

- The U.S. Internal Revenue Service uses data collected from tax returns to sort taxpayers into two groups: those that will be audited and those that will not. (See [31].)

Investigation of the dependence among variables

- Data on several variables were used to identify factors that were responsible for client success in hiring external consultants. (See [12].)
- Measurements of variables related to innovation, on the one hand, and variables related to the business environment and business organization, on the other hand, were used to discover why some firms are product innovators and some firms are not. (See [3].)
- Measurements of pulp fiber characteristics and subsequent measurements of characteristics of the paper made from them are used to examine the relations between pulp fiber properties and the resulting paper properties. The goal is to determine those fibers that lead to higher quality paper. (See [17].)
- The associations between measures of risk-taking propensity and measures of socioeconomic characteristics for top-level business executives were used to assess the relation between risk-taking behavior and performance. (See [18].)

Prediction

- The associations between test scores, and several high school performance variables, and several college performance variables were used to develop predictors of success in college. (See [10].)
- Data on several variables related to the size distribution of sediments were used to develop rules for predicting different depositional environments. (See [7] and [20].)
- Measurements on several accounting and financial variables were used to develop a method for identifying potentially insolvent property-liability insurers. (See [28].)
- cDNA microarray experiments (gene expression data) are increasingly used to study the molecular variations among cancer tumors. A reliable classification of tumors is essential for successful diagnosis and treatment of cancer. (See [9].)

Hypotheses testing

- Several pollution-related variables were measured to determine whether levels for a large metropolitan area were roughly constant throughout the week, or whether there was a noticeable difference between weekdays and weekends. (See Exercise 1.6.)
- Experimental data on several variables were used to see whether the nature of the instructions makes any difference in perceived risks, as quantified by test scores. (See [27].)
- Data on many variables were used to investigate the differences in structure of American occupations to determine the support for one of two competing sociological theories. (See [16] and [25].)
- Data on several variables were used to determine whether different types of firms in newly industrialized countries exhibited different patterns of innovation. (See [15].)

The preceding descriptions offer glimpses into the use of multivariate methods in widely diverse fields.

1.3 The Organization of Data

Throughout this text, we are going to be concerned with analyzing measurements made on several variables or characteristics. These measurements (commonly called *data*) must frequently be arranged and displayed in various ways. For example, graphs and tabular arrangements are important aids in data analysis. Summary numbers, which quantitatively portray certain features of the data, are also necessary to any description.

We now introduce the preliminary concepts underlying these first steps of data organization.

Arrays

Multivariate data arise whenever an investigator, seeking to understand a social or physical phenomenon, selects a number $p \geq 1$ of *variables* or *characters* to record. The values of these variables are all recorded for each distinct *item*, *individual*, or *experimental unit*.

We will use the notation x_{jk} to indicate the particular value of the k th variable that is observed on the j th item, or trial. That is,

x_{jk} = measurement of the k th variable on the j th item

Consequently, n measurements on p variables can be displayed as follows:

| | Variable 1 | Variable 2 | ... | Variable k | ... | Variable p |
|------------|------------|------------|-----|--------------|-----|--------------|
| Item 1: | x_{11} | x_{12} | ... | x_{1k} | ... | x_{1p} |
| Item 2: | x_{21} | x_{22} | ... | x_{2k} | ... | x_{2p} |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| Item j : | x_{j1} | x_{j2} | ... | x_{jk} | ... | x_{jp} |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| Item n : | x_{n1} | x_{n2} | ... | x_{nk} | ... | x_{np} |

Or we can display these data as a rectangular array, called \mathbf{X} , of n rows and p columns:

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2k} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{j1} & x_{j2} & \cdots & x_{jk} & \cdots & x_{jp} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} & \cdots & x_{np} \end{bmatrix}$$

The array \mathbf{X} , then, contains the data consisting of all of the observations on all of the variables.

Example 1.1 (A data array) A selection of four receipts from a university bookstore was obtained in order to investigate the nature of book sales. Each receipt provided, among other things, the number of books sold and the total amount of each sale. Let the first variable be total dollar sales and the second variable be number of books sold. Then we can regard the corresponding numbers on the receipts as four measurements on two variables. Suppose the data, in tabular form, are

Variable 1 (dollar sales): 42 52 48 58
 Variable 2 (number of books): 4 5 4 3

Using the notation just introduced, we have

$$x_{11} = 42 \quad x_{21} = 52 \quad x_{31} = 48 \quad x_{41} = 58$$

$$x_{12} = 4 \quad x_{22} = 5 \quad x_{32} = 4 \quad x_{42} = 3$$

and the data array \mathbf{X} is

$$\mathbf{X} = \begin{bmatrix} 42 & 4 \\ 52 & 5 \\ 48 & 4 \\ 58 & 3 \end{bmatrix}$$

with four rows and two columns.

Considering data in the form of arrays facilitates the exposition of the subject matter and allows numerical calculations to be performed in an orderly and efficient manner. The efficiency is twofold, as gains are attained in both (1) *describing* numerical calculations as operations on arrays and (2) the *implementation* of the calculations on computers, which now use many languages and statistical packages to perform array operations. We consider the manipulation of arrays of numbers in Chapter 2. At this point, we are concerned only with their value as devices for displaying data.

Descriptive Statistics

A large data set is bulky, and its very mass poses a serious obstacle to any attempt to visually extract pertinent information. Much of the information contained in the data can be assessed by calculating certain summary numbers, known as *descriptive statistics*. For example, the arithmetic average, or sample mean, is a descriptive statistic that provides a measure of location—that is, a “central value” for a set of numbers. And the average of the squares of the distances of all of the numbers from the mean provides a measure of the spread, or variation, in the numbers.

We shall rely most heavily on descriptive statistics that measure location, variation, and linear association. The formal definitions of these quantities follow.

Let $x_{11}, x_{21}, \dots, x_{n1}$ be n measurements on the first variable. Then the arithmetic average of these measurements is

$$\bar{x}_1 = \frac{1}{n} \sum_{j=1}^n x_{j1}$$

If the n measurements represent a subset of the full set of measurements that might have been observed, then \bar{x}_1 is also called the *sample mean* for the first variable. We adopt this terminology because the bulk of this book is devoted to procedures designed to analyze samples of measurements from larger collections.

The sample mean can be computed from the n measurements on each of the p variables, so that, in general, there will be p sample means:

$$\bar{x}_k = \frac{1}{n} \sum_{j=1}^n x_{jk} \quad k = 1, 2, \dots, p \quad (1-1)$$

A measure of spread is provided by the *sample variance*, defined for n measurements on the first variable as

$$s_1^2 = \frac{1}{n} \sum_{j=1}^n (x_{j1} - \bar{x}_1)^2$$

where \bar{x}_1 is the sample mean of the x_{j1} 's. In general, for p variables, we have

$$s_k^2 = \frac{1}{n} \sum_{j=1}^n (x_{jk} - \bar{x}_k)^2 \quad k = 1, 2, \dots, p \quad (1-2)$$

Two comments are in order. First, many authors define the sample variance with a divisor of $n - 1$ rather than n . Later we shall see that there are theoretical reasons for doing this, and it is particularly appropriate if the number of measurements, n , is small. The two versions of the sample variance will always be differentiated by displaying the appropriate expression.

Second, although the s^2 notation is traditionally used to indicate the sample variance, we shall eventually consider an array of quantities in which the sample variances lie along the main diagonal. In this situation, it is convenient to use double subscripts on the variances in order to indicate their positions in the array. Therefore, we introduce the notation s_{kk} to denote the same variance computed from measurements on the k th variable, and we have the notational identities

$$s_k^2 = s_{kk} = \frac{1}{n} \sum_{j=1}^n (x_{jk} - \bar{x}_k)^2 \quad k = 1, 2, \dots, p \quad (1-3)$$

The square root of the sample variance, $\sqrt{s_{kk}}$, is known as the *sample standard deviation*. This measure of variation uses the same units as the observations.

Consider n pairs of measurements on each of variables 1 and 2:

$$\begin{bmatrix} x_{11} \\ x_{12} \end{bmatrix}, \begin{bmatrix} x_{21} \\ x_{22} \end{bmatrix}, \dots, \begin{bmatrix} x_{n1} \\ x_{n2} \end{bmatrix}$$

That is, x_{j1} and x_{j2} are observed on the j th experimental item ($j = 1, 2, \dots, n$). A measure of linear association between the measurements of variables 1 and 2 is provided by the *sample covariance*

$$s_{12} = \frac{1}{n} \sum_{j=1}^n (x_{j1} - \bar{x}_1)(x_{j2} - \bar{x}_2)$$

or the average product of the deviations from their respective means. If large values for one variable are observed in conjunction with large values for the other variable, and the small values also occur together, s_{12} will be positive. If large values from one variable occur with small values for the other variable, s_{12} will be negative. If there is no particular association between the values for the two variables, s_{12} will be approximately zero.

The *sample covariance*

$$s_{jk} = \frac{1}{n} \sum_{j=1}^n (x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k) \quad i = 1, 2, \dots, p, \quad k = 1, 2, \dots, p \quad (1-4)$$

measures the association between the i th and k th variables. We note that the covariance reduces to the sample variance when $i = k$. Moreover, $s_{jk} = s_{kj}$ for all i and k .

The final descriptive statistic considered here is the *sample correlation coefficient* (or *Pearson's product-moment correlation coefficient*; see [14]). This measure of the linear association between two variables does not depend on the units of measurement. The sample correlation coefficient for the i th and k th variables is defined as

$$r_{ik} = \frac{s_{ik}}{\sqrt{s_{ii} s_{kk}}} = \frac{\sum_{j=1}^n (x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k)}{\sqrt{\sum_{j=1}^n (x_{ji} - \bar{x}_i)^2} \sqrt{\sum_{j=1}^n (x_{jk} - \bar{x}_k)^2}} \quad (1-5)$$

for $i = 1, 2, \dots, p$ and $k = 1, 2, \dots, p$. Note $r_{ik} = r_{ki}$ for all i and k .

The sample correlation coefficient is a standardized version of the sample covariance, where the product of the square roots of the sample variances provides the standardization. Notice that r_{ik} has the same value whether n or $n - 1$ is chosen as the common divisor for s_{ij} , s_{kk} , and s_{ik} .

The sample correlation coefficient r_{ik} can also be viewed as a *sample covariance*. Suppose the original values x_{ji} and x_{jk} are replaced by *standardized* values $(x_{ji} - \bar{x}_i)/\sqrt{s_{ii}}$ and $(x_{jk} - \bar{x}_k)/\sqrt{s_{kk}}$. The standardized values are commensurable because both sets are centered at zero and expressed in standard deviation units. The sample correlation coefficient is just the sample covariance of the standardized observations.

Although the signs of the sample correlation and the sample covariance are the same, the correlation is ordinarily easier to interpret because its magnitude is bounded. To summarize, the sample correlation r has the following properties:

1. The value of r must be between -1 and $+1$ inclusive.
2. Here r measures the strength of the linear association. If $r = 0$, this implies a lack of linear association between the components. Otherwise, the sign of r indicates the direction of the association: $r < 0$ implies a tendency for one value in the pair to be larger than its average when the other is smaller than its average; and $r > 0$ implies a tendency for one value of the pair to be large when the other value is large and also for both values to be small together.
3. The value of r_{ik} remains unchanged if the measurements of the i th variable are changed to $y_i = ax_i + b$, $j = 1, 2, \dots, n$, and the values of the k th variable are changed to $y_k = cx_k + d$, $j = 1, 2, \dots, n$, provided that the constants a and c have the same sign.

The quantities s_{jk} and r_{ik} do not, in general, convey all there is to know about the association between two variables. Nonlinear associations can exist that are not revealed by these descriptive statistics. Covariance and correlation provide measures of *linear* association, or association along a line. Their values are less informative for other kinds of association. On the other hand, these quantities can be very sensitive to "wild" observations ("outliers") and may indicate association when, in fact, little exists. In spite of these shortcomings, covariance and correlation coefficients are routinely calculated and analyzed. They provide cogent numerical summaries of association when the data do not exhibit obvious nonlinear patterns of association and when wild observations are not present.

Suspect observations must be accounted for by correcting obvious recording mistakes and by taking actions consistent with the identified causes. The values of s_{ik} and r_{ik} should be quoted both with and without these observations.

The sum of squares of the deviations from the mean and the sum of cross-product deviations are often of interest themselves. These quantities are

$$w_{kk} = \sum_{j=1}^n (x_{jk} - \bar{x}_k)^2 \quad k = 1, 2, \dots, p \quad (1-6)$$

and

$$w_{ik} = \sum_{j=1}^n (x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k) \quad i = 1, 2, \dots, p, \quad k = 1, 2, \dots, p \quad (1-7)$$

The descriptive statistics computed from n measurements on p variables can also be organized into arrays.

Arrays of Basic Descriptive Statistics

Sample means $\bar{\mathbf{x}} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{bmatrix}$

Sample variances and covariances $\mathbf{S}_n = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{bmatrix} \quad (1-8)$

Sample correlations $\mathbf{R} = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{bmatrix}$

The sample mean array is denoted by \bar{x} , the sample variance and covariance array by the capital letter S_n , and the sample correlation array by \mathbf{R} . The subscript n on the array S_n is a mnemonic device used to remind you that n is employed as a divisor for the elements s_{jk} . The size of all of the arrays is determined by the number of variables, p .

The arrays S_n and \mathbf{R} consist of p rows and p columns. The array \bar{x} is a single column with p rows. The first subscript on an entry in arrays S_n and \mathbf{R} indicates the row; the second subscript indicates the column. Since $s_{jk} = s_{kj}$ and $r_{jk} = r_{kj}$ for all j and k , the entries in symmetric positions about the main northwest-southeast diagonals in arrays S_n and \mathbf{R} are the same, and the arrays are said to be *symmetric*.

Example 1.2 (The arrays \bar{x} , S_n , and \mathbf{R} for bivariate data) Consider the data introduced in Example 1.1. Each receipt yields a pair of measurements, total dollar sales, and number of books sold. Find the arrays \bar{x} , S_n , and \mathbf{R} .

Since there are four receipts, we have a total of four measurements (observations) on each variable.

The sample means are

$$\bar{x}_1 = \frac{1}{4} \sum_{j=1}^4 x_{j1} = \frac{1}{4}(42 + 52 + 48 + 58) = 50$$

$$\bar{x}_2 = \frac{1}{4} \sum_{j=1}^4 x_{j2} = \frac{1}{4}(4 + 5 + 4 + 3) = 4$$

$$\bar{\mathbf{x}} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \end{bmatrix} = \begin{bmatrix} 50 \\ 4 \end{bmatrix}$$

The sample variances and covariances are

$$s_{11} = \frac{1}{4} \sum_{j=1}^4 (x_{j1} - \bar{x}_1)^2 = \frac{1}{4}((42 - 50)^2 + (52 - 50)^2 + (48 - 50)^2 + (58 - 50)^2) = 34$$

$$s_{22} = \frac{1}{4} \sum_{j=1}^4 (x_{j2} - \bar{x}_2)^2 = \frac{1}{4}((4 - 4)^2 + (5 - 4)^2 + (4 - 4)^2 + (3 - 4)^2) = .5$$

$$s_{12} = \frac{1}{4} \sum_{j=1}^4 (x_{j1} - \bar{x}_1)(x_{j2} - \bar{x}_2) = \frac{1}{4}((42 - 50)(4 - 4) + (52 - 50)(5 - 4) + (48 - 50)(4 - 4) + (58 - 50)(3 - 4)) = -1.5$$

$$s_{21} = s_{12}$$

and

$$S_n = \begin{bmatrix} 34 & -1.5 \\ -1.5 & .5 \end{bmatrix}$$

The sample correlation is

$$r_{12} = \frac{s_{12}}{\sqrt{s_{11}} \sqrt{s_{22}}} = \frac{-1.5}{\sqrt{34} \sqrt{.5}} = -.36$$

$$r_{21} = r_{12}$$

so

$$\mathbf{R} = \begin{bmatrix} 1 & -.36 \\ -.36 & 1 \end{bmatrix}$$

Graphical Techniques

Plots are important, but frequently neglected, aids in data analysis. Although it is impossible to simultaneously plot *all* the measurements made on several variables and study the configurations, plots of individual variables and plots of pairs of variables can still be very informative. Sophisticated computer programs and display equipment allow one the luxury of visually examining data in one, two, or three dimensions with relative ease. On the other hand, many valuable insights can be obtained from the data by constructing plots with paper and pencil. Simple, yet elegant and effective, methods for displaying data are available in [29]. It is good statistical practice to plot pairs of variables and visually inspect the pattern of association. Consider, then, the following seven pairs of measurements on two variables:

| | | | | | | | |
|-----------------------|---|-----|---|---|----|---|-----|
| Variable 1 (x_1): | 3 | 4 | 2 | 6 | 8 | 2 | 5 |
| Variable 2 (x_2): | 5 | 5.5 | 4 | 7 | 10 | 5 | 7.5 |

These data are plotted as seven points in two dimensions (each axis representing a variable) in Figure 1.1. The coordinates of the points are determined by the paired measurements: (3, 5), (4, 5.5), ..., (5, 7.5). The resulting two-dimensional plot is known as a *scatter diagram* or *scatter plot*.

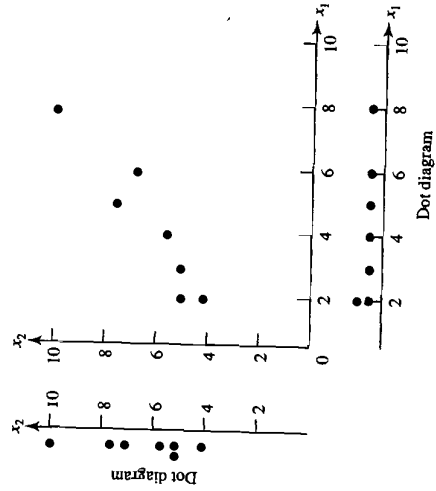


Figure 1.1 A scatter plot and marginal dot diagrams