

## 8 Descriptive Statistics: Bivariate Data

*Dr. Frink:* Hey, Do you know there is a direct link between the decline in Spirograph and the rise in gang activity...Think about it!!!

From: *The Simpsons*

### 8.1 Scatter Plots

If each of a series of observation produces two measurements we say the collected data is bivariate. For example, suppose the height and weight are recorded for each person in a study. In this case we have continuous bivariate data since the values can in principle take on arbitrarily precise values. By contrast, if for each person in a survey we record the person's sex and a particular voting preference, we also have bivariate data. However, this data is categorical, since the possible values are restricted to only a few possibilities. We postpone consideration of such data until later in the course.

Construction of a scatter plot is the first step in understanding continuous bivariate data. To create a scatter plot we plot a point  $(x, y)$  for each observation, where the abscissa  $x$  is the first data value in the observation and the ordinate  $y$  is the second. Generally, the points are not connected to each other.

**Example 8.1:** Using Table 8.1 below prepare a scatter plot for the grades of individual students on the first and second hourly exams in a certain course.

Exam 1	Exam 2	Exam 1	Exam 2	Exam 1	Exam 2	Exam 1	Exam 2
55	54	41	47	53	30	42	82
75	85	76	23	51	18	43	21
95	85	71	61	45	58	53	74
45	56	81	88	65	33	73	60
63	58	72	83	13	20	86	83
70	23	<b>82</b>	<b>46</b>	74	83	49	57
47	78	88	98	30	33	75	59
40	44	49	52	92	100		
74	79	33	38	74	83		
72	75	75	97	56	74		

**Table 8.1**

*Solution:*

In the scatter plot the first exam grade for each student is plotted along the horizontal axis and the second grade along the vertical axis. For example, we have labeled the point with coordinates  $(82, 46)$ , which appears in bold italics in Table 8.1.

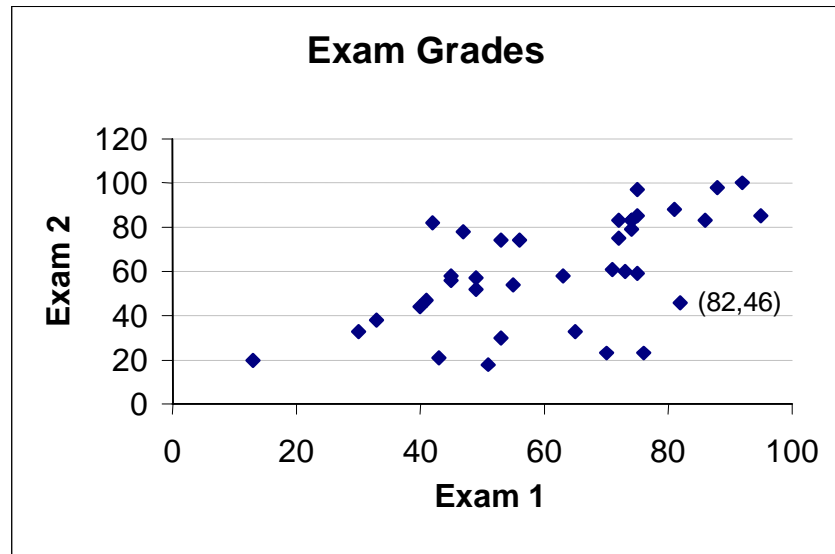


Figure 8.1

We are looking for a general pattern or trend in the plot. In this case there is a moderate trend upward and to the right. In other words, a high grade on the first exam tends to be associated with a high grade on the second exam, although this pattern is by no means universal. The pair (82, 46) certainly does not follow this trend. ■

## 8.2 The Correlation Coefficient

We could argue about the apparent trend in Figure 8.1 without arriving at a conclusive answer. We would like to find a quantitative measure to help us assess the degree of association between the two scores. We asserted that a high score on the first exam tended to be followed by high score on the second exam. What do we mean by “high score” or “low score”? A score is high if it is above the average score for that exam, low if it is below the average. Thus, a positive value of  $x - \bar{x}$  denotes a high score and a negative value a low score, similarly for  $y - \bar{y}$ .

We then consider for each data point  $(x, y)$  the product  $(x - \bar{x})(y - \bar{y})$ . If  $x$  is greater than  $\bar{x}$  and at the same time  $y$  is greater than  $\bar{y}$ , this product is positive. If  $x$  is a low score (below average) and the second grade  $y$  is also below average, the product is again positive. Therefore, whenever the  $x$  and  $y$  values follow the observed trend, the product  $(x - \bar{x})(y - \bar{y})$  is positive. If a high score on exam 1 is followed by a low grade on exam 2 (or vice-versa) the product  $(x - \bar{x})(y - \bar{y})$  is negative. It seems reasonable then to add these contributions from each data point and then average the result to obtain a numerical measure of the association between the two scores. This leads to

**Definition 8.1:** For a bivariate data set the *covariance*  $c_{xy}$  between the variables  $x$  and  $y$  is

$$c_{xy} = \frac{(x_1 - \bar{x})(y_1 - \bar{y}) + (x_2 - \bar{x})(y_2 - \bar{y}) + \cdots + (x_n - \bar{x})(y_n - \bar{y})}{n-1} \quad (8.1)$$

where  $n > 1$  is the number of data points in the set. ■

As in the definition of the standard deviation, we compute the average by dividing by  $n-1$ , rather than by  $n$ . The technical reason for this need not concern us. Computing the covariance by hand is rather tedious and error prone, as is the case for all the statistical quantities we will study in this chapter. We assume the reader has access to a computer or calculator in which these functions are available.

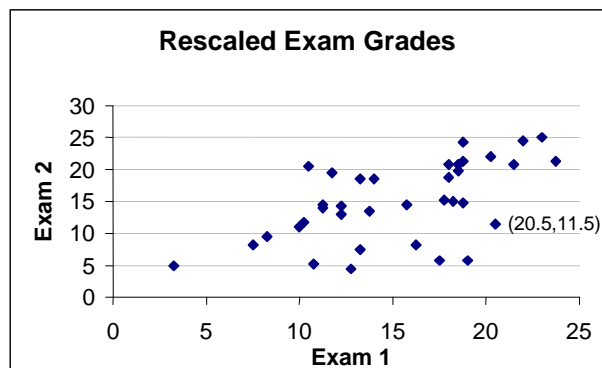
**Example 8.2:** Find the covariance of the exam grades in Table 8.1.

*Solution:*

To find the covariance for the exam grades in Example 8.1 we first must compute the averages for each exam. These are  $\bar{x} = 61.6$  and  $\bar{y} = 60.5$ . We then compute the various products  $(55 - 61.6)(54 - 60.5) = 42.9$ ,  $(75 - 61.6)(85 - 60.5) = 328.3$ , etc. These are summed and the total divided by  $n-1 = 36$ , giving for the covariance a value of 272.7. This is positive and seems to confirm our suspicion regarding the trend of the data. But is there any significance to the size of this number? ■

Suppose the instructor had scaled back each score to the range  $[0, 25]$ , because four exams were to be given and the final total score for the term should add up to a maximum of 100. How would this scaling affect the covariance? Since we are dividing each score by 4, the average on each exam would also be divided by 4. Every term in the numerator of the covariance would be divided by 16 and the new covariance would be 17.0.

However, the scatter plot of the scaled scores would look identical to Figure 8.1, except for a change in the scale on each axis.



**Figure 8.2**

Thus, a change of scale does not affect our perception of a trend in the scatter plot. Since the scaling does affect the numerical value (but not the sign) of the covariance, the numerical value cannot have any significance as a measure of this trend.

To obtain a useful numerical measure of a trend, we need to modify the covariance. We must compare the covariance with a quantity that measures the spread of each data variable.

**Definition 8.2:** The *correlation coefficient*  $r = \frac{c_{xy}}{s_x s_y}$ , where  $s_x$  and  $s_y$  denote the standard deviations of the individual data variables  $x$  and  $y$ . ■

Most statistical packages have a function that computes the correlation coefficient (also known as Pearson's  $r$ ), without the need for the user to compute separately the covariance and the two standard deviations.

**Example 8.3:** Compute the correlation coefficient for the data in Example 8.1.

*Solution:*

For the data in Example 8.1 we found that  $c_{xy} = 272.7$ . The two standard deviations are  $s_x = 19.1$  and  $s_y = 24.5$ . We thus obtain that  $r = \frac{272.7}{19.05 \times 24.47} = 0.585$ . Observe that this value is not affected by scaling the grade scores. As was noted in Chapter 7, Exercise 5, multiplying or dividing each value in a data set by a positive constant  $k$  multiplies or divides the standard deviation by the same number. Hence, the scaling in the numerator of  $r$  is canceled by exactly the same scaling in the denominator. Thus  $r$  is a true measure of trend. ■

In addition to being unaffected by multiplicative scaling, the correlation coefficient is unaffected by additive translation. Precisely,

**Correlation Property 8.1:**

- a) If a constant  $k$  is added to each  $x$  data value and a constant  $l$  is added to each  $y$  data value, then the new numbers have the same correlation coefficient as the old numbers.
- b) If each  $x$  value is multiplied by a positive constant  $k$  and each  $y$  value is multiplied by a positive constant  $l$ , then the new numbers have the same correlation coefficient as the old numbers.

*Proof:*

For a) note that when the same constant is added to each data value in a set the mean is shifted by the same amount. Thus, terms of the form  $x - \bar{x}$  do not change and so the covariance  $c_{xy}$  remains

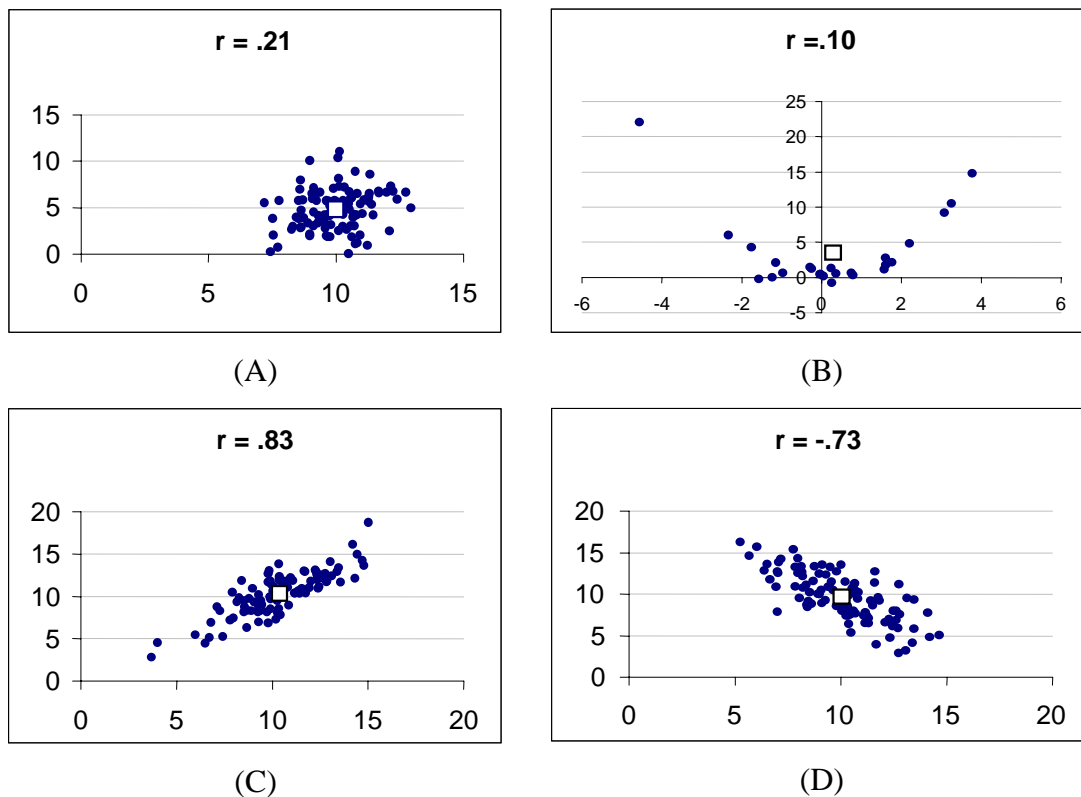
the same. Similarly the standard deviations  $s_x$  and  $s_y$  are unchanged and therefore, also the correlation coefficient. We have already discussed the invariance of  $r$  due to change in scale b).■

We remark that the correlation coefficient, unlike the covariance and standard deviation, has no units. It is a pure number. In fact, this number satisfies the following property:

**Correlation Property 8.2:** For any bivariate data, the correlation coefficient is a number in the interval  $[-1, 1]$ .■

We consider some further examples of data sets to get some understanding of the correlation coefficient. The scatter plots below exhibit computer-generated data with the corresponding value of  $r$ . For the reader who wishes not to take everything on faith, we present a simple proof of this statement in section 8.6.

**Example 8.4:** Describe the relationship between each of the following scatter plots and the corresponding values of  $r$ .



**Figure 8.3**

In each plot the point  $(\bar{x}, \bar{y})$  has been marked with a  $\square$ . The correlation coefficient measures the distribution of the data pairs around this point.

- Picture (A) illustrates typical bivariate data with low correlation coefficient. The data forms a cloud with a roughly circular shape around  $(\bar{x}, \bar{y})$  and no well-defined axis. There is no clear association between the size of  $x$  relative to  $\bar{x}$  and the size of  $y$  relative to  $\bar{y}$ .
- In (B) we see a data set that also has a low value of the correlation coefficient. In this case the data does exhibit an obvious pattern, but it is not linear. ***This brings out the important point that the correlation coefficient only measures the strength of a linear relation between the variables.*** A value of  $r$  close to zero does not rule out that non-linear trends may exist. Once again, this emphasizes the importance of plotting the data and not simply relying on a single numerical value.
- Pictures (C) and (D) illustrate data with strong positive and negative correlation. The data lie along a well-defined linear axis that passes through or near the point  $(\bar{x}, \bar{y})$ . The negative value of  $r$  in (D) implies that values of  $x$  larger than  $\bar{x}$  are associated with values of  $y$  smaller than  $\bar{y}$ . ■

We summarize the discussion in

**Correlation Property 8.3:**

- $r$  close to +1  $\Leftrightarrow$  data shows a strong positive linear correlation (large  $x$  values go with large  $y$ ).
- $r$  close to -1  $\Leftrightarrow$  data shows a strong negative linear correlation (large  $x$  values go with small  $y$ ).
- $r$  close to 0  $\Leftrightarrow$  data shows no or weak linear correlation. Other non-linear trends may be possible. ■

These should be taken as approximate guidelines. Without a plot of the data you can be misled by only the value of  $r$ . For some justification of Correlation Property 8.3 see section 8.6.

### 8.3 Association and Causality

Depending on the nature of the variables, there is a tendency to interpret a high correlation between two variables as indicative of some causal connection between them. This may be correct, but it is not established by the statistical correlation. In fact, no causal connection can be established by statistics alone. A causal connection requires the support of an underlying theory with an established experimentally verified body of knowledge. Unfortunately, for many problems that interest us, particularly in ecology, medicine, and epidemiology, the underlying scientific principles are not sufficiently advanced to enable us to evaluate from basic theory many of the associations suggested by statistical investigations.

Lacking a clear scientific basis with which to evaluate a statistical association, the prudent investigator will want to rule out other possibilities that might account for the observed

association. One such is a so-called hidden variable. For example, you will find a strong positive correlation between the median price of a new home and the rate of property crimes during the ten-year period from 1983 to 1992. (See exercise 13.) Maybe people were committing more crimes so they could afford those more expensive homes. Sounds silly, but a great many positive associations are dressed up in this way to a cause and effect relation. Of course, here there is a hidden variable, time. Over a course of a number of years many sociological and economic indices will exhibit a steady directional pattern, perhaps by random, perhaps due to some underlying reason. If we plot two such variables together we will find a correlation. Most such correlations are meaningless.

Conversely, a lack of statistical correlation does not mean there is no association between the variables. We have already seen an example in Figure 8.3(B). More seriously, a causal connection may exist, but the subjects used in the data collection may be affected by other variables that, by adding to the variability, mask the true association.


#### 8.4 Tech Notes

We give directions for constructing a scatter plot using *Excel*. The steps are similar to those used to create a graph of a function, as described in lesson 3 of *tutorial.xls*. To describe the method we will go through a specific example.

**Example 8.5:** Using the file *airfares.xls*, construct a scatter plot of airfare versus flight distance in miles.

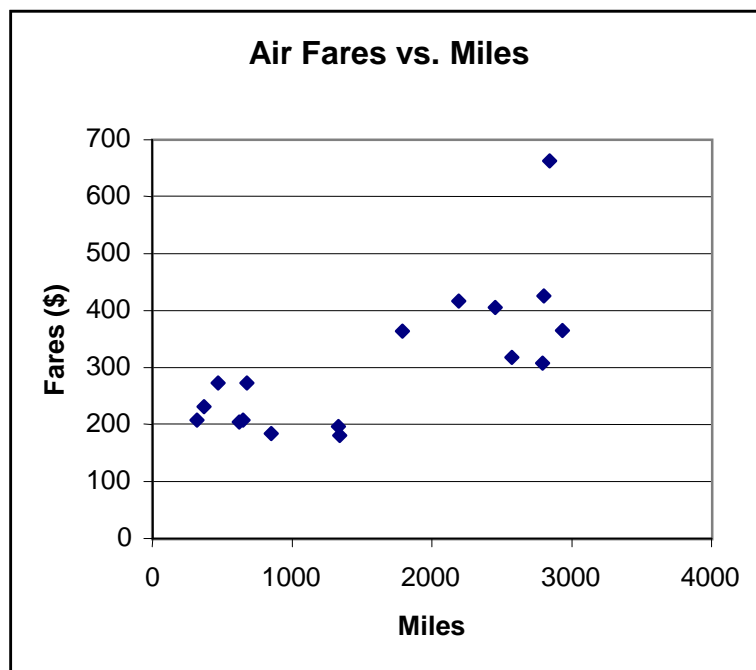
*Solution:*

Open the file and use these instructions to construct the plot.

1. We will use the Chart Wizard. This is activated from the toolbar using a button that shows a bar chart (). Click on this button to select the Wizard. The wizard opens with the first of 4 steps.
2. Wizard Step 1: **Select Chart Type.** Click on the icon for the  $x, y$  scatter plot [XY (Scatter)] and the sub-type showing non-connected points. (This is the default choice.) Click the *Next* button to continue.
3. Wizard Step 2: **Select Chart Source Data.** In the Data Range box you want to enter b7:c23 (or \$b\$7:\$c\$23). You can do this by first clearing the box and then using the mouse to select the range on the worksheet or you can type the range. When you type the range, you can omit the \$ signs, but do not leave any spaces when you type b7:c23. The option *Series in* should have *columns* selected, since our data are arranged in columns. Click on the *Next* button to proceed. You can also go back to modify previously made choices.

4. Wizard Step 3: **Chart Options.** Add a title by typing *Airfares vs. Miles* in the title box. Add titles to the axes by typing for Value (X), *Miles* and typing for Value (Y), *Fares(\$)*. A legend is not needed here since the plot shows only one series. Eliminate the legend by clicking the tab marked *Legend* and then deleting the check from the option *Show Legend*. Click *Next* to continue.
5. Wizard Step 4: **Chart Location.** Select the second option, *as object in...*. This places the chart in the current worksheet. Select the *Finish* button if you are satisfied with the appearance of the graph as shown by the Wizard or the *Back* button if you want to make additional changes. You can also modify the graph once it has been placed into the spreadsheet.

Your graph should look like the figure below.



**Figure 8.4**

■

**Example 8.6:** Use *Excel* to find the correlation coefficient for the data plotted in Example 8.5.


*Solution:*

To compute the *correlation coefficient* for the data, pick a blank cell and click the mouse on this cell. Type `=correl(b7:b23,c7:c23)` and press Enter. The result 0.7328 should appear in the box. To compute the *covariance* you can enter the expression `=covar(b7:b23, c7:c23)`. Remember to place a label in a box adjacent to the computed result to remind you what number appears in the cell. ■





*As a technical aside, note that Excel uses a slightly different expression for the covariance than our equation (8.1). Namely, instead of dividing the numerator of (8.1) by  $n-1$ , the formula used by Excel has a denominator of  $n$ . Thus, multiplying Excel's covariance by  $n/(n-1)$  will yield the value of the covariance used in this chapter. Excel's correlation coefficient coincides in value with that produced by Definition 8.2.*

It is very easy to forget the abbreviation and syntax for the many functions available in *Excel*. The function wizard, activated using the icon, , can guide you through this process. The *Excel* function collection is organized by topic and then alphabetically within topic. The initial topic category is a list of the most recently used functions. After selecting the wizard, choose the list of Statistical functions and let the wizard guide you through the steps needed to correctly enter the commands for the correlation coefficient and covariance.

## 8.5 Summary

In this chapter we have considered bivariate data obtained when two numerical values are associated with each observation in a data set. With such data, one is primarily interested in relationships between the two values. A first step in such an analysis is the construction of a *scatter plot* in which we plot each pair of values  $(x_i, y_i)$  as a point in the plane. When the scatter plot shows a linear trend in the data we can measure the strength of that association with the *correlation coefficient*,  $r$ . The latter is a number satisfying  $-1 \leq r \leq 1$ . Values of  $r$  that are close to  $\pm 1$  indicate a strong linear correlation. It is important to bear in mind that a strong correlation between two variables does not necessarily indicate a causal relationship, although further investigation may ultimately reveal the truth of such an explanation.

## 8.6 Mathematical Excursions

We have stated some mathematical properties of the correlation coefficient that are certainly not obvious, though we have confirmed them in several numerical examples. For the more adventurous reader we provide some additional details regarding the proof of Correlation Property 8.2 and Correlation Property 8.3.

The key to Correlation Property 8.2 is the following algebraic lemma:

**Lemma 8.1:** Suppose  $a_1, a_2, \dots, a_n$  satisfies  $a_1^2 + a_2^2 + \dots + a_n^2 = 1$  and  $b_1, b_2, \dots, b_n$  also satisfies  $b_1^2 + b_2^2 + \dots + b_n^2 = 1$ , then  $|a_1 b_1 + a_2 b_2 + \dots + a_n b_n| \leq 1$ . Equality holds only when either  $a_1 = b_1, a_2 = b_2, \dots, a_n = b_n$  or  $a_1 = -b_1, a_2 = -b_2, \dots, a_n = -b_n$ .

*Proof:*

A sum of squares of real numbers is always non-negative. Therefore,

$$(a_1 - b_1)^2 + (a_2 - b_2)^2 + \cdots + (a_n - b_n)^2 \geq 0 \quad (8.2)$$

with equality only when  $a_1 = b_1, a_2 = b_2, \dots, a_n = b_n$ . Now expand each term on the left side of (8.2). After rearranging terms we obtain,

$$(a_1^2 + a_2^2 + \cdots + a_n^2) + (b_1^2 + b_2^2 + \cdots + b_n^2) - 2(a_1b_1 + a_2b_2 + \cdots + a_nb_n) \geq 0.$$

Using that  $a_1^2 + a_2^2 + \cdots + a_n^2 = 1$  and  $b_1^2 + b_2^2 + \cdots + b_n^2 = 1$ , this yields

$$2 \geq 2(a_1b_1 + a_2b_2 + \cdots + a_nb_n),$$

or, dividing by 2,  $a_1b_1 + a_2b_2 + \cdots + a_nb_n \leq 1$ . Note that equality holds only when there is equality in (8.2), so that  $a_1 = b_1, a_2 = b_2, \dots, a_n = b_n$ . This establishes half of the desired inequality. We have also to show that  $-1 \leq a_1b_1 + a_2b_2 + \cdots + a_nb_n$ , with equality only when  $a_1 = -b_1, a_2 = -b_2, \dots, a_n = -b_n$ . The reader can show this by working with the inequality

$$(a_1 + b_1)^2 + (a_2 + b_2)^2 + \cdots + (a_n + b_n)^2 \geq 0$$

instead of (8.2). ■

Correlation Property 8.2 follows easily from the previous lemma.

*Proof of Correlation Property 8.2:*

Recall the definition of the correlation coefficient:

$$r = \frac{c_{xy}}{s_x s_y} = \frac{1}{n-1} \left( \frac{(x_1 - \bar{x})(y_1 - \bar{y}) + (x_2 - \bar{x})(y_2 - \bar{y}) + \cdots + (x_n - \bar{x})(y_n - \bar{y})}{s_x s_y} \right).$$

Now we break up the right side above into individual terms of the form  $\left( \frac{x_i - \bar{x}}{s_x \sqrt{n-1}} \right) \left( \frac{y_i - \bar{y}}{s_y \sqrt{n-1}} \right)$ .

This gives

$$r = \left( \frac{x_1 - \bar{x}}{s_x \sqrt{n-1}} \right) \left( \frac{y_1 - \bar{y}}{s_y \sqrt{n-1}} \right) + \left( \frac{x_2 - \bar{x}}{s_x \sqrt{n-1}} \right) \left( \frac{y_2 - \bar{y}}{s_y \sqrt{n-1}} \right) + \cdots + \left( \frac{x_n - \bar{x}}{s_x \sqrt{n-1}} \right) \left( \frac{y_n - \bar{y}}{s_y \sqrt{n-1}} \right). \quad (8.3)$$

Let

$$a_1 = \frac{x_1 - \bar{x}}{s_x \sqrt{n-1}}, a_2 = \frac{x_2 - \bar{x}}{s_x \sqrt{n-1}}, \dots, a_n = \frac{x_n - \bar{x}}{s_x \sqrt{n-1}},$$

and similarly let

$$b_1 = \frac{y_1 - \bar{y}}{s_y \sqrt{n-1}}, b_2 = \frac{y_2 - \bar{y}}{s_y \sqrt{n-1}}, \dots, b_n = \frac{y_n - \bar{y}}{s_y \sqrt{n-1}}.$$

Then from (8.3) we have  $r = a_1 b_1 + a_2 b_2 + \dots + a_n b_n$ . The conclusion that  $r$  lies in the interval  $[-1, 1]$  will follow immediately from Lemma 8.1, if we show that the numbers  $a_1, a_2, \dots, a_n$  and  $b_1, b_2, \dots, b_n$  satisfy the conditions  $a_1^2 + a_2^2 + \dots + a_n^2 = 1$  and  $b_1^2 + b_2^2 + \dots + b_n^2 = 1$ . Indeed,

$$\begin{aligned} a_1^2 + a_2^2 + \dots + a_n^2 &= \frac{(x_1 - \bar{x})^2}{s_x^2(n-1)} + \frac{(x_2 - \bar{x})^2}{s_x^2(n-1)} + \dots + \frac{(x_n - \bar{x})^2}{s_x^2(n-1)} \\ &= \frac{1}{s_x^2} \left( \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n-1} \right) = 1, \end{aligned}$$

since by Definition 7.6,  $s_x^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n-1}$ . A similar argument applies to the sum  $b_1^2 + b_2^2 + \dots + b_n^2$ . ■

We can also use Lemma 8.1 to establish the extreme cases mentioned in Correlation Property 8.3. Namely when  $r = \pm 1$ , we show that the data values fall precisely on a straight line. It is then plausible that when  $r$  is close to either of these values the data will show a nearly linear trend.

*Partial proof of Correlation Property 8.3:*

Suppose  $r = 1$ . Using the quantities  $a_i$  and  $b_i$  defined above, we have that  $1 = r = a_1 b_1 + a_2 b_2 + \dots + a_n b_n$ . Consequently, equality holds in Lemma 8.1 so that for all  $i$ ,  $a_i = b_i$ . This means, however, that

$$\frac{x_i - \bar{x}}{s_x \sqrt{n-1}} = \frac{y_i - \bar{y}}{s_y \sqrt{n-1}}$$

or

$$\frac{x_i - \bar{x}}{s_x} = \frac{y_i - \bar{y}}{s_y}.$$

The latter equation implies that every data point lies on the line whose equation is

$$\frac{x - \bar{x}}{s_x} = \frac{y - \bar{y}}{s_y} \quad \text{or} \quad y = \frac{s_y}{s_x}(x - \bar{x}) + \bar{y}.$$

A similar argument works if  $r = -1$ . ■

### 8.7 Exercises

1. Consider the data in the following table:

$x$	1	3	-2	1	4	2	0
$y$	-2	3	0	1	-1	1	-3

- a) Using a calculator or computer find  $s_x$ ,  $s_y$ ,  $c_{xy}$ . According to *Excel* the value of  $r$  is approximately 0.30. Verify this.
- b) Using the result in a) and Correlation Property 8.1 what will be the value of  $s_x$ ,  $s_y$ ,  $c_{xy}$ , and  $r$  for the following data:

$x$	3	5	0	3	6	4	2
$y$	-1	4	1	2	0	2	-2

- c) Using the result in a) and Correlation Property 8.1 what will be the value of  $s_x$ ,  $s_y$ ,  $c_{xy}$ , and  $r$  for the following data:

$x$	2	6	-4	2	8	4	0
$y$	-6	9	0	3	-3	3	-9

2. The figure below shows a scatter plot for a set of 25 data values. The correlation coefficient is  $-0.98$ .

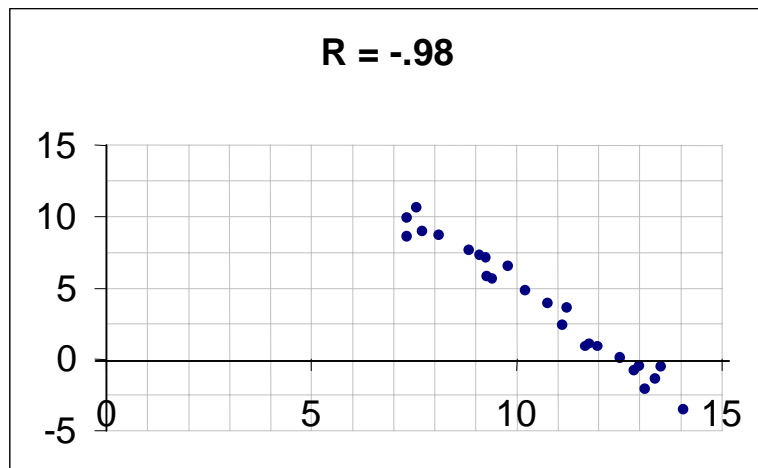
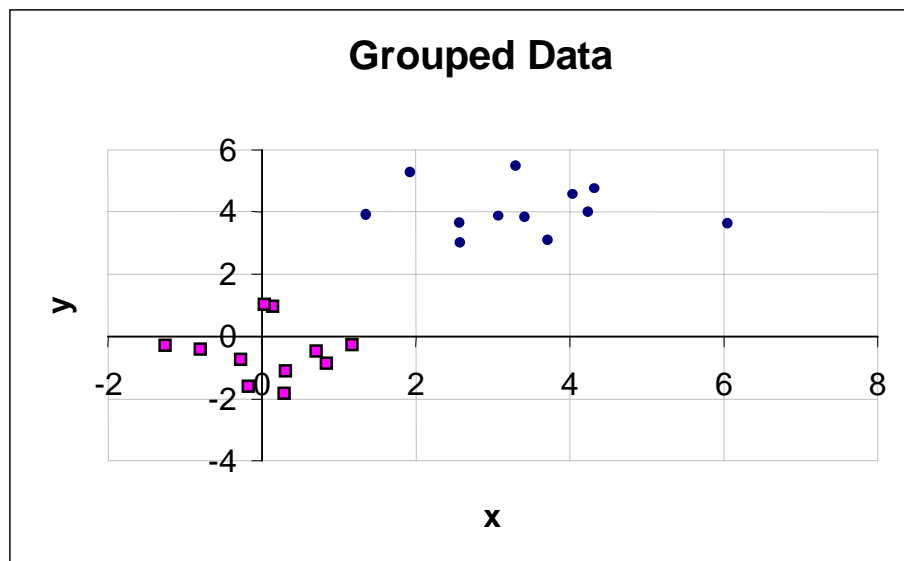


Figure 8.5

- a) The value of  $\bar{x} = 10.6$  and  $\bar{y} = 3.9$ . Mark the approximate location of the point  $(\bar{x}, \bar{y})$  on the graph. How many data values have both  $x$  and  $y$  greater than their respective means or both less than these means? How is this related to the value of  $r$ ?
- b) Estimate the range of  $x$  and the range of  $y$  from the graph.

## 8 Bivariate Data

- c) Can we use the Range Rule from Chapter 7 and the answer in b) to estimate the standard deviation for each data set? State reasons (The true values are  $s_x = 2.1$  and  $s_y = 4.2$ ).
3. Repeat the questions in exercise 2 for the grade data in Example 8.1. The values of  $\bar{x} = 61.6$ ,  $s_x = 19.1$ ,  $\bar{y} = 60.5$ , and  $s_y = 24.5$ .
4. The height and weight are recorded for a random group of adults. Which of the following would you expect for the correlation coefficient? Justify your choice.
- i) near +1      ii) moderately positive      iii) close to zero      iv) moderately negative      v) near -1
5. You purchase a new car and keep it for 10 years. Every year you record the amount you spend on repairs, including routine maintenance. Which of the following would you expect for the correlation coefficient? Justify your choice.
- i) near +1      ii) moderately positive      iii) close to zero      iv) moderately negative      v) near -1
6. 50 students take a multiple-choice exam with 10 questions. The number of correct and incorrect answers is recorded for each student. Which of the following values would you consider most likely for the correlation coefficient? Justify your choice.
- i) +1      ii) 0.5      iii) 0      iv) -0.5      v) -1
7. The figure below shows a data set which is comprised of two subgroups (one plotted with a circle the other with a square). These subgroups might represent say male and female subjects in a study of a pair of numerical characteristics.



**Figure 8.6**

## 8 Bivariate Data

- a) Given that the means for the individual groups are: Squares:  $\bar{x} \approx 0$ ,  $\bar{y} \approx -0.5$ ,  $n = 11$ ; Circles:  $\bar{x} \approx 3.4$ ,  $\bar{y} \approx 4.1$ ,  $n = 12$ ; plot the pair  $(\bar{x}, \bar{y})$  for each group. What value do you expect for the correlation coefficient for each group? Justify your answer.
- b) Using the information in a) compute the value of  $\bar{x}$  and  $\bar{y}$  if the data are combined. Mark the point on the graph and assess the size of the correlation coefficient for the combined data.
- c) Is it a good idea to use  $r$  to describe the strength of the association for the entire data set? Explain your answer.



8. Open the file *acorn.xls*. Prepare a scatter plot of tree height vs. acorn size and compute the correlation coefficient? How strong is the association between these variables?



9. Open the file *emission.xls*.

- a) Prepare a scatter plot of CO emissions vs. hydrocarbon emissions and find the correlation coefficient. How strong is the association between these variables?
- b) Prepare a scatter plot of NO emissions vs. hydrocarbon emissions and find the correlation coefficient. How strong is the association between these variables?
- c) What relevance do parts a) and b) have to the setting of emissions standards for engines of the type tested?



10. Open the file *sat.xls*. Prepare a scatter plot of the average SAT score vs. the percent of students taking the test. Find the correlation coefficient. How strong is the association? What might explain this association?



11. Open the file *auto.xls*. Prepare a scatter plot of auto mileage vs. auto weight for the given models. Find the correlation coefficient. How would you assess the strength of the association between these variables? What is the physical explanation for this association?



12. Open the file *cities.xls*.

- a) Prepare a scatter plot of # of Law enforcement officers vs. Population size. Find the correlation coefficient. Does there appear to be a strong association between the two variables?
- b) Copy the data to the blank sheet1 and then eliminate all cities with populations greater than 1.5 million from the data. Prepare the scatter plot as in part a) for the remaining cities and compute the correlation coefficient. Does the association appear as pronounced as it did in a)?

Remark: The points eliminated in b) are known as *high leverage points*. They have  $x$  values far from the mean (789,000) and exert a heavy influence on the value of  $r$ , perhaps exaggerating the linear association that holds for the bulk of the data values. To reiterate, always plot (and replot) the data!

## 8 Bivariate Data



13. Open the file *crime&prices.xls*. Prepare a scatter plot of crime rate vs. housing prices and determine the correlation coefficient.



14. Open the file *ozone.xls*. Prepare a scatter plot of the customer's ozone report ( $y$ ) and the level determined by the NIST monitor ( $x$ ). What is the correlation coefficient for these two variables?



15. Open the file *hibernat.xls*. Prepare a scatter plot of the age at death vs. % of lifetime in hibernation. Assess the degree of linear association between these variables.