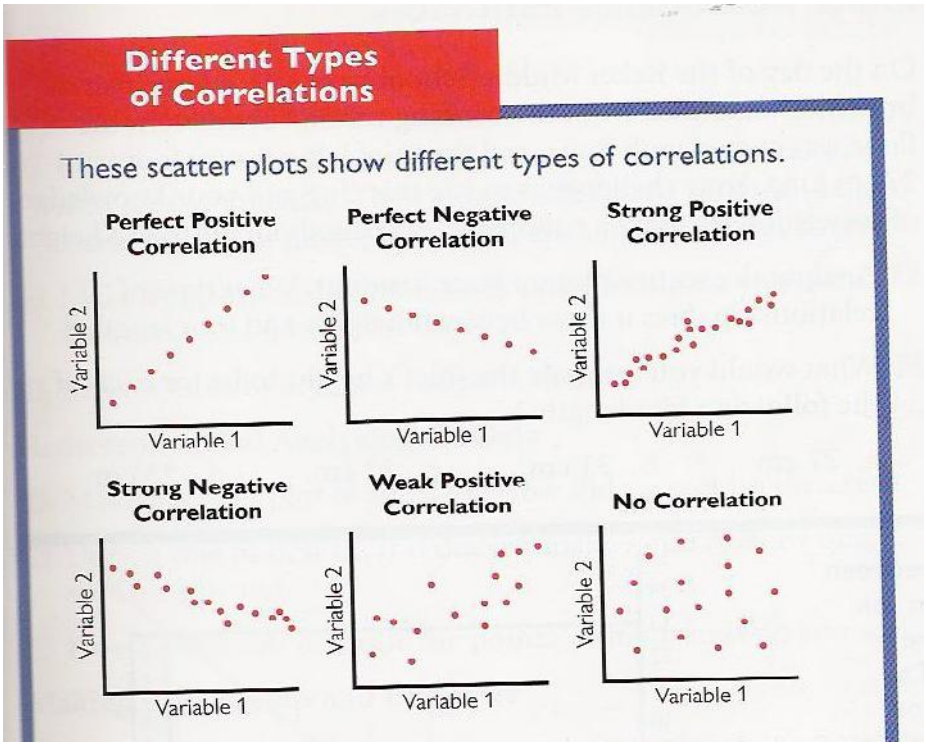


Figure 4.1.2 *Different Types of Correlation*

Notice that in the first two scatter plots, those with perfect correlation, the data points lie exactly in a straight line. If the line has positive slope, the correlation is positive. If the line has negative slope, the correlation is negative. When the correlation is less than perfect, the points do not lie exactly on a line, but might be thought of as being near a line. The more scattered the points are, the farther they are from this imaginary line, the weaker the correlation.

A different way to look at correlation relates to Exploration 4.1 above. For a line with positive slope, the points get higher as we move from left to right. In other words, y increases as x increases. When looking at a scatter plot, if the points tend to rise as you look from left to right, the correlation is positive. The stronger this tendency is, the stronger the correlation.

Likewise, for a line with negative slope, the points get lower as we move from left to right; y decreases as x increases. If a scatter plot has this same tendency, the correlation is negative. The stronger this tendency is, the stronger the correlation.

As *MathThematics, Book 3* (p. 538) defines it:

> Two variables that are related in some way are said to be *correlated*. There is a **positive correlation** if one variable tends to increase as the other increases. There is a **negative correlation** if one variable tends to decrease as the other increases.

Figure 4.1.3 below reproduces page 42 from *Mathematics In Context: Insights Into Data*.

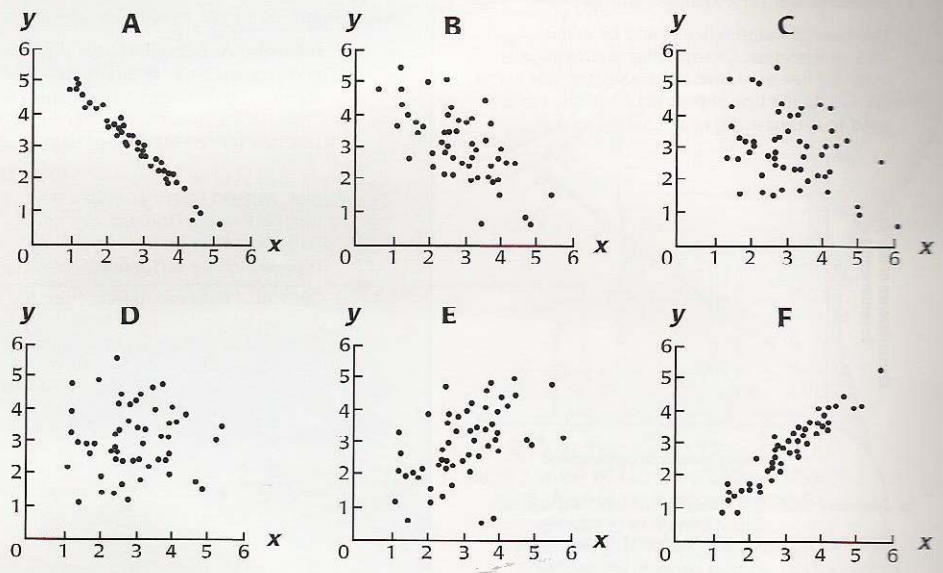
Focus on Understanding

1. Answer the questions in Figure 4.1.3.

2. Do you think that, in general, it will always be clear from a scatter plot whether the correlation between two variables is strong or weak? If we added a third category, moderate, would that eliminate disagreement about the strength of correlations?

Scatter plots are often made to show the relationship between two variables. The points on a scatter plot, or scatter diagram, look like a "cloud" of points. When there is a strong relationship between the two variables, the graph resembles a line. If the points appear as a straight line, you can say there is a strong *correlation* between the two variables.

8. Describe the correlation in the following scatter plots. Indicate whether there is no correlation, a weak correlation, or a strong correlation.



9. For which of the above scatter plots can you best predict the value of y when $x = 4$? Explain your reasoning.

In the diagrams above, you can see that a correlation can be weak or strong. Correlations can also be positive or negative.

10. a. What do you think is meant by the phrase "negative correlation"?

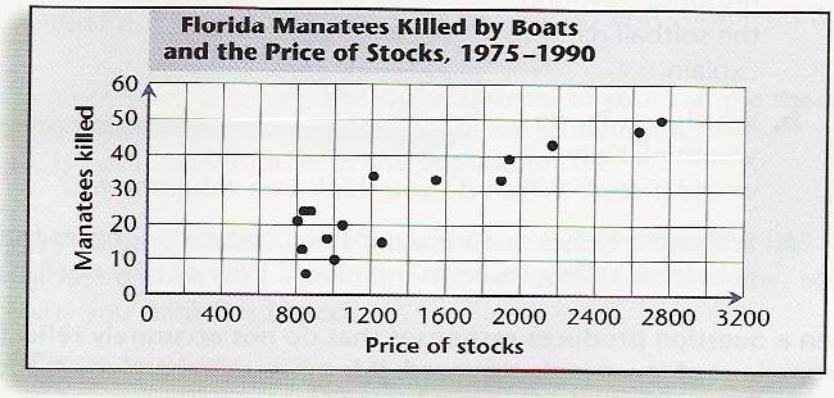
b. Which of the above scatter plots show a negative correlation?

Figure 4.1.3 *Describing Correlation*

Correlation and Cause-and-Effect

It is important to understand that correlation does not imply a cause-and-effect relationship between the variables. In other words, just because one variable tends to increase (or decrease) as the other increases, that does not necessarily mean that a change in one variable causes the other to change. Consider the exercise in Figure 4.1.4 below. It is taken from *MathThematics, Book 3* (p. 539).

13 The scatter plot below shows the number of Florida manatees killed by boats and the price of stocks as measured by the Dow Jones Industrial Average for the years 1975–1990.



a. Is there a *positive correlation*, a *negative correlation*, or *no correlation* between the price of stocks and the number of manatees killed by boats?

b. Would it be correct to say that a rise in stock prices tends to *cause* an increase in the number of manatees killed by boats?

c. Discussion If a correlation exists between two variables, does that necessarily mean there is a cause-and-effect relationship between the variables? Explain.

Figure 4.1.4 *Stock Prices and Manatees Killed*

Notice that the price of stocks and the number of manatees killed appear to have a positive correlation, but there is no cause-and-effect relationship between them. An increase in the number of manatees killed does not cause stock prices to rise, nor do rising stock prices cause more manatees to be killed. Correlation does not imply a cause-and-effect relationship.

So far, everything discussed in this section is well within the capabilities of middle school students. In the newer middle school curricula, students are expected to:

- Understand the idea of positive and negative correlation and be able to identify examples of these from a scatter plot.
- Understand that correlations between variables can have different strengths. Some variables are strongly correlated; others are weakly correlated. They should be able to identify examples of strong and weak correlation from scatter plots.
- Understand that correlation does not necessarily imply a cause-and-effect relationship. Two variables can be correlated without a change in one causing a change in the other.

College students and middle school teachers should have a deeper understanding of correlation, as something that can be measured. The next section deals with Pearson's correlation coefficient, a formula statisticians use to measure correlation between two variables. The value of this

measurement indicates not only whether one variable tends to increase or decrease as the other increases, but also how strong this tendency is.

4.2 Pearson's Correlation Coefficient

This section will explore the ideas behind Pearson's correlation coefficient. To help illustrate these ideas without getting too bogged down in computations, we'll use the small, highly rigged data set at the right, where most of the computations work out nicely. By the end of the section, you should be able to calculate the value of Pearson's coefficient for more complex data sets, interpret what this value indicates about the data, and have some understanding of why it works.

x	*y*
1	1
2	2
3	4
5	2
6	4
7	5

To begin to explore the idea of measuring correlation between two variables, we need to define the three S's:

The Three S's

$$S_{xx} = \sum (x - \bar{x})^2 \qquad S_{yy} = \sum (y - \bar{y})^2 \qquad S_{xy} = \sum (x - \bar{x})(y - \bar{y})$$

You've seen S_{xx} before. Remember from Chapter 3 that variance is given by the formula:

$$s^2 = \frac{\sum (x - \bar{x})^2}{n - 1}$$

S_{xx} is the numerator of this variance formula. S_{yy} is similar, but for the *y*-values instead of the *x*-values. We need both of these because we are now working with two variables instead of one as in Chapter 3.

The third of the S's, S_{xy}, is the key to measuring correlation. To understand S_{xy}, we start with a scatter plot of the data and draw in two additional lines, called mean lines: a horizontal line $y = \bar{y}$, where the *y*-value is the average of the *y*-values from the data points, and a vertical line $x = \bar{x}$, where the *x*-value is the average of the *x*-values from the data points. In this case, $\bar{y} = 3$ and $\bar{x} = 4$, so the two mean lines are $y = 3$ and $x = 4$. The two mean lines divide the *xy*-plane (and the data) into four sections. See Figure 4.2.1 below.

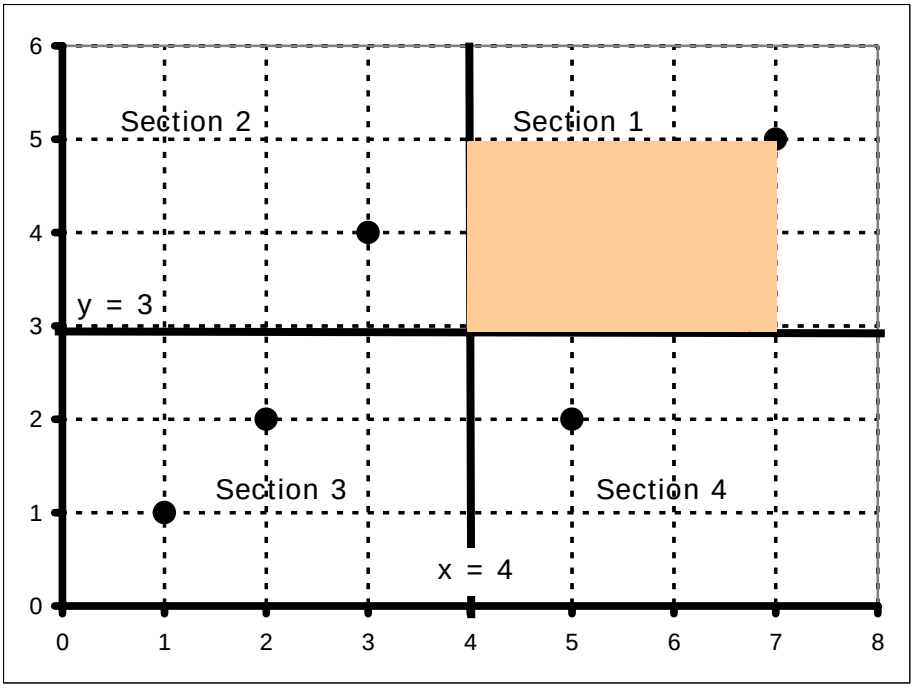


Figure 4.2.1 *Dividing the plane into four sections*

Consider the data point (7,5) in Section 1. For this point, $x - \bar{x} = 7 - 4 = 3$, a positive number. Also for this data point, $y - \bar{y} = 5 - 3 = 2$, also a positive number. The product, $(x - \bar{x})(y - \bar{y}) = 3 \cdot 2 = 6$, represents the area of the gray rectangle between the data point and the two mean lines. (See Figure 4.2.1.)

For any data point (x,y) in Section 1, $x > \bar{x}$, so $(x - \bar{x})$ will be positive. Likewise, $y > \bar{y}$, so $(y - \bar{y})$ will be positive. Therefore, $(x - \bar{x})(y - \bar{y})$ will be positive. In the same way, you can determine the sign of $(x - \bar{x})(y - \bar{y})$ in each of the other sections. Fill in the blanks for Sections 2, 3, and 4 in Figure 4.1.5, just as we have already done for Section 1.

Section 2 $(x - \bar{x})$ _____ $(y - \bar{y})$ _____ $(x - \bar{x})(y - \bar{y})$ _____	Section 1 $(x - \bar{x})$ positive $(y - \bar{y})$ positive $(x - \bar{x})(y - \bar{y})$ positive
Section 3 $(x - \bar{x})$ _____ $(y - \bar{y})$ _____ $(x - \bar{x})(y - \bar{y})$ _____	Section 4 $(x - \bar{x})$ _____ $(y - \bar{y})$ _____ $(x - \bar{x})(y - \bar{y})$ _____

Figure 4.2.2 *The sign of* $(x - \bar{x})(y - \bar{y})$

You should have concluded that $(x-\bar{x})(y-\bar{y})$ is positive for data points in Section 1 and Section 3, and $(x-\bar{x})(y-\bar{y})$ is negative in Section 2 and Section 4. For each data point, $(x-\bar{x})(y-\bar{y})$ represents the area of the rectangle between the data point and the two mean lines, except that in Sections 2 and 4, this area is counted as negative. See Figure 4.2.3 below.

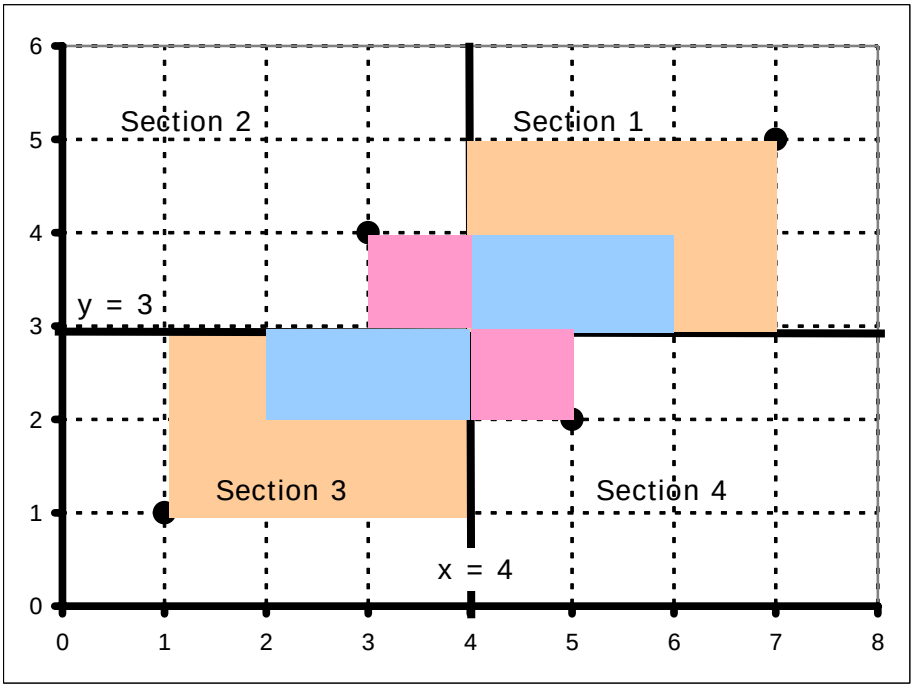


Figure 4.2.3 *Rectangle areas represented by* $(x-\bar{x})(y-\bar{y})$

In this case, there are two rectangles in Section 1 with areas 6 and 2 for a total area of 8. Likewise Section 3 has a total area of 8 while Sections 2 and 4 have a rectangle area of 1 each. The value of S_{xy} is:

$$S_{xy} = \sum(x-\bar{x})(y-\bar{y}) = 8+8+(-1)+(-1) = 14, \text{ a positive number.}$$

Whenever, as in this case, the correlation is positive, most of the points will tend to be in Sections 1 and 3, so most of the rectangle area will be counted as positive. Whenever there is positive correlation, S_{xy} will work out to be positive. Conversely, if the correlation is negative, most of the points will tend to be in Sections 2 and 4, so most of the rectangle area will be counted as negative. Hence, S_{xy} will work out to be negative whenever there is negative correlation.

S_{xy} might be sufficient to determine whether two variables have a positive or negative correlation, but S_{xy} alone is not a good measure of the strength of the relationship between the variables. There are two reasons for this:

1. The magnitude of S_{xy} is affected by how much spread (variance) x and y have. For example, if x was a distance, we would get different values for S_{xy}, depending on whether we measured x in feet or in inches. We would hesitate to say that converting x from feet to inches strengthens the relationship between x and y.

2. The magnitude of S_{xy} tends to increase as the number of data points increases. We wouldn't want to say that the relationship between two variables should be twice as strong if we collect twice as much data.

For these reasons, S_{xy} is divided by something so that the resulting value is not affected by the variances of x and y or by the number of data points. The result is known as *Person's correlation coefficient*.

Pearson's Correlation Coefficient: $$r = \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}}$$

Pearson's correlation coefficient r has the following properties:

1. The value of r is always between -1 and 1. If the data points lie exactly on a line with positive slope, then $r = 1$. If the data points lie exactly on a line with negative slope, then $r = -1$. In general, the closer the value of r is to 1 or -1, the stronger the linear relationship between the variables. The table in Figure 4.2.4 below might be helpful in classifying the strength of the relationship between two variables. We should note, however, that sample size also plays a role in interpreting the value of Pearson's correlation coefficient. An r-value of .9 computed from a sample of 3 data points does not indicate as strong a relationship between variables as it would for a sample of 300 data points.

$-1 < r < -.8$	$-.8 < r < -.5$	$-.5 < r < .5$	$.5 < r < .8$	$.8 < r < 1$
strong negative correlation	moderate negative correlation	weak or no correlation	moderate positive correlation	strong positive correlation

Figure 4.2.4 *Values of r and strength of correlation*

2. The value of r is not affected by changing the units that the variables are measured in. For example, changing measurements from feet to inches has no effect on the value of r.

3. The value of r is not affected by which variable is called x and which variable is called y.

The formulas for the three S's given above are most useful in interpreting what they represent, but for computing the values of the three S's from data, there are some other formulas which are usually easier.

Computing Formulas for the Three S's

$$S_{xx} = \sum x^2 - \frac{\left(\sum x\right)^2}{n}$$

$$S_{yy} = \sum y^2 - \frac{\left(\sum y\right)^2}{n}$$

$$S_{xy} = \sum xy - \frac{\left(\sum x\right)\left(\sum y\right)}{n}$$

Let's compute the value of Pearson's correlation coefficient both ways and compare the results. The table in Figure 4.2.5 below shows the raw data and some of the computations. The bottom row gives the total for each column. Remember from earlier that $\bar{x} = 4$ and $\bar{y} = 3$.

x	y	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$	x^2	y^2	xy
1	1	9	4	6	1	1	1
2	2	4	1	2	4	4	4
3	4	1	1	-1	9	16	12
5	2	1	1	-1	25	4	10
6	4	4	1	2	36	16	24
7	5	9	4	6	49	25	35
24	18	28	12	14	124	66	86

Figure 4.2.5 *Computations for Pearson's correlation coefficient*

Without going through all of the details, let's just highlight where a few of the numbers come from. The first number in the $(x-\bar{x})^2$ column uses $x = 1$ and $\bar{x} = 4$:

$$(x-\bar{x})^2 = (1-4)^2 = (-3)^2 = 9$$

The next number in that column uses the x-value for that row ($x = 2$):

$$(x-\bar{x})^2 = (2-4)^2 = (-2)^2 = 4$$

The total for that column, 28, is $\sum (x-\bar{x})^2$, which is S_{xx}. In the same way, 12, the total for the column headed $(y-\bar{y})^2$, is S_{yy}, and 14, the total for the $(x-\bar{x})(y-\bar{y})$ column, is S_{xy}.

Let's check that we get the same values by using the computing formulas for the S's.

$$S_{xx} = \sum x^2 - \frac{\left(\sum x\right)^2}{n} = 124 - \frac{24^2}{6} = 28$$

$$S_{yy} = \sum y^2 - \frac{\left(\sum y\right)^2}{n} = 66 - \frac{18^2}{6} = 12$$

$$S_{xy} = \sum xy - \frac{\left(\sum x\right)\left(\sum y\right)}{n} = 86 - \frac{(24)(18)}{6} = 14$$

As you can see, the results are the same. For those of you thinking that the original formulas are easier than the computing formulas remember two things. First, we've done most of the computations for you; it's different when you do them all yourself. Second, this data is highly rigged so that the numbers work out nice. Ordinarily, $\bar{x}$ and $\bar{y}$ are not nice whole numbers, but long, messy decimals. The computing formulas avoid the problem of plugging in these long, messy decimals for $\bar{x}$ and $\bar{y}$.

Finally, let's compute Pearson's coefficient:

$$r = \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}} = \frac{14}{\sqrt{28 \cdot 12}} = .764$$

This value indicates moderate (almost strong) positive correlation.

A Short Cut Using the TI-83 Calculator

You can use a calculator like the TI-83 Plus to get the values for $\sum x$, $\sum y$, etc. Hit STAT and choose 1: EDIT. Enter the *x*-values as L1 and the *y*-values as L2. Then hit STAT and then the right arrow key (to move to CALC). Notice that "2: 2-Var Stats" is on the list of options. Choose that option, and "2-Var Stats" appears on your screen. Hit enter and scroll down to see values for $\sum x$, $\sum x^2$, $\sum y$, $\sum y^2$, and $\sum xy$.

Note 1: The value S_x given by the calculator is not one of the three S's in this section, but rather the standard deviation of the *x*-values. We will discuss σ_x in Chapter 7.

Note 2: You'll see an even shorter short cut in the next section.

The first two columns of the table in Figure 4.2.6 below gives data and some computations from *Math in Context: Statistics and the Environment, Murre Island Bats* (p. 5). The data in the first column (*x*) is the air temperature in Celsius. The second column (*y*) is the number of minutes

that bats spend outside their caves. Some of the computations have been done for you, but there are some blank spaces left for you to fill in.

x	y	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$	x^2	y^2	xy
20	16	2.469	19.612	6.959	400	256	320
22	21	0.184	0.327	0.245	484	441	462
19	15	6.612	29.469	13.959	361	225	285
22	24	0.184	12.755	1.531	484	576	528
26	30	19.612	91.612	42.388	676	900	780
23	23	2.041	6.612	3.673	529	529	529
19	14				361	196	266
151	143	37.714	201.714	85.286	3295	3123	3170

Figure 4.2.6 *Data and computations for Murre Island Bats*

Focus on Understanding: Murre Island Bats

1. Compute the values of $\bar{x}$ and $\bar{y}$.

2. Compute the values that go in the three blank spaces in the table.

3. Compute the three S's using the original formulas.

4. Compute the three S's using the computing formulas. Check to see whether you got the same values as in (c).

5. Use the three S's to compute Pearson's Correlation Coefficient *r*. What does this value indicate about the relationship between temperature and the time that bats spend outside their caves? Draw a scatter plot to see if this seems reasonable.

6. What would happen to the value of *r* if the temperatures were given in Fahrenheit, rather than in Celsius?

7. In this case, do you think that there is a cause-and-effect relationship between *x* and *y*? Explain.