24. If your software includes access to a computer algebra system (CAS), use it as follows: Let $f(\mathbf{u}) = A\mathbf{u}$ be the matrix transformation defined by

$$A = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}.$$

(a) Find the (symbolic) matrix that defines the matrix transformation $f(f(f(f(\mathbf{u}))))$.

(b) Use CAS commands to simplify the matrix obtained in part (a) so that you obtain the identities for $\sin(4\theta)$ and $\cos(4\theta)$.

1.8 Correlation Coefficient (Optional)

As we noted in Section 1.2, we can use an n-vector to provide a listing of data. In this section we provide a statistical application of the dot product to measure the strength of a linear relationship between two data vectors.

Before presenting this application, we must note two additional properties that vectors possess: length (also known as magnitude) and direction. These notions will be carefully developed in Chapter 4; in this section we merely give the properties without justification.

The length of the n-vector

$$\mathbf{v} = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_{n-1} \\ v_n \end{bmatrix},$$

denoted as $\|\mathbf{v}\|$, is defined as

$$\|\mathbf{v}\| = \sqrt{v_1^2 + v_2^2 + \cdots + v_{n-1}^2 + v_n^2}. \tag{1}$$

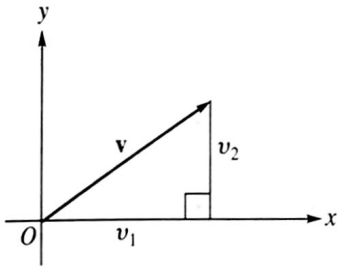


FIGURE 1.21 Length of $\mathbf{v}$.

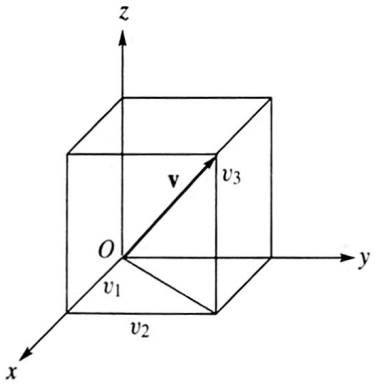


FIGURE 1.22 Length of $\mathbf{v}$.

If $n = 2$, the definition given in Equation (1) can be established easily as follows: From Figure 1.21 we see by the Pythagorean theorem that the length of the directed line segment from the origin to the point (v_1, v_2) is $\sqrt{v_1^2 + v_2^2}$. Since this directed line segment represents the vector $\mathbf{v} = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$, we agree that $\|\mathbf{v}\|$, the length of the vector $\mathbf{v}$, is the length of the directed line segment. If $n = 3$, a similar proof can be given by applying the Pythagorean theorem twice in Figure 1.22.

It is easiest to determine the direction of an n-vector by defining the angle between two vectors. In Sections 5.1 and 5.4, we define the angle θ between the nonzero vectors $\mathbf{u}$ and $\mathbf{v}$ as the angle determined by the expression

$$\cos\theta = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\| \, \|\mathbf{v}\|}.$$

In those sections we show that

$$-1 \leq \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\| \, \|\mathbf{v}\|} \leq 1.$$

Hence, this quantity can be viewed as the cosine of an angle $0 \leq \theta \leq \pi$.

We now turn to our statistical application. We compare two data n-vectors $\mathbf{x}$ and $\mathbf{y}$ by examining the angle θ between the vectors. The closeness of $\cos\theta$ to -1 or 1 measures how near the two vectors are to being parallel, since the angle between parallel vectors is either 0 or π radians. Nearly parallel indicates a strong relationship between the vectors. The smaller $|\cos\theta|$ is, the less likely it is that the vectors are parallel, and hence the weaker the relationship is between the vectors.

Table 1.1 contains data about the ten largest U.S. corporations, ranked by market value for 2004. In addition, we have included the corporate revenue for 2004. All figures are in billions of dollars and have been rounded to the nearest billion.

TABLE 1.1

Corporation	*Market Value (in $ billions)*	*Revenue (in $ billions)*
General Electric Corp.	329	152
Microsoft	287	37
Pfizer	285	53
Exxon Mobile	277	271
Citigroup	255	108
Wal-Mart Stores	244	288
Intel	197	34
American International Group	195	99
IBM Corp.	172	96
Johnson & Johnson	161	47

Source: *Time Almanac 2006, Information Please*, Pearson Education, Boston, Mass., 2005; and http://www.geohive.com/charts.

To display the data in Table 1.1 graphically, we form ordered pairs, (market value, revenue), for each of the corporations and plot this set of ordered pairs. The display in Figure 1.23 is called a **scatter plot**. This display shows that the data are spread out more vertically than horizontally. Hence there is wider variability in the revenue than in the market value.

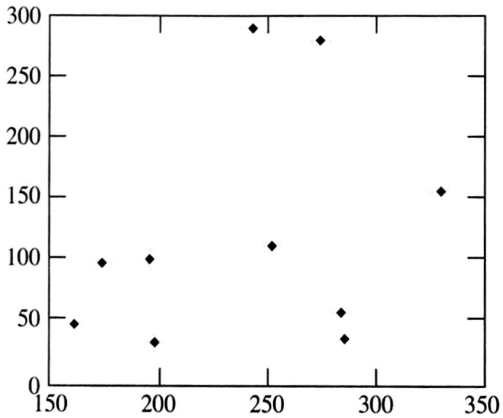


FIGURE 1.23

If we are interested only in how individual values of market value and revenue go together, then we can rigidly translate (shift) the plot so that the pattern of points does not change. One translation that is commonly used is to move the center of the plot to the origin. (If we think of the dots representing the ordered pairs as weights, we see that this amounts to shifting the center of mass to the origin.) To perform this translation, we compute the mean of the market value observations and subtract it from each market value; similarly, we compute the mean of the revenue observations and subtract it from each revenue value. We have (rounded to a whole number)

$$\text{Mean of market values} = 240, \quad \text{mean of revenues} = 119.$$

Subtracting the mean from each observation is called **centering the data**, and the corresponding centered data are displayed in Table 1.2. The corresponding scatter plot of the centered data is shown in Figure 1.24.

TABLE 1.2

Centered Market Value (in $ billions)	*Centered Revenue (in $ billions)*
89	33
47	−82
45	−66
37	152
15	−11
4	169
−43	−85
−45	−20
−68	−23
−79	−72

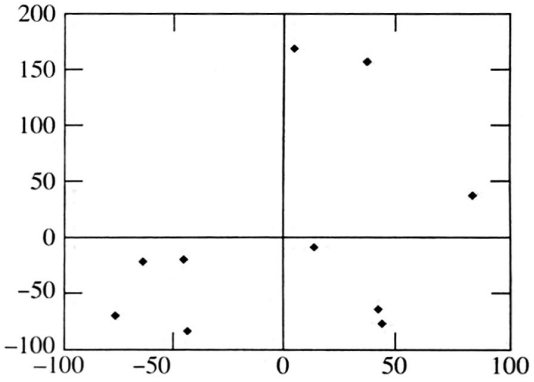


FIGURE 1.24

Note that the arrangement of dots in Figures 1.23 and 1.24 is the same; the scales of the respective axes have changed.

A scatter plot places emphasis on the observed data, not on the variables involved as general entities. What we want is a new way to plot the information that focuses on the variables. Here the variables involved are market value and revenue, so we want one axis for each corporation. This leads to a plot with ten axes, which we are unable to draw on paper. However, we visualize this situation by considering 10-vectors, that is, vectors with ten components, one for each corporation. Thus we define a vector $\mathbf{v}$ as the vector of centered market values and a

vector **w** as the vector of centered revenues:

$$\mathbf{v} = \begin{bmatrix} 89 \\ 47 \\ 45 \\ 37 \\ 15 \\ 4 \\ -43 \\ -45 \\ -68 \\ -79 \end{bmatrix}, \quad \mathbf{w} = \begin{bmatrix} 33 \\ -82 \\ -66 \\ 152 \\ -11 \\ 169 \\ -85 \\ -20 \\ -23 \\ -72 \end{bmatrix}.$$

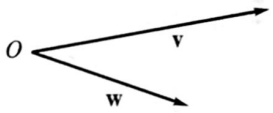


FIGURE 1.25

The best we can do schematically is to imagine **v** and **w** as directed line segments emanating from the origin, which is denoted by **0** (Figure 1.25).

The representation of the centered information by vectors, as in Figure 1.25, is called a **vector plot**. From statistics, we have the following conventions:

- In a vector plot, the length of a vector indicates the variability of the corresponding variable.
- In a vector plot, the angle between vectors measures how similar the variables are to each other.

The statistical terminology for "how similar the variables are" is "how highly correlated the variables are." Vectors that represent highly correlated variables have either a small angle or an angle close to π radians between them. Vectors that represent uncorrelated variables are nearly perpendicular; that is, the angle between them is near $\pi/2$.

The following chart summarizes the statistical terminology applied to the geometric characteristics of vectors in a vector plot.

Geometric Characteristics	*Statistical Interpretation*
Length of a vector.	Variability of the variable represented.
Angle between a pair of vectors is small.	The variables represented by the vectors are highly positively correlated.
Angle between a pair of vectors is near π.	The variables represented by the vectors are highly negatively correlated.
Angle between a pair of vectors is near $\pi/2$.	The variables represented by the vectors are uncorrelated or unrelated. The variables are said to be perpendicular or orthogonal.

From statistics we have the following measures of a sample of data $\{x_1, x_2, \ldots, x_{n-1}, x_n\}$:

Sample size $= n$, the number of data.

Sample mean $= \bar{x} = \dfrac{\sum_{i=1}^{n} x_i}{n}$, the average of the data.

Correlation coefficient: If the n-vectors $\mathbf{x}$ and $\mathbf{y}$ are data vectors where the data have been centered, then the correlation coefficient, denoted by $Cor(\mathbf{x}, \mathbf{y})$, is computed by

$$Cor(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \, \|\mathbf{y}\|}.$$

Geometrically, $Cor(\mathbf{x}, \mathbf{y})$ is the cosine of the angle between vectors $\mathbf{x}$ and $\mathbf{y}$.

For the centered data in Table 1.2, the sample size is $n = 10$, the mean of the market value variable is 240, and the mean of the revenue variable is 119. To determine the correlation coefficient for $\mathbf{v}$ and $\mathbf{w}$, we compute

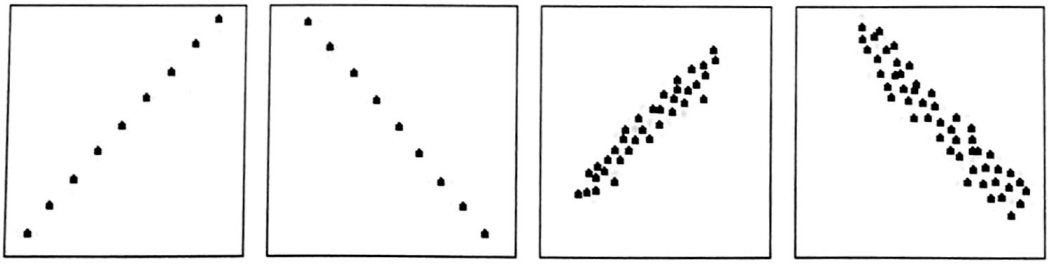
$$Cor(\mathbf{v}, \mathbf{w}) = \cos\theta = \frac{\mathbf{v} \cdot \mathbf{w}}{\|\mathbf{v}\| \, \|\mathbf{w}\|} = 0.2994,$$

and thus

$$\theta = \arccos(0.2994) = 1.2667 \text{ radians} \approx 72.6°.$$

This result indicates that the variables market value and revenue are not highly correlated. This seems to be reasonable, given the physical meaning of the variables from a financial point of view. Including more than the ten top corporations may provide a better measure of the correlation between market value and revenue. Another approach that can be investigated based on the scatter plots is to omit data that seem far from the grouping of the majority of the data. Such data are termed **outliers**, and this approach has validity for certain types of statistical studies.

Figure 1.26 shows scatter plots that geometrically illustrate various cases for the value of the correlation coefficient. This emphasizes that the correlation coefficient is a measure of linear relationship between a pair of data vectors $\mathbf{x}$ and $\mathbf{y}$. The closer all the data points are to the line (in other words, the less scatter), the higher the correlation between the data.



(a) Perfect positive correlation; $Cor\,(\mathbf{x}, \mathbf{y}) = 1$.

(b) Perfect negative correlation; $Cor\,(\mathbf{x}, \mathbf{y}) = -1$.

(c) Less than perfect positive correlation; $0 \leq Cor\,(\mathbf{x}, \mathbf{y}) \leq 1$.

(d) Less than perfect negative correlation; $-1 \leq Cor\,(\mathbf{x}, \mathbf{y}) \leq 0$.

FIGURE 1.26

To compute the correlation coefficient of a set of ordered pairs (x_i, y_i), $i = 1, 2, \ldots, n$, where

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{n-1} \\ x_n \end{bmatrix} \quad \text{and} \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{n-1} \\ y_n \end{bmatrix},$$

we use the steps in Table 1.3. The computational procedure in Table 1.3 is called the **Pearson product–moment correlation coefficient** in statistics.

TABLE 1.3

1. Compute the sample means for each data vector:

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}, \qquad \overline{y} = \frac{\sum_{i=1}^{n} y_i}{n}.$$

2. Determine the centered x-data and the centered y-data as the vectors $\mathbf{x}_c$ and $\mathbf{y}_c$, respectively, where

$$\mathbf{x}_c = \begin{bmatrix} x_1 - \overline{x} & x_2 - \overline{x} & \cdots & x_n - \overline{x} \end{bmatrix}^T$$
$$\mathbf{y}_c = \begin{bmatrix} y_1 - \overline{y} & y_2 - \overline{y} & \cdots & y_n - \overline{y} \end{bmatrix}^T.$$

3. Compute the correlation coefficient as

$$Cor(\mathbf{x}_c, \mathbf{y}_c) = \frac{\mathbf{x}_c \cdot \mathbf{y}_c}{\|\mathbf{x}_c\| \, \|\mathbf{y}_c\|}.$$

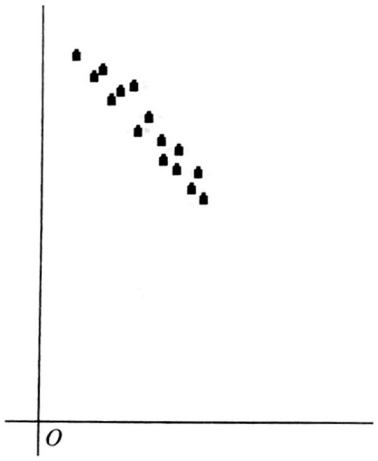


FIGURE 1.27

The correlation coefficient can be an informative statistic in a wide variety of applications. However, care must be exercised in interpreting this numerical estimation of a relationship between data. As with many applied situations, the interrelationships can be much more complicated than they are perceived to be at first glance. Be warned that statistics can be misused by confusing relationship with cause and effect. Here we have provided a look at a computation commonly used in statistical studies that employ dot products and the length of vectors. A much more detailed study is needed to use the correlation coefficient as part of a set of information for hypothesis testing. We emphasize such warnings with the following discussion, adapted from *Misused Statistics*, by A. J. Jaffe and H. F. Spirer (Marcel Dekker, Inc., New York, 1987):

> Data involving divorce rate per 1000 of population versus death rate per 1000 of population were collected from cities in a certain region. Figure 1.27 shows a scatter plot of the data. This diagram suggests that

divorce rate is highly (negatively) correlated with death rate. Based on this measure of relationship, should we infer that

(i) divorces cause death?
(ii) reducing the divorce rate will reduce the death rate?

Certainly we have not proved any cause-and-effect relationship; hence we must be careful to guard against statements based solely on a numerical measure of relationship.

Key Terms

Dot product
Length of a vector
Direction of a vector
Angle between vectors
Parallel vectors
Perpendicular vectors
Scatter plot
Vector plot
Correlated/uncorrelated variables
Sample size
Sample mean
Correlation coefficient
Outliers

1.8 Exercises

1. The data sets displayed in Figures A, B, C, and D have one of the following correlation coefficients; 0.97, 0.93, 0.88, 0.76. Match the figure with its correlation coefficient.

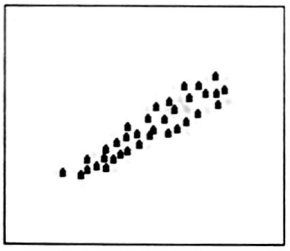


Figure A.

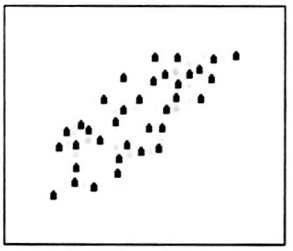


Figure B.

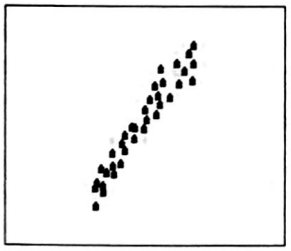


Figure C.

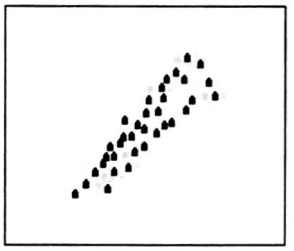


Figure D.

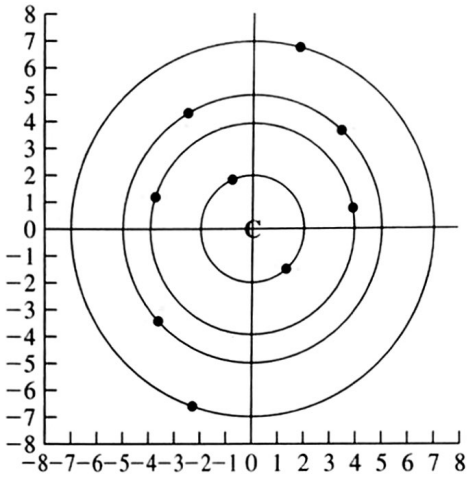
2. A meter that measures flow rates is being calibrated. In this initial test, $n = 8$ flows are sent to the meter and the corresponding meter readings are recorded. Let the set of flows and the corresponding meter readings be given by the 8-vectors $\mathbf{x}$ and $\mathbf{y}$, respectively, where

$$\mathbf{x} = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \end{bmatrix} \quad \text{and} \quad \mathbf{y} = \begin{bmatrix} 1.2 \\ 2.1 \\ 3.3 \\ 3.9 \\ 5.2 \\ 6.1 \\ 6.9 \\ 7.7 \end{bmatrix}.$$

Compute the correlation coefficient between the input flows in $\mathbf{x}$ and the resultant meter readings in $\mathbf{y}$.

3. An experiment to measure the amplitude of a shock wave resulting from the detonation of an explosive charge is conducted by placing recorders at various distances from the charge. (Distances are 100s of feet.) A common arrangement for the recorders is shown in the accompanying figure. The distance of a recorder from the charge and the amplitude of the recorded shock wave are shown in the table. Compute the correlation coefficient between

the distance and amplitude data.



• – Recorder C – Explosive charge

Distance	*Amplitude*
200	12.6
200	19.9
400	9.3
400	9.5
500	7.9
500	7.8
500	8.0
700	6.0
700	6.4

4. An equal number of two-parent families, each with three children younger than ten years old were interviewed in cities of populations ranging from 25,000 to 75,000. Interviewers collected data on (average) yearly living expenses for housing (rental/mortgage payments), food, and clothing. The collected living expense data were rounded to the nearest 100 dollars. Compute the correlation coefficient between the population data and living expense data shown in the following table:

City Population (in 1000s)	*Average Yearly Living Expense (in $ 100s)*
25	72
30	65
35	78
40	70
50	79
60	85
65	83
75	88

Supplementary Exercises

1. Determine the number of entries on or above the main diagonal of a $k \times k$ matrix when

(a) $k = 2$, **(b)** $k = 3$, **(c)** $k = 4$, **(d)** $k = n$.

2. Let $A = \begin{bmatrix} 0 & 2 \\ 0 & 5 \end{bmatrix}$.

(a) Find a $2 \times k$ matrix $B \neq O$ such that $AB = O$ for $k = 1, 2, 3, 4$.

(b) Are your answers to part (a) unique? Explain.

3. Find all 2×2 matrices with real entries of the form

$$A = \begin{bmatrix} a & b \\ 0 & c \end{bmatrix}$$

such that $A^2 = I_2$.

4. An $n \times n$ matrix A (with real entries) is called a **square root** of the $n \times n$ matrix B (with real entries) if $A^2 = B$.

(a) Find a square root of $B = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$.

(b) Find a square root of $B = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$.

(c) Find a square root of $B = I_4$.

(d) Show that there is no square root of

$$B = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}.$$