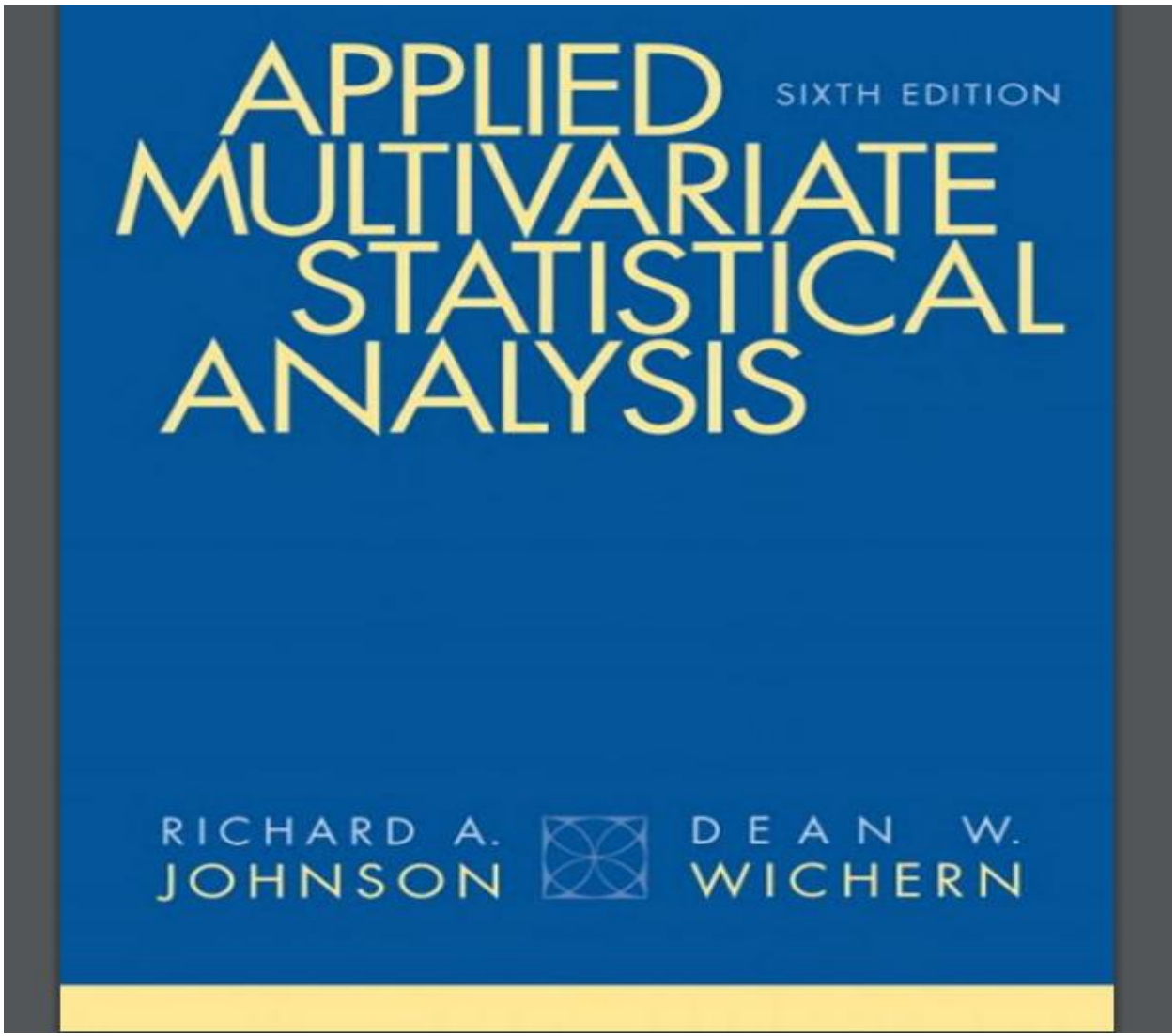
Advanced Multivariate Methods



Chapter 1: Steps in Multivariate Analysis

- Univariate – Architecture of Data, Variable Structure, Minimum and Maximum, Measures of Centrality and Dispersion
- Bivariate – Correlation, ANOVA, Simply Linear Regression
- Multivariate – Multivariate Regression, Classification Methods
- STATISTICS: Science of Variation, Comparison of Unit of Analysis and How Different

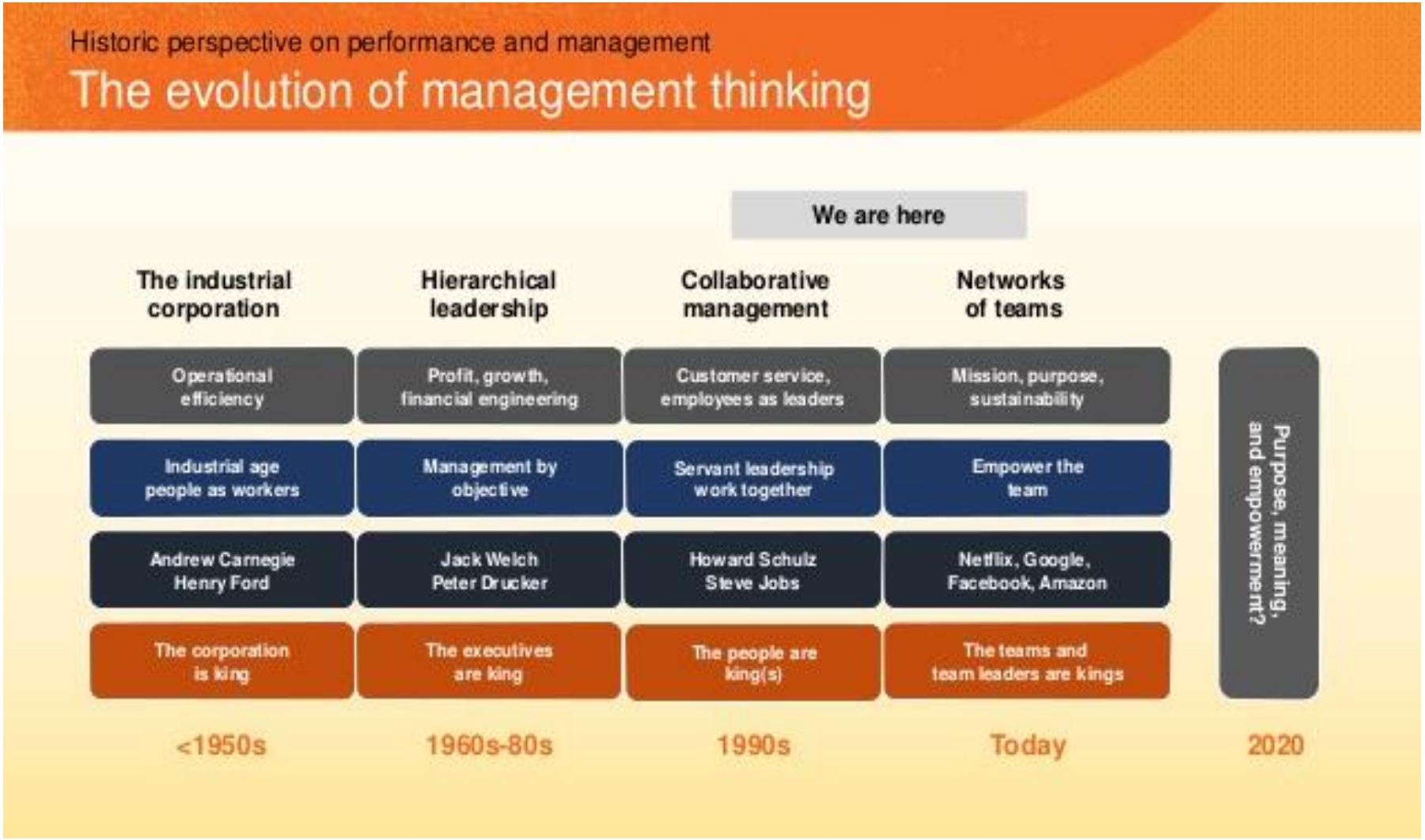
Chapter 1: Aspects of Multivariate Analysis

People Analytics: Taking Off

Progress Rapidly Accelerating

		2015	2016	% Change
Plan	Performing multi-year workforce planning	38%	48%	+ 26%
Correlate	Correlating people data to business performance	24%	39%	+ 63%
	Correlating people data to business performance (% excellent)	5%	11%	+ 120%
Predict	Using people data to predict business performance	28%	36%	+ 29%
	Using people data to predict performance (% excellent)	4%	9%	+ 125%

History of Data Analysis in Any Discipline: i.e. Business Analytics, Computational Biology, Proteomics and Genomics



1.1 Introduction: Application of Multivariate Methods to Scientific Investigation (Research Projects)

The objectives of scientific investigations to which multivariate methods most naturally lend themselves include the following:

1. *Data reduction or structural simplification.* The phenomenon being studied is represented as simply as possible without sacrificing valuable information. It is hoped that this will make interpretation easier.
2. *Sorting and grouping.* Groups of "similar" objects or variables are created, based upon measured characteristics. Alternatively, rules for classifying objects into well-defined groups may be required.
3. *Investigation of the dependence among variables.* The nature of the relationships among variables is of interest. Are all the variables mutually independent or are one or more variables dependent on the others? If so, how?
4. *Prediction.* Relationships between variables must be determined for the purpose of predicting the values of one or more variables on the basis of observations on the other variables.
5. *Hypothesis construction and testing.* Specific statistical hypotheses, formulated in terms of the parameters of multivariate populations, are tested. This may be done to validate assumptions or to reinforce prior convictions.

Multivariate Analysis – Outcomes (Definable, Measurable, Evidence-Based)

1.2 Applications of Multivariate Techniques: Data Reduction or Simplification (Factor Analysis)

Data reduction or simplification

- Using data on several variables related to cancer patient responses to radiotherapy, a simple measure of patient response to radiotherapy was constructed. (See Exercise 1.15.)
- Track records from many nations were used to develop an index of performance for both male and female athletes. (See [8] and [22].)
- Multispectral image data collected by a high-altitude scanner were reduced to a form that could be viewed as images (pictures) of a shoreline in two dimensions. (See [23].)
- Data on several variables relating to yield and protein content were used to create an index to select parents of subsequent generations of improved bean plants. (See [13].)
- A matrix of tactic similarities was developed from aggregate data derived from professional mediators. From this matrix the number of dimensions by which professional mediators judge the tactics they use in resolving disputes was determined. (See [21].)

1.2 Applications of Multivariate Techniques: Data Reduction or Simplification (Clustering)

Sorting and grouping

- Data on several variables related to computer use were employed to create clusters of categories of computer jobs that allow a better determination of existing (or planned) computer utilization. (See [2].)
- Measurements of several physiological variables were used to develop a screening procedure that discriminates alcoholics from nonalcoholics. (See [26].)
- Data related to responses to visual stimuli were used to develop a rule for separating people suffering from a multiple-sclerosis-caused visual pathology from those not suffering from the disease. (See Exercise 1.14.)

- 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].)

1.2 Applications of Multivariate Techniques: Investigation of Dependence Among Variables(Discriminant Analysis)

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].)

1.2 Applications of Multivariate Techniques: Prediction (Regression)

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].)

1.2 Applications of Multivariate Techniques: Hypothesis Testing (z-test; t-test; confidence intervals about the mean)

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].)

1.3 Organization of Data: Measures of Dispersion and Frequency Counts

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.

1.3 Organization of Data: Array; Rows and Column Display

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 kth variable that is observed on the jth item, or trial. That is,

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

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

	Variable 1	Variable 2	$\cdots$	Variable k	$\cdots$	Variable p
Item 1:	x_{11}	x_{12}	$\cdots$	x_{1k}	$\cdots$	x_{1p}
Item 2:	x_{21}	x_{22}	$\cdots$	x_{2k}	$\cdots$	x_{2p}
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
Item j:	x_{j1}	x_{j2}	$\cdots$	x_{jk}	$\cdots$	x_{jp}
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
Item n:	x_{n1}	x_{n2}	$\cdots$	x_{nk}	$\cdots$	x_{np}

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

Data File Format For Analysis

Suppose that the response variable y in an experiment is expected to be influenced by two input variables x_1 and x_2, and the data relevant to these input variables are recorded along with the measurements of y. With n runs of an experiment, we have a data set of the form shown in Table 5.

TABLE 5 Data Structure for Multiple Regression with Two Input Variables

Experimental	Input Variables		Response
Run	x_1	x_2	y
1	x_{11}	x_{12}	y_1
2	x_{21}	x_{22}	y_2
.	.	.	.
.	.	.	.
.	.	.	.
i	x_{i1}	x_{i2}	y_i
.	.	.	.
.	.	.	.
.	.	.	.
n	x_{n1}	x_{n2}	y_n

By analogy with the simple linear regression model, we can then tentatively formulate:

TABLE 5 Data Structure for Multiple Regression with Two Input Variables

Experimental	Input Variables		Response
Run	x_1	x_2	y
1	x_{11}	x_{12}	y_1
2	x_{21}	x_{22}	y_2
.	.	.	.
.	.	.	.
.	.	.	.
i	x_{i1}	x_{i2}	y_i
.	.	.	.
.	.	.	.
.	.	.	.
n	x_{n1}	x_{n2}	y_n

Table 5 (p. 504)

Data Structure for Multiple Regression with Two Input Variables

2.1 Data from an Environmental Survey

The data set shown in Figure 2.3 represents 30 responses from a questionnaire concerning the president's environmental policies. (See the file **Questionnaire Data.xlsx**.) Identify the variables and observations.

	A	B	C	D	E	F	G
1	Person	Age	Gender	State	Children	Salary	Opinion
2	1	35	Male	Minnesota	1	$65,400	5
3	2	61	Female	Texas	2	$62,000	1
4	3	35	Male	Ohio	0	$63,200	3
5	4	37	Male	Florida	2	$52,000	5
6	5	32	Female	California	3	$81,400	1
7	6	33	Female	New York	3	$46,300	5
8	7	65	Female	Minnesota	2	$49,600	1
9	8	45	Male	New York	1	$45,900	5
10	9	40	Male	Texas	3	$47,700	4
11	10	32	Female	Texas	1	$59,900	4
12	11	57	Male	New York	1	$48,100	4
13	12	38	Female	Virginia	0	$58,100	3
14	13	37	Female	Illinois	2	$56,000	1
15	14	42	Female	Virginia	2	$53,400	1
16	15	38	Female	New York	2	$39,000	2
17	16	48	Male	Michigan	1	$61,500	2
18	17	40	Male	Ohio	0	$37,700	1
19	18	57	Female	Michigan	2	$36,700	4
20	19	44	Male	Florida	2	$45,200	3
21	20	40	Male	Michigan	0	$59,000	4
22	21	21	Female	Minnesota	2	$54,300	2
23	22	49	Male	New York	1	$62,100	4
24	23	34	Male	New York	0	$78,000	3
25	24	49	Male	Arizona	0	$43,200	5
26	25	40	Male	Arizona	1	$44,500	3
27	26	38	Male	Ohio	1	$43,300	1
28	27	27	Male	Illinois	3	$45,400	2
29	28	63	Male	Michigan	2	$53,900	1
30	29	52	Male	California	1	$44,100	3
31	30	48	Female	New York	2	$31,000	4

1.3 Organization of Data: Rectangular Array (Linear Algebra)

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

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

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

1.3 Organization of Data: Unit of Analysis (What is Being Compared) and 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. ■

1.3 Organization of the Data: Array as Implementation in Computer (2 variables: 2 vectors)

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.

1.3 Organization of Data: Descriptive Statistics - Mean

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

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

Describing a Data Set of Measurements, p. 25.

Describing a Data Set of Measurements

1. Summarization and description of the overall pattern.
   (a) Presentation of tables and graphs.
   (b) Noting important features of the graphed data including symmetry or departures from it.
   (c) Scanning the graphed data to detect any observations that seem to stick far out from the major mass of the data—the outliers.
2. Computation of numerical measures.
   (a) A typical or representative value that indicates the center of the data.
   (b) The amount of spread or variation present in the data.

1.3 Organization of Data: Descriptive Statistics – Mean and Variance

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:

$$\bar{x}_k = \frac{1}{n}\sum_{j=1}^{n} x_{jk} \qquad k = 1, 2, \ldots, p \tag{1-1}$$

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

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

The sum $x_1 + x_2 + \cdots + x_n$ is denoted as $\sum_{i=1}^{n} x_i$.

Read this as "the sum of all x_i with i ranging from 1 to n."

The **sample mean** of a set of n measurements $x_1, x_2, \ldots, x_n$ is the sum of these measurements divided by n. The sample mean is denoted by $\bar{x}$.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \quad \text{or} \quad \frac{\sum x_i}{n}$$

Sample Mean, p. 42.

Calculating the Mean

According to the concept of "average," the mean represents a center of a data set. If we picture the dot diagram of a data set as a thin weightless horizontal bar on which balls of equal size and weight are placed at the positions of the data points, then the mean $\overline{x}$ represents the point on which the bar will balance. The computation of the sample mean and its physical interpretation are illustrated in Example 7.

Example 7

Calculating and Interpreting the Sample Mean

The birth weights in pounds of five babies born one day in the same hospital are 9.2, 6.4, 10.5, 8.1, and 7.8. Obtain the sample mean and create a dot diagram.

SOLUTION

The mean birth weight for these data is

$$\overline{x} = \frac{9.2 + 6.4 + 10.5 + 8.1 + 7.8}{5} = \frac{42.0}{5} = 8.4 \text{ pounds}$$

The dot diagram of the data appears in Figure 9, where the sample mean (marked by Δ) is the balancing point or center of the picture.

Figure 9 Dot diagram and the sample mean for the birth-weight data.

1.3 Organization of Data: Sample Variance and Standard Deviation

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 kth variable, and we have the notational identities

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

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:

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

Variance and Standard Deviation (Square Root of Variance)

$$\sum (\text{Deviations}) = \sum (x_i - \bar{x}) = 0$$

To obtain a measure of spread, we must eliminate the signs of the deviations before averaging. One way of removing the interference of signs is to square the numbers. A measure of spread, called the **sample variance**, is constructed by adding the squared deviations and dividing the total by the number of observations minus one.

Sample variance of n observations:

$$s^2 = \frac{\text{sum of squared deviations}}{n-1}$$

$$= \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

TABLE 9 Calculation of Deviations

Observation x	Deviation $x - \bar{x}$
3	−3
5	−1
7	1
7	1
8	2

Table 9, Calculation of Deviations, p. 50.

Standard Deviation Formula

Because the variance involves a sum of squares, its unit is the square of the unit in which the measurements are expressed. For example, if the data pertain to measurements of weight in pounds, the variance is expressed in $(\text{pounds})^2$. To obtain a measure of variability in the same unit as the data, we take the positive square root of the variance, called the **sample standard deviation**. The standard deviation rather than the variance serves as a basic measure of variability.

Sample Standard Deviation

$$s = \sqrt{\text{Variance}} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1}}$$

Example 12

Calculating the Sample Standard Deviation

Calculate the standard deviation for the data of Example 11.

SOLUTION

We already calculated the variance $s^2 = 4$ so the standard deviation is $s = \sqrt{4} = 2$.

1.3 Organization of Data: Sample Covariance

That is, x_{j1} and x_{j2} are observed on the jth experimental item $(j = 1, 2, \ldots, n)$. 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*

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

measures the association between the ith and kth variables. We note that the covariance reduces to the sample variance when $i = k$. Moreover, $s_{ik} = s_{ki}$ for all i and k.

1.3 Organization of Data: Pearson's Sample Correlation Coefficient

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 ith and kth variables is defined as

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

for $i = 1, 2, \ldots, p$ and $k = 1, 2, \ldots, p$. 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_{ii}, s_{kk}, and s_{ik}.

The sample correlation coefficient r_{ik} can also be viewed as a sample *covariance*.

Sample Correlation Coefficient

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

Box, Sample Correlation Coefficient, p. 98.

Correlation: Sums of Squared Deviations of x and y observations

When the pair (x_i, y_i) has both components above their sample means or both below their sample means, the product of standardized observations will be positive; otherwise it will be negative. Consequently, if most pairs have both components simultaneously above or simultaneously below their means, r will be positive.

An alternative formula for r is used for calculation. It is obtained by canceling the common term $n-1$.

Calculation Formula for the Sample Correlation Coefficient

$$r = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}}$$

where

$$S_{xy} = \sum (x-\overline{x})(y-\overline{y})$$

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

The quantities S_{xx} and S_{yy} are the sums of squared deviations of the x observations and the y observations, respectively. S_{xy} is the sum of cross products of the x deviations with the y deviations. This formula is examined in more detail in Chapter 11.

Calculation Formula for the Sample Correlation Coefficient

$$r = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}}$$

where

$$S_{xy} = \sum (x - \bar{x})(y - \bar{y})$$

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

Box, Calculation Formula for the Sample Correlation Coefficient, p. 99.

Sample Correlation as Sample Covariance: Standardized Z-Score Values, Mean = 0 and Standard Deviation or σ of values from 0 to 3

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.

Sample Correlation 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 ith variable are changed to $y_{ji} = ax_{ji} + b, j = 1, 2, \ldots, n$, and the values of the kth variable are changed to $y_{jk} = cx_{jk} + d, j = 1, 2, \ldots, n$, provided that the constants a and c have the same sign.

Scatter Plots or Diagrams: Degrees of Linear Correlation

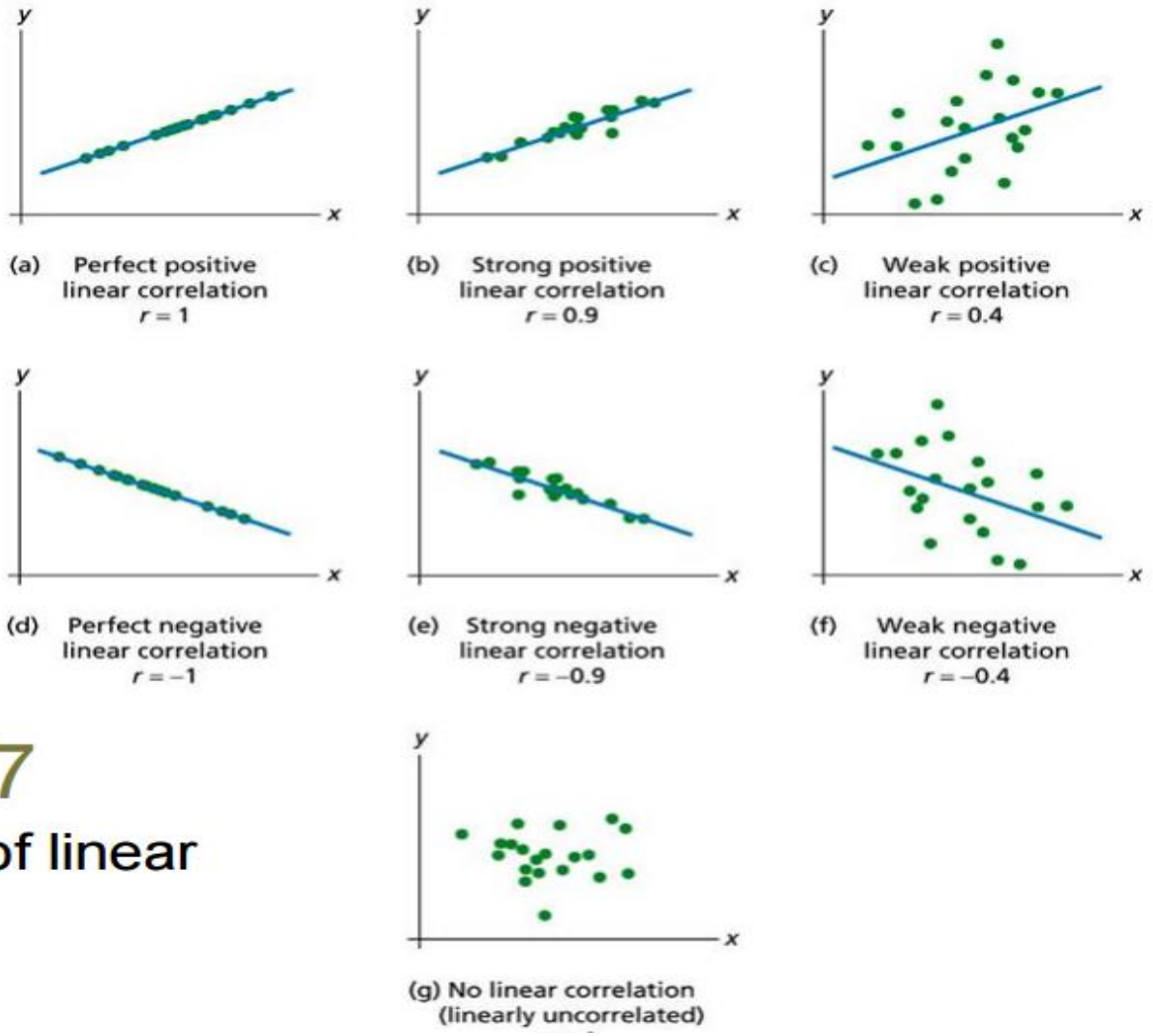


Figure 14.17

Various degrees of linear correlation

Covariance and Pearson Correlation – Linear Association: Sensitive to Outliers

The quantities s_{ik} 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.

Sum of Squares Deviations and Sum of Cross-Products Deviations

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

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

and

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

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

Arrays of Basic Descriptive Statistics

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

Array or Matrix or Matrices – Means, Variances & Covariances, Correlations

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

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

Sample correlations

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

Array Notation – Mean ($\bar{x}$) Covariance (S_n) and Correlation (R)

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

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

Bivariate Data Arrays: (1) Mean

Example 1.2 (The arrays $\bar{\mathbf{x}}$, $\mathbf{S}_n$, 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 $\bar{\mathbf{x}}$, $\mathbf{S}_n$, and $\mathbf{R}$.

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

The sample means are

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

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

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

Bivariate Data Arrays: (2) Variances and Co-variances

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

$$\mathbf{S}_n = \begin{bmatrix} 34 & -1.5 \\ -1.5 & .5 \end{bmatrix}$$

Bivariate Data Arrays: (3) Correlation

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: Bivariate Data

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: $(3, 5)$, $(4, 5.5), \ldots, (5, 7.5)$. The resulting two-dimensional plot is known as a *scatter diagram* or *scatter plot.*

Graphical Techniques: Bivariate Data
Scatter Plot of 2 Variables

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: $(3, 5)$, $(4, 5.5), \ldots, (5, 7.5)$. The resulting two-dimensional plot is known as a *scatter diagram* or *scatter plot.*

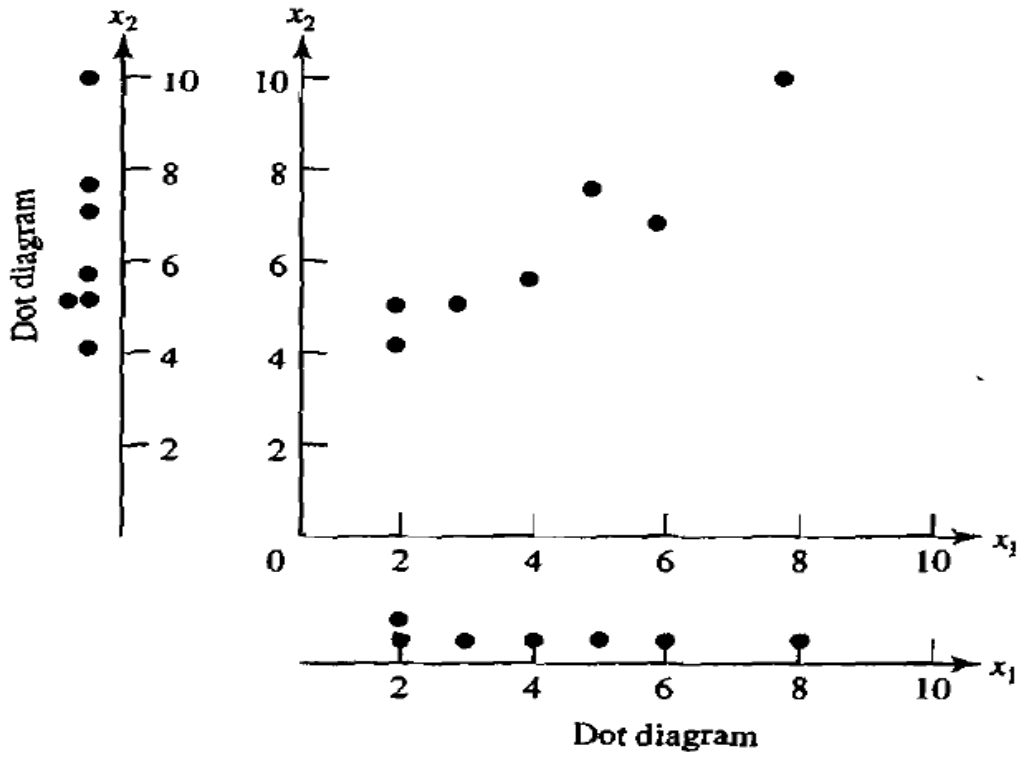


Figure 1.1 A scatter plot and marginal dot diagrams.

Marginal Dot Diagrams or Dot Plot: Univariate Analysis

Also shown in Figure 1.1 are separate plots of the observed values of variable 1 and the observed values of variable 2, respectively. These plots are called (*marginal*) *dot diagrams*. They can be obtained from the original observations or by projecting the points in the scatter diagram onto each coordinate axis.

The information contained in the single-variable dot diagrams can be used to calculate the sample means $\bar{x}_1$ and $\bar{x}_2$ and the sample variances s_{11} and s_{22}. (See Exercise 1.1.) The scatter diagram indicates the orientation of the points, and their coordinates can be used to calculate the sample covariance s_{12}. In the scatter diagram of Figure 1.1, large values of x_1 occur with large values of x_2 and small values of x_1 with small values of x_2. Hence, s_{12} will be positive.

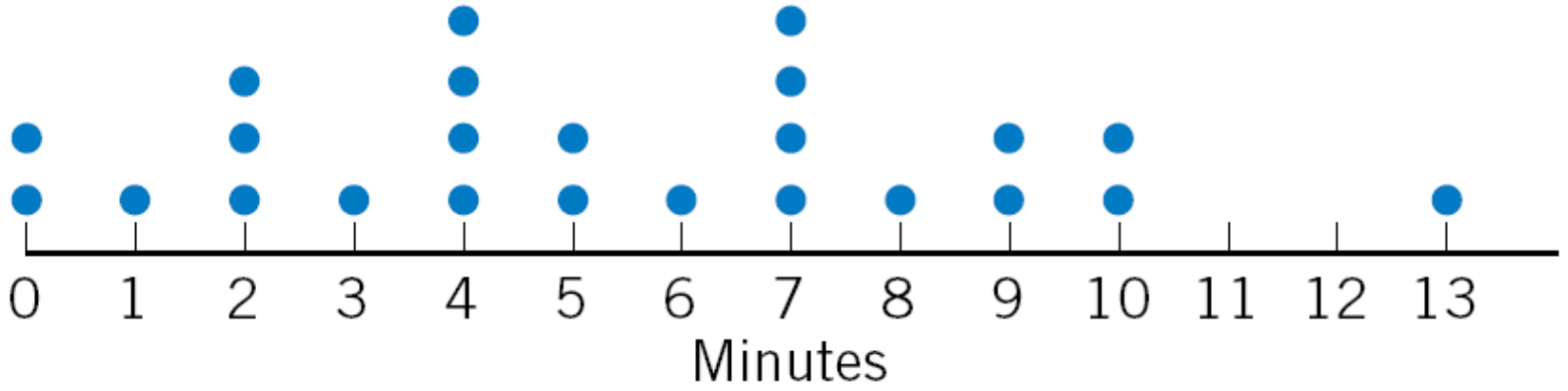


Figure 6, Time not concentrating on the mathematics assignment (out of 20 minutes), p. 32.

Difference Between Dot Diagrams (Univariate) and Scatter Plot (Bivariate)

Dot diagrams and scatter plots contain different kinds of information. The information in the marginal dot diagrams is not sufficient for constructing the scatter plot. As an illustration, suppose the data preceding Figure 1.1 had been paired differently, so that the measurements on the variables x_1 and x_2 were as follows:

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

(We have simply rearranged the values of variable 1.) The scatter and dot diagrams for the "new" data are shown in Figure 1.2. Comparing Figures 1.1 and 1.2, we find that the marginal dot diagrams are the same, but that the scatter diagrams are decidedly different. In Figure 1.2, large values of x_1 are paired with small values of x_2 and small values of x_1 with large values of x_2. Consequently, the descriptive statistics for the individual variables $\bar{x}_1$, $\bar{x}_2$, s_{11}, and s_{22} remain unchanged, but the sample covariance s_{12}, which measures the association between pairs of variables, will now be negative.

Dot Plot and Scatter Plot Overlay

The different orientations of the data in Figures 1.1 and 1.2 are not discernible from the marginal dot diagrams alone. At the same time, the fact that the marginal dot diagrams are the same in the two cases is not immediately apparent from the scatter plots. The two types of graphical procedures complement one another; they are not competitors.

The next two examples further illustrate the information that can be conveyed by a graphic display.

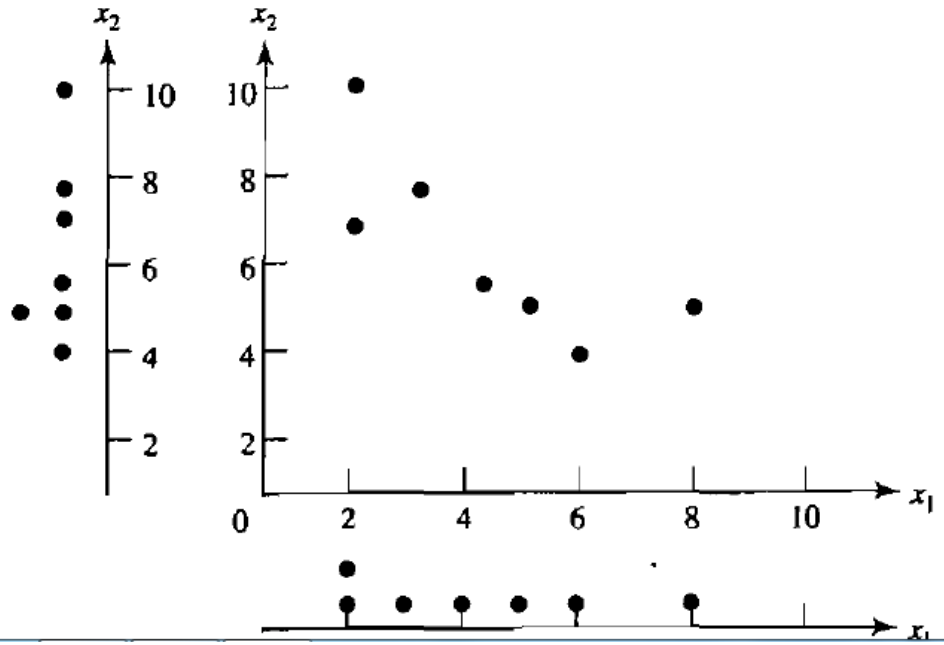


Figure 1.2 Scatter plot and dot diagrams for rearranged data.

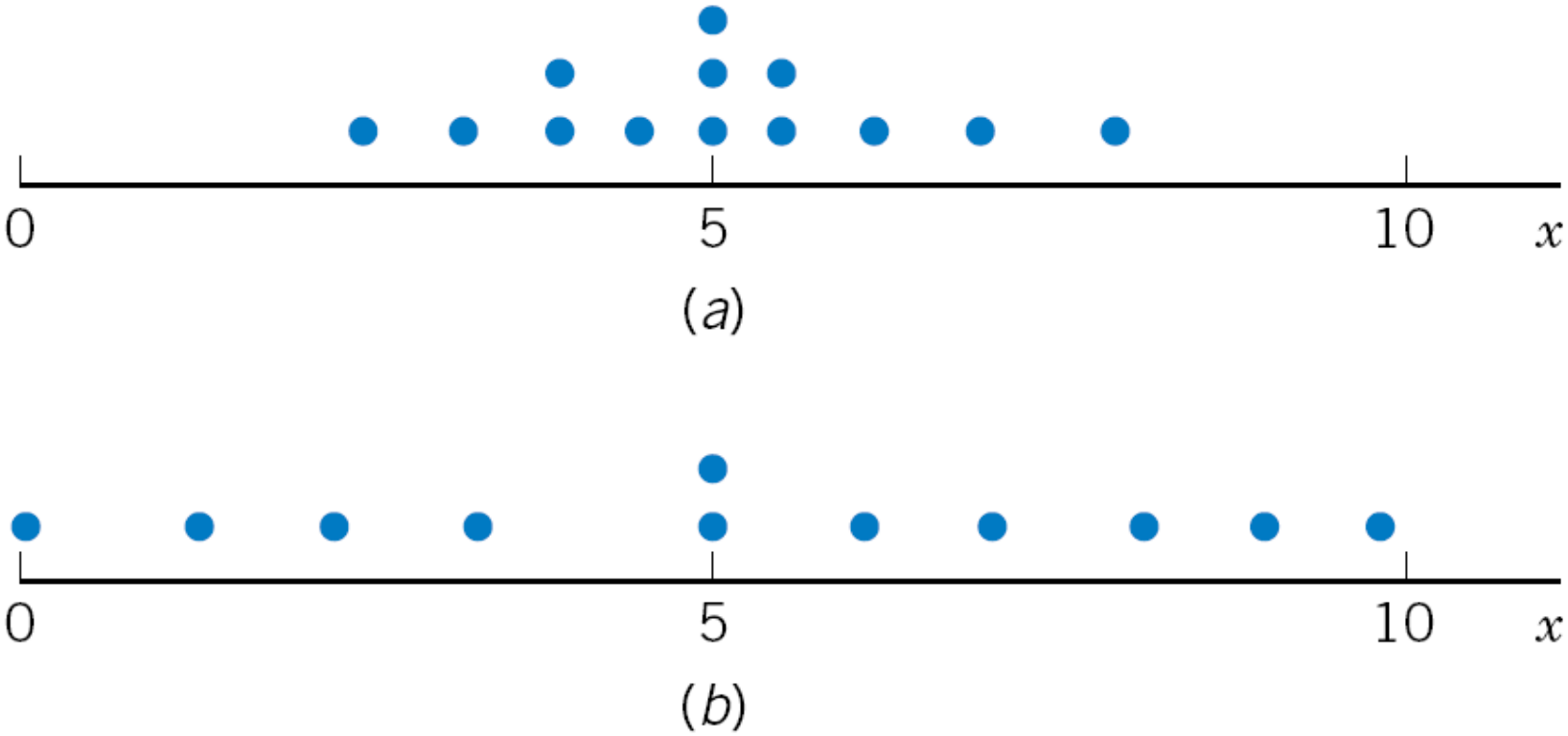


Figure 10, Dot diagrams with similar center values but different amounts of variation, p. 49.

Scatter Plot Illustrates Correlation

Example 1.3 (The effect of unusual observations on sample correlations) Some financial data representing jobs and productivity for the 16 largest publishing firms appeared in an article in *Forbes* magazine on April 30, 1990. The data for the pair of variables x_1 = employees (jobs) and x_2 = profits per employee (productivity) are graphed in Figure 1.3. We have labeled two "unusual" observations. Dun & Bradstreet is the largest firm in terms of number of employees, but is "typical" in terms of profits per employee. Time Warner has a "typical" number of employees, but comparatively small (negative) profits per employee.

Correlation Between Profits Per Employee and No. of Employees (N = 16 Corporations or Firms)

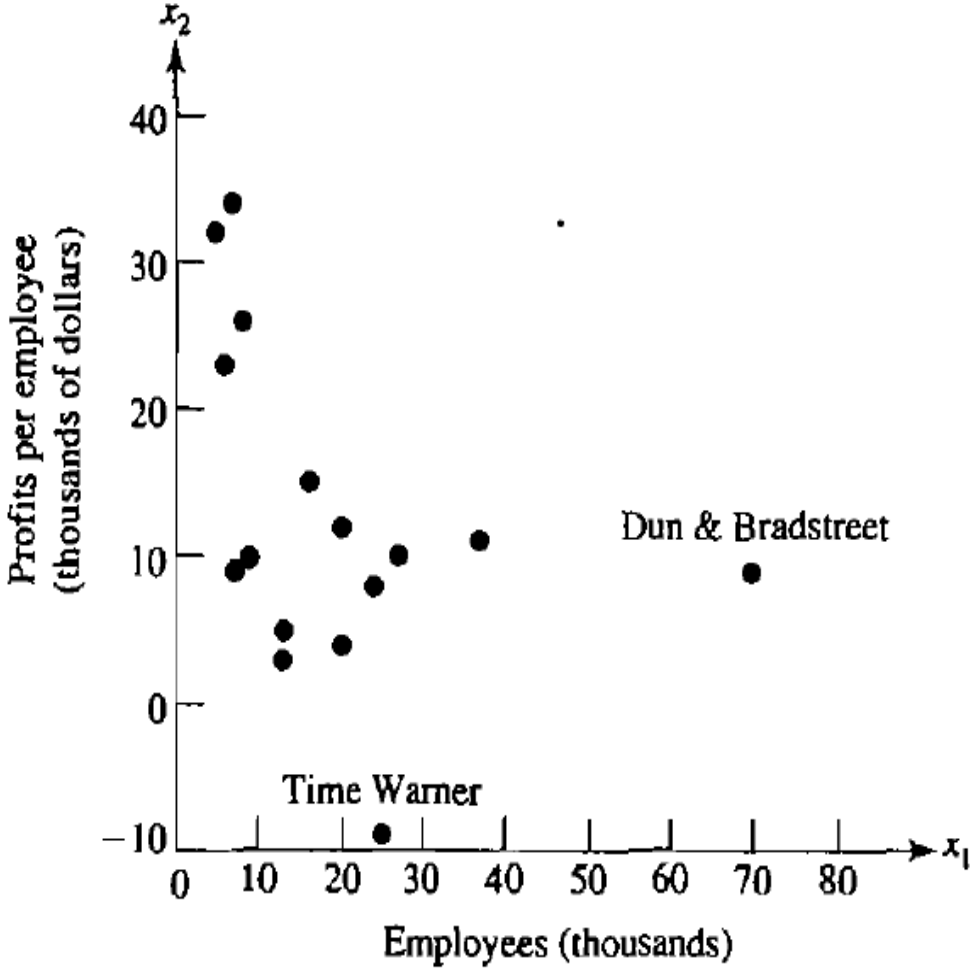


Figure 1.3 Profits per employee and number of employees for 16 publishing firms.

Effect of Outliers or Leverage Points on Correlation Coefficient

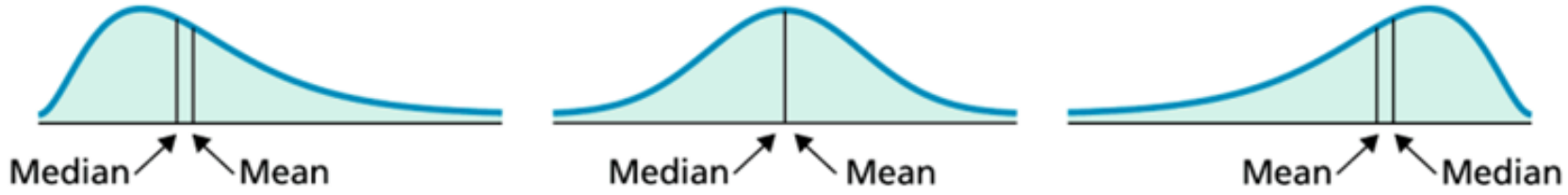
The sample correlation coefficient computed from the values of x_1 and x_2 is

$$r_{12} = \begin{cases} -.39 & \text{for all 16 firms} \\ -.56 & \text{for all firms but Dun \& Bradstreet} \\ -.39 & \text{for all firms but Time Warner} \\ -.50 & \text{for all firms but Dun \& Bradstreet and Time Warner} \end{cases}$$

It is clear that atypical observations can have a considerable effect on the sample correlation coefficient. ■

Skewness: Location of Mean and Median

Relative positions of the mean and median for (a) right-skewed, (b) symmetric, and (c) left-skewed distributions



(a) Right skewed (b) Symmetric (c) Left skewed

Scatter Plot for Baseball Data

Example 1.4 (A scatter plot for baseball data) In a July 17, 1978, article on money in sports, *Sports Illustrated* magazine provided data on x_1 = player payroll for National League East baseball teams.

We have added data on x_2 = won–lost percentage for 1977. The results are given in Table 1.1.

The scatter plot in Figure 1.4 supports the claim that a championship team can be bought. Of course, this cause–effect relationship cannot be substantiated, because the experiment did not include a random assignment of payrolls. Thus, statistics cannot answer the question: Could the Mets have won with $4 million to spend on player salaries?

Selected Teams from National League East (n = 6) and Variables

Table 1.1 1977 Salary and Final Record for the National League East

Team	x_1 = player payroll	x_2 = won–lost percentage
Philadelphia Phillies	3,497,900	.623
Pittsburgh Pirates	2,485,475	.593
St. Louis Cardinals	1,782,875	.512
Chicago Cubs	1,725,450	.500
Montreal Expos	1,645,575	.463
New York Mets	1,469,800	.395

Scatter Plot Between 2 Variables Player Payroll and Won-Lost %age

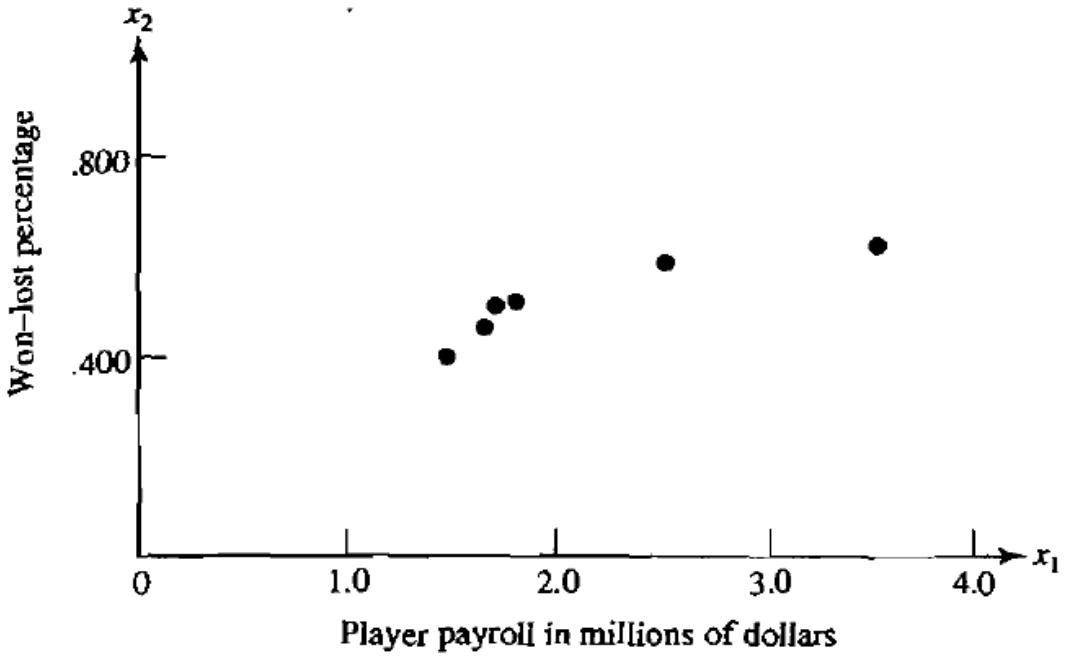


Figure 1.4 Salaries and won–lost percentage from Table 1.1.

To construct the scatter plot in Figure 1.4, we have regarded the six paired observations in Table 1.1 as the coordinates of six points in two-dimensional space. The figure allows us to examine visually the grouping of teams with respect to the variables total payroll and won–lost percentage. ■

Scatter Plots for Several Variables

Example 1.5 (Multiple scatter plots for paper strength measurements) Paper is manufactured in continuous sheets several feet wide. Because of the orientation of fibers within the paper, it has a different strength when measured in the direction produced by the machine than when measured across, or at right angles to, the machine direction. Table 1.2 shows the measured values of

x_1 = density (grams/cubic centimeter)

x_2 = strength (pounds) in the machine direction

x_3 = strength (pounds) in the cross direction

A novel graphic presentation of these data appears in Figure 1.5, page 16. The scatter plots are arranged as the off-diagonal elements of a covariance array and box plots as the diagonal elements. The latter are on a different scale with this

Paper Quality Variables

Table 1.2 Paper-Quality Measurements

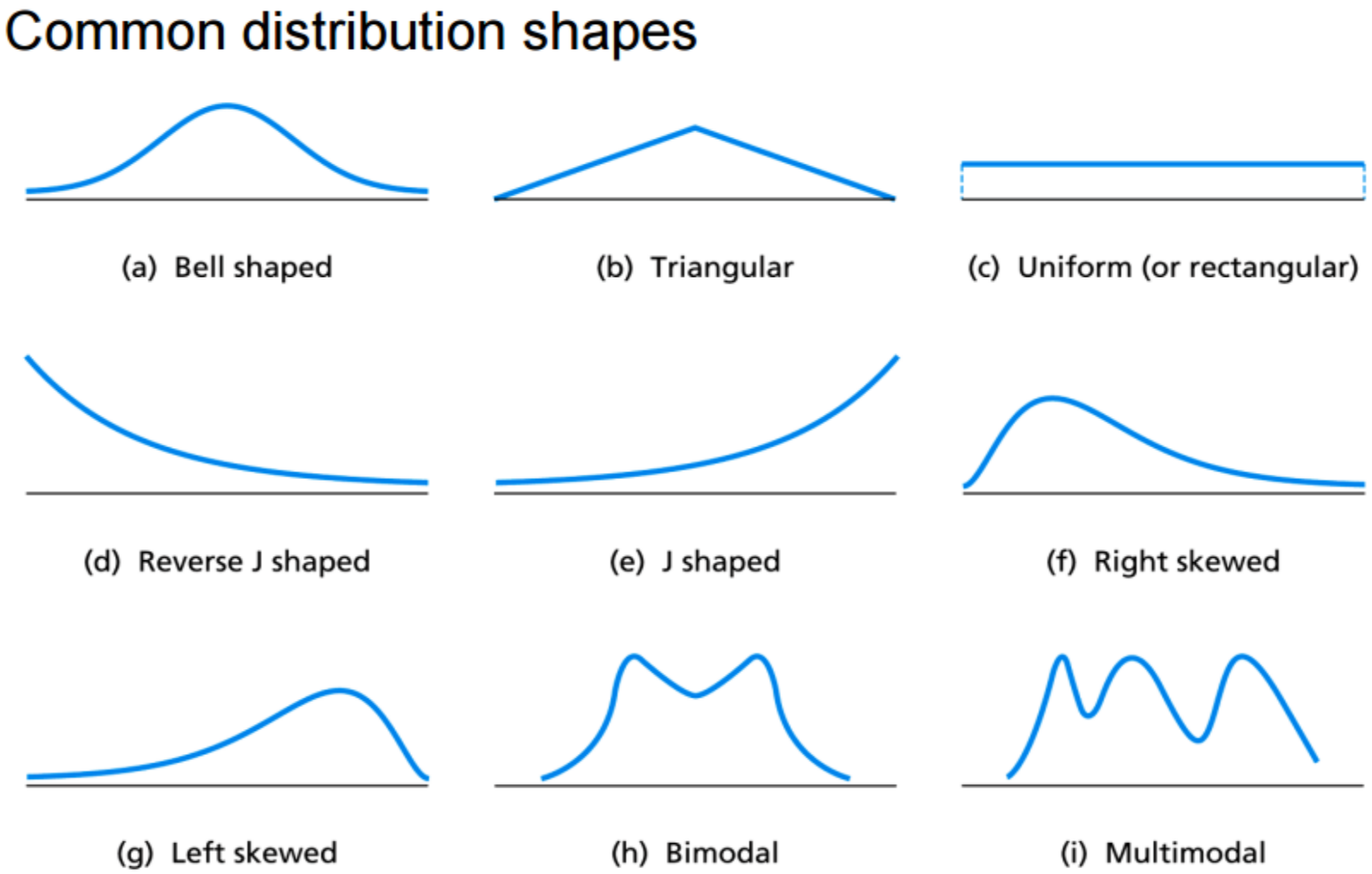
Specimen	Density	Strength	
		Machine direction	Cross direction
1	.801	121.41	70.42
2	.824	127.70	72.47
3	.841	129.20	78.20
4	.816	131.80	74.89
5	.840	135.10	71.21
6	.842	131.50	78.39
7	.820	126.70	69.02
8	.802	115.10	73.10
9	.828	130.80	79.28
10	.819	124.60	76.48
11	.826	118.31	70.25
12	.802	114.20	72.88
13	.810	120.30	68.23
14	.802	115.70	68.12
15	.832	117.51	71.62
16	.796	109.81	53.10
17	.759	109.10	50.85
18	.770	115.10	51.68

Specimen Paper Quality Variables

19	.759	118.31	50.60
20	.772	112.60	53.51
21	.806	116.20	56.53
22	.803	118.00	70.70.
23	.845	131.00	74.35
24	.822	125.70	68.29
25	.971	126.10	72.10
26	.816	125.80	70.64
27	.836	125.50	76.33
28	.815	127.80	76.75
29	.822	130.50	80.33
30	.822	127.90	75.68
31	.843	123.90	78.54
32	.824	124.10	71.91
33	.788	120.80	68.22
34	.782	107.40	54.42
35	.795	120.70	70.41
36	.805	121.91	73.68
37	.836	122.31	74.93
38	.788	110.60	53.52
39	.772	103.51	48.93
40	.776	110.71	53.67
41	.758	113.80	52.42

Source: Data courtesy of SONOCO Products Company.

Single Variable or Univariate Analysis



Paper Quality – Comparison Boxplot and Scatter Plot for Three Variables

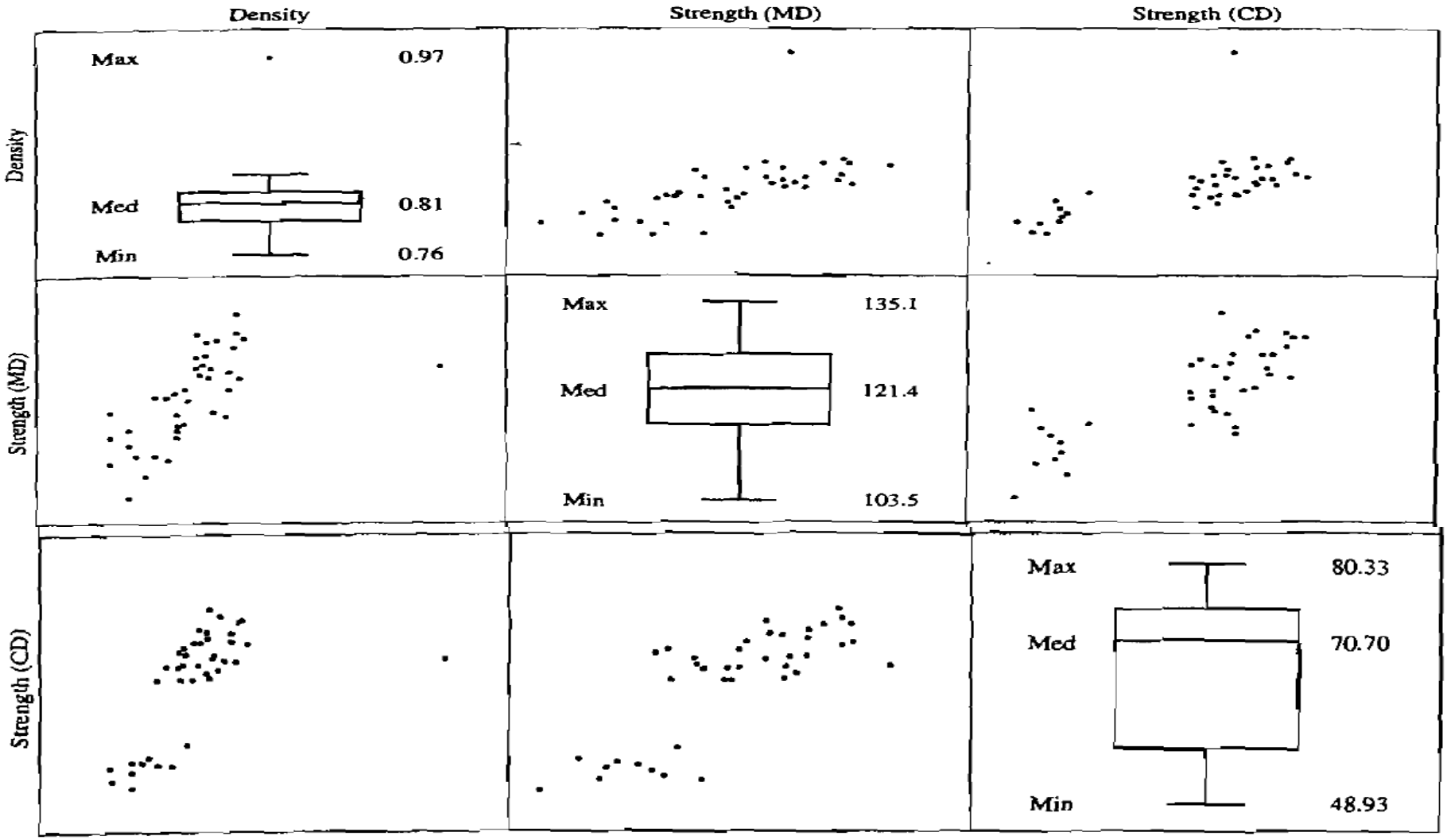
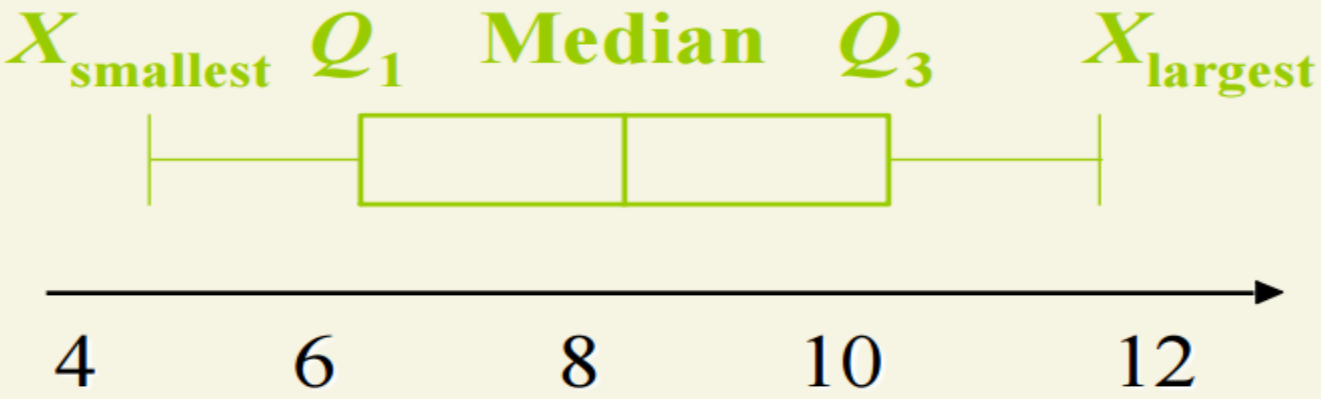


Figure 1.5 Scatter plots and boxplots of paper-quality data from Table 1.2.

Box Plot and 5-Number Summary

Box Plot

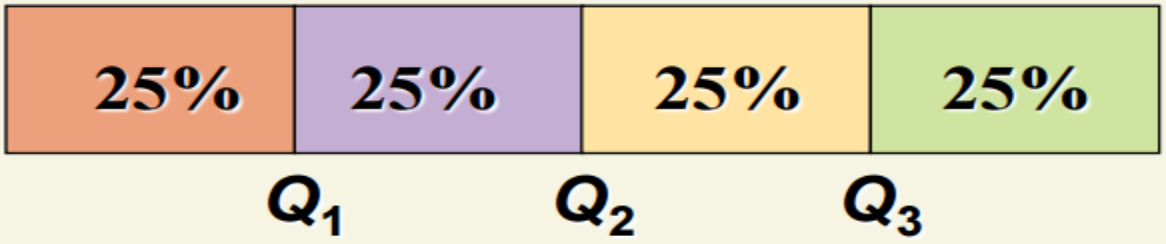
1. Graphical display of data using 5-number summary



Quartiles: Noncentral Tendency

Quartiles

Measure of *noncentral* tendency
Split ordered data into 4 quarters



Lower quartile Q_L is 25th percentile.

Middle quartile *m* is the median.

Upper quartile Q_U is 75th percentile.

Interquartile range: $\text{IQR} = Q_U - Q_L$

Detect Outliers with Boxplot (Univariate); Patters with Scatter Plot (Bivariate)

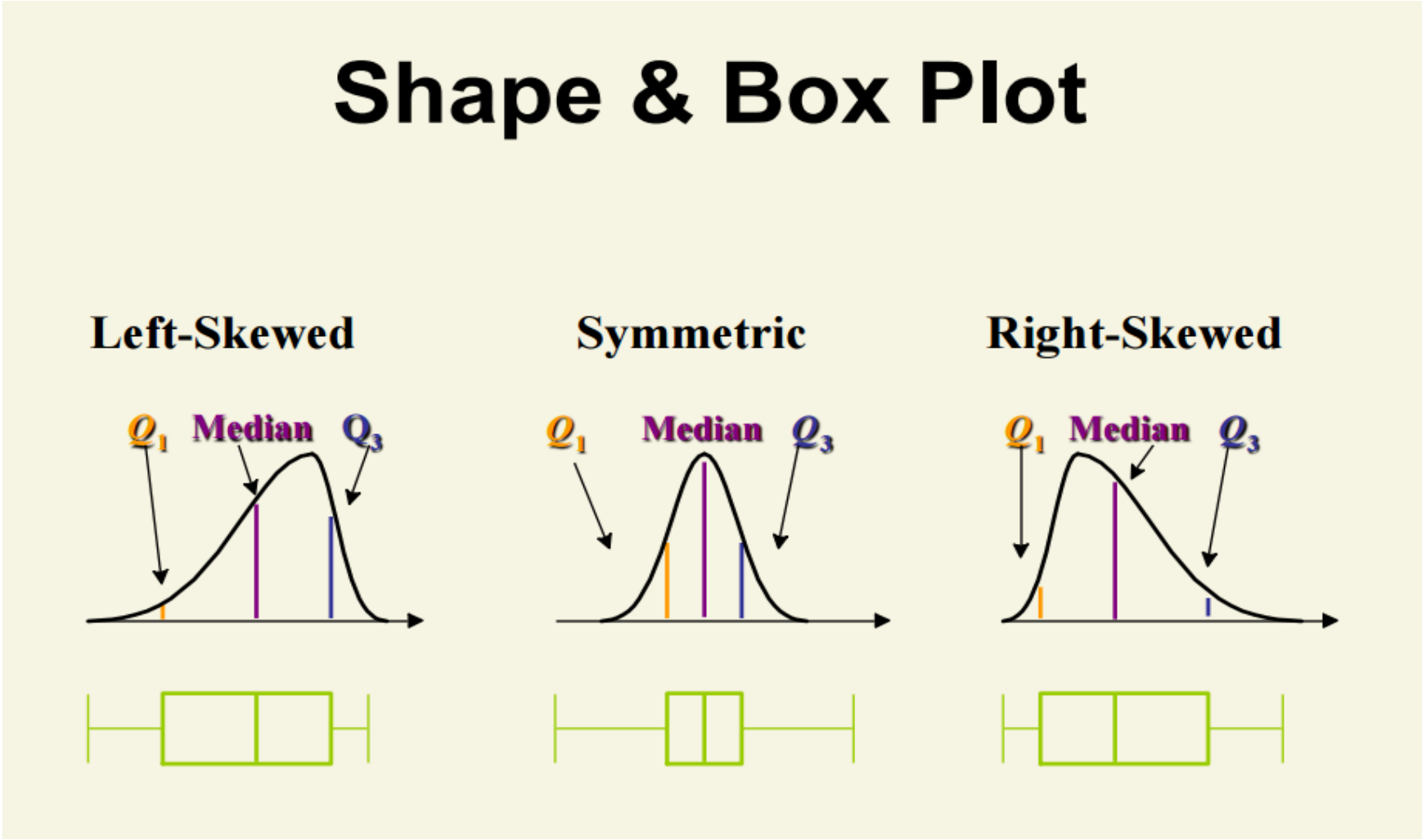
software, so we use only the overall shape to provide information on symmetry and possible outliers for each individual characteristic. The scatter plots can be inspected for patterns and unusual observations. In Figure 1.5, there is one unusual observation: the density of specimen 25. Some of the scatter plots have patterns suggesting that there are two separate clumps of observations.

These scatter plot arrays are further pursued in our discussion of new software graphics in the next section. ■

In the general multiresponse situation, p variables are simultaneously recorded on n items. Scatter plots should be made for pairs of important variables and, if the task is not too great to warrant the effort, for all pairs.

Limited as we are to a three-dimensional world, we cannot always picture an entire set of data. However, two further geometric representations of the data provide an important conceptual framework for viewing multivariable statistical methods. In cases where it is possible to capture the essence of the data in three dimensions, these representations can actually be graphed.

Box Plot and Normal Distribution vs. Skewed Distribution



P-Dimensional Scatter Plot

***n* Points in *p* Dimensions (*p*-Dimensional Scatter Plot).** Consider the natural extension of the scatter plot to p dimensions, where the p measurements

$$(x_{j1}, x_{j2}, \ldots, x_{jp})$$

on the jth item represent the coordinates of a point in p-dimensional space. The coordinate axes are taken to correspond to the variables, so that the jth point is x_{j1} units along the first axis, x_{j2} units along the second, ..., x_{jp} units along the pth axis. The resulting plot with n points not only will exhibit the overall pattern of variability, but also will show similarities (and differences) among the n items. Groupings of *items* will manifest themselves in this representation.

The next example illustrates a three-dimensional scatter plot.

Unstandardized vs. Standardized Values in Scatter Plot

Example 1.6 (Looking for lower-dimensional structure) A zoologist obtained measurements on $n = 25$ lizards known scientifically as *Cophosaurus texanus*. The weight, or mass, is given in grams while the snout-vent length (SVL) and hind limb span (HLS) are given in millimeters. The data are displayed in Table 1.3.

Although there are three size measurements, we can ask whether or not most of the variation is primarily restricted to two dimensions or even to one dimension.

To help answer questions regarding reduced dimensionality, we construct the three-dimensional scatter plot in Figure 1.6. Clearly most of the variation is scatter about a one-dimensional straight line. Knowing the position on a line along the major axes of the cloud of poinfs would be almost as good as knowing the three measurements Mass, SVL, and HLS.

However, this kind of analysis can be misleading if one variable has a much larger variance than the others. Consequently, we first calculate the standardized values, $z_{jk} = (x_{jk} - \bar{x}_k)/\sqrt{s_{kk}}$, so the variables contribute equally to the variation

Lizard Size Data with 2 Variables

Table 1.3 Lizard Size Data

Lizard	Mass	SVL	HLS	Lizard	Mass	SVL	HLS
1	5.526	59.0	113.5	14	10.067	73.0	136.5
2	10.401	75.0	142.0	15	10.091	73.0	135.5
3	9.213	69.0	124.0	16	10.888	77.0	139.0
4	8.953	67.5	125.0	17	7.610	61.5	118.0
5	7.063	62.0	129.5	18	7.733	66.5	133.5
6	6.610	62.0	123.0	19	12.015	79.5	150.0
7	11.273	74.0	140.0	20	10.049	74.0	137.0
8	2.447	47.0	97.0	21	5.149	59.5	116.0
9	15.493	86.5	162.0	22	9.158	68.0	123.0
10	9.004	69.0	126.5	23	12.132	75.0	141.0
11	8.199	70.5	136.0	24	6.978	66.5	117.0
12	6.601	64.5	116.0	25	6.890	63.0	117.0
13	7.622	67.5	135.0				

Source: Data courtesy of Kevin E. Bonine.

Scatter Plot - Unstandardized

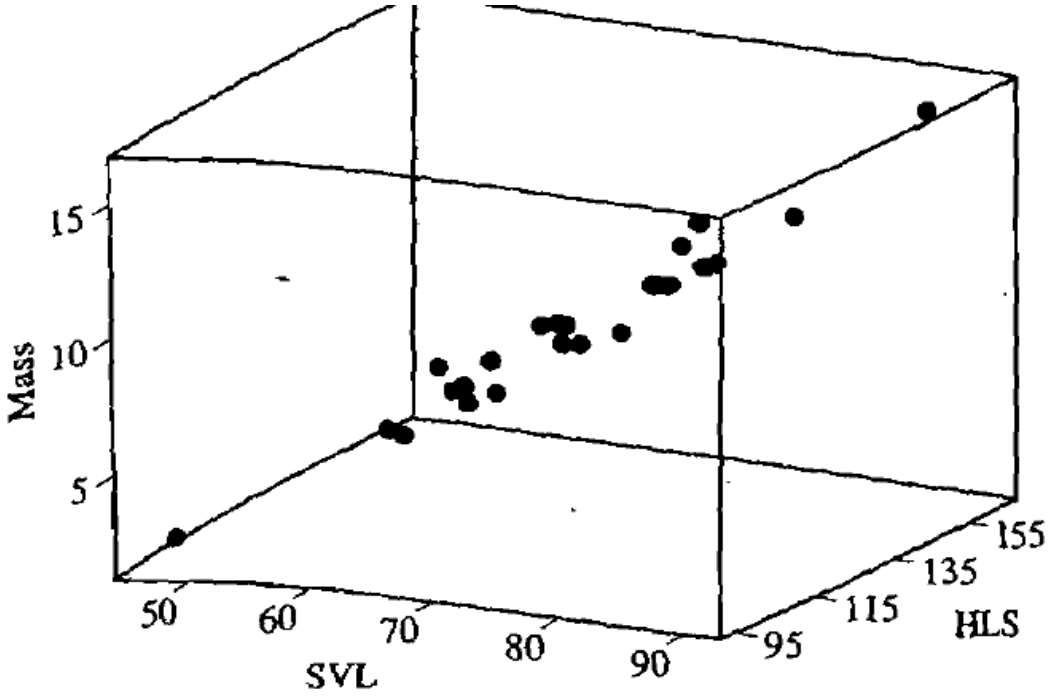


Figure 1.6 3D scatter plot of lizard data from Table 1.3.

in the scatter plot. Figure 1.7 gives the three-dimensional scatter plot for the standardized variables. Most of the variation can be explained by a single variable determined by a line through the cloud of points.

Scatter Plot - Standardized

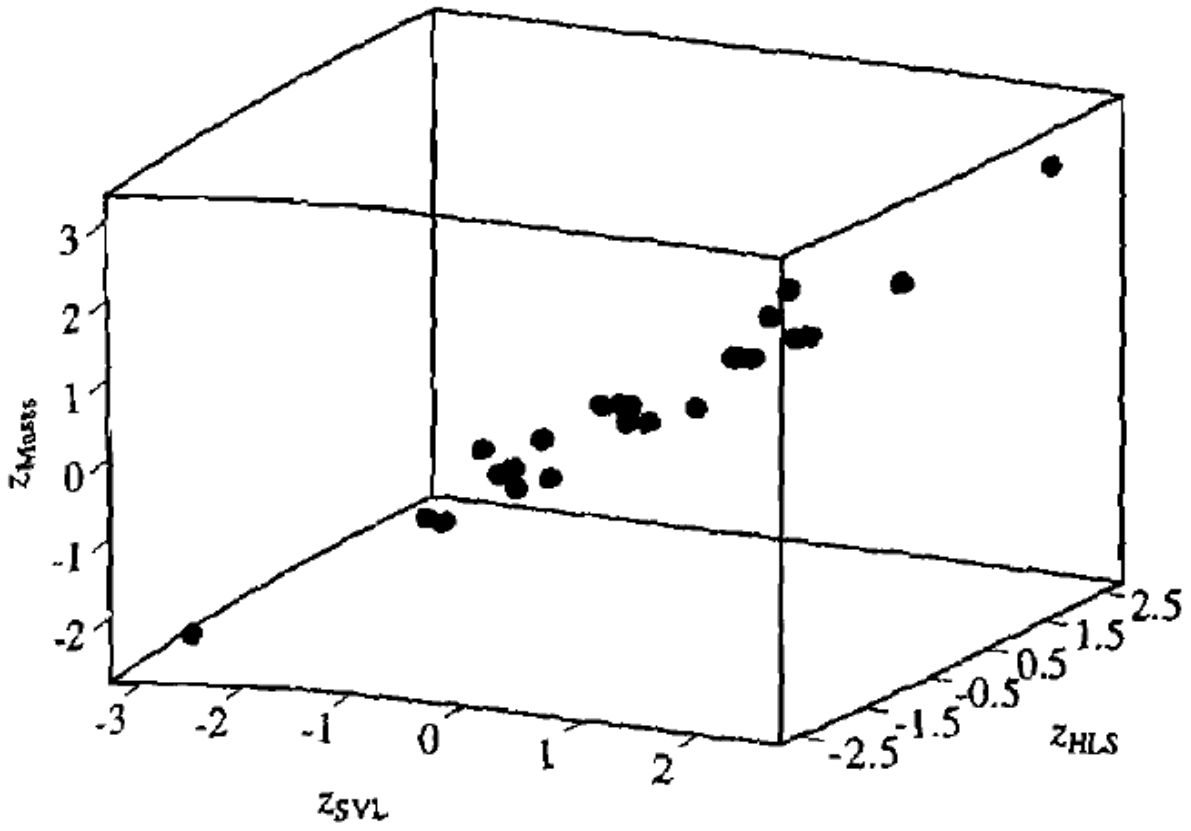


Figure 1.7 3D scatter plot of standardized lizard data. ■

A three-dimensional scatter plot can often reveal group structure.

Comparison of Male and Female Lizard Data Using Scatter Plot

Example 1.7 (Looking for group structure in three dimensions) Referring to Example 1.6, it is interesting to see if male and female lizards occupy different parts of the three-dimensional space containing the size data. The gender, by row, for the lizard data in Table 1.3 are

f m f f m f m f m f m f m

m m m f m m m f f m f f

Comparison Male and Female Lizards by Body Size (Mass and Length)

Figure 1.8 repeats the scatter plot for the original variables but with males marked by solid circles and females by open circles. Clearly, males are typically larger than females.

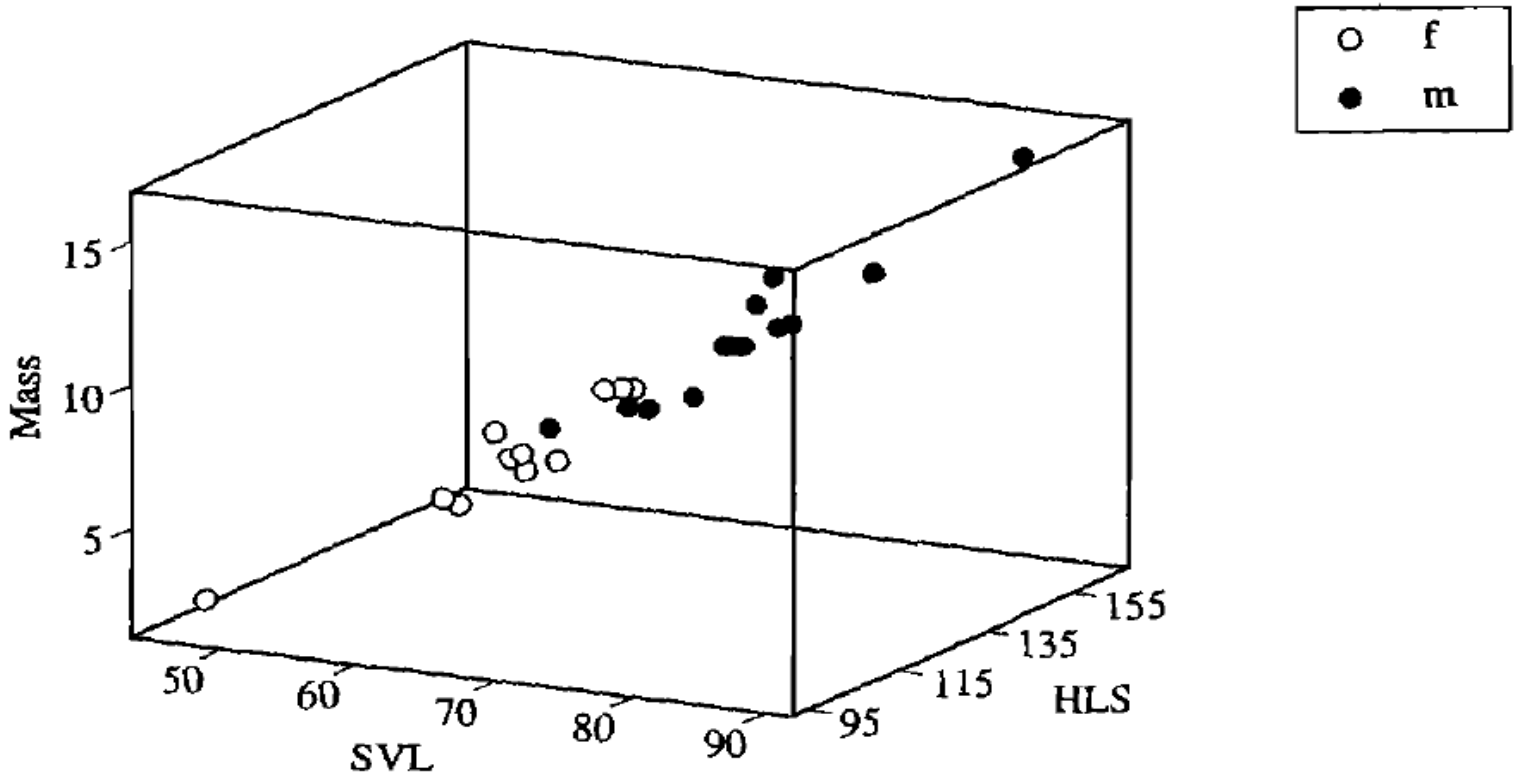


Figure 1.8 3D scatter plot of male and female lizards.

P Points in n Dimensions – Variable as Vector in an Array

***p* Points in *n* Dimensions.** The n observations of the p variables can also be regarded as p points in n-dimensional space. Each column of $\mathbf{X}$ determines one of the points. The ith column,

$$\begin{bmatrix} x_{1i} \\ x_{2i} \\ \vdots \\ x_{ni} \end{bmatrix}$$

consisting of all n measurements on the ith variable, determines the ith point.

In Chapter 3, we show how the closeness of points in n dimensions can be related to measures of association between the corresponding *variables*.

Data Display – Picture Snapshot in Time

1.4 Data Displays and Pictorial Representations

The rapid development of powerful personal computers and workstations has led to a proliferation of sophisticated statistical software for data analysis and graphics. It is often possible, for example, to sit at one's desk and examine the nature of multidimensional data with clever computer-generated pictures. These pictures are valuable aids in understanding data and often prevent many false starts and subsequent inferential problems.

As we shall see in Chapters 8 and 12, there are several techniques that seek to represent p-dimensional observations in few dimensions such that the original distances (or similarities) between pairs of observations are (nearly) preserved. In general, if multidimensional observations can be represented in two dimensions, then outliers, relationships, and distinguishable groupings can often be discerned by eye. We shall discuss and illustrate several methods for displaying multivariate data in two dimensions. One good source for more discussion of graphical methods is [11].

Comparison of Several 2-Dimensional Scatter Plots

Linking Multiple Two-Dimensional Scatter Plots

One of the more exciting new graphical procedures involves electronically connecting many two-dimensional scatter plots.

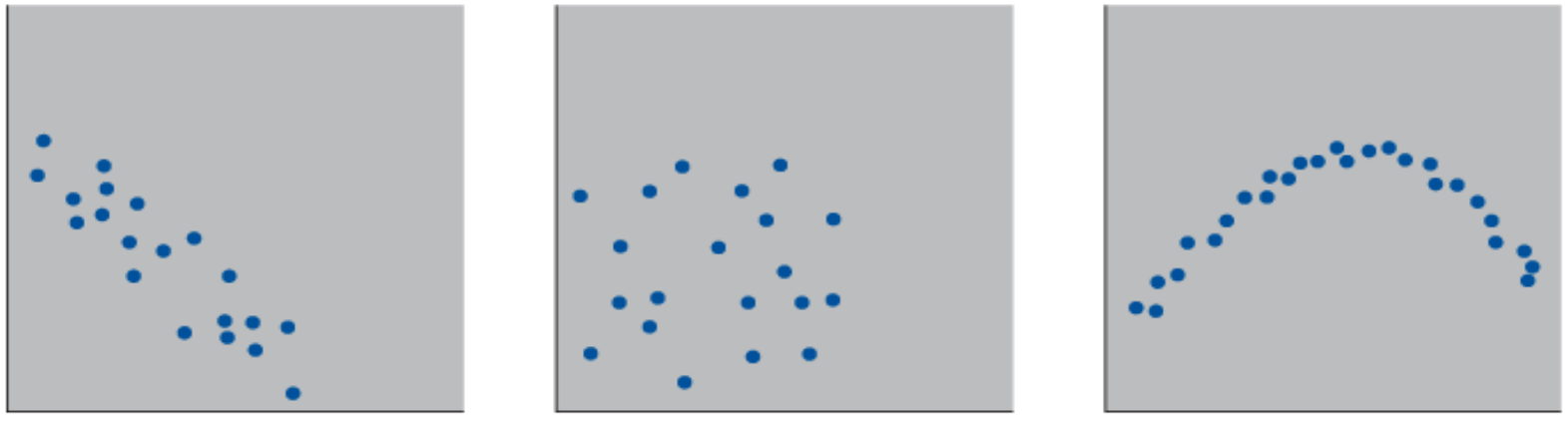
Example 1.8 (Linked scatter plots and brushing) To illustrate *linked* two-dimensional scatter plots, we refer to the paper-quality data in Table 1.2. These data represent measurements on the variables x_1 = density, x_2 = strength in the machine direction, and x_3 = strength in the cross direction. Figure 1.9 shows two-dimensional scatter plots for pairs of these variables organized as a 3×3 array. For example, the picture in the upper left-hand corner of the figure is a scatter plot of the pairs of observations (x_1, x_3). That is, the x_1 values are plotted along the horizontal axis, and the x_3 values are plotted along the vertical axis. The lower right-hand corner of the figure contains a scatter plot of the observations (x_3, x_1). That is, the axes are reversed. Corresponding interpretations hold for the other scatter plots in the figure. Notice that the variables and their three-digit ranges are indicated in the boxes along the SW–NE diagonal. The

Purpose: Compare Scatter Plots for Outliers and Effect on Correlation

operation of marking (*selecting*), the obvious outlier in the (x_1, x_3) scatter plot of Figure 1.9 creates Figure 1.10(a), where the outlier is labeled as specimen 25 and the same data point is highlighted in all the scatter plots. Specimen 25 also appears to be an outlier in the (x_1, x_2) scatter plot but not in the (x_2, x_3) scatter plot. The operation of *deleting* this specimen leads to the modified scatter plots of Figure 1.10(b).

From Figure 1.10, we notice that some points in, for example, the (x_2, x_3) scatter plot seem to be disconnected from the others. Selecting these points, using the (dashed) rectangle (see page 22), highlights the selected points in all of the other scatter plots and leads to the display in Figure 1.11(a). Further checking revealed that specimens 16–21, specimen 34, and specimens 38–41 were actually specimens

Calculation of Correlation Coefficient or r: Based on Standard Deviation x and y



(d) $r = -.9$ (e) $r = 0$ (f) $r = 0$

Figure 3 Correspondence between the values of r and the amount of scatter.

4.1 CALCULATION OF *r*

The sample correlation coefficient quantifies the association between two numerically valued characteristics. It is calculated from n pairs of observations on the two characteristics

$$(x_1, y_1),\ (x_2, y_2),\ \ldots,\ (x_n, y_n)$$

The correlation coefficient is best interpreted in terms of the **standardized observations**, or **sample z values**

$$\frac{\text{Observation} - \text{Sample mean}}{\text{Sample standard deviation}} = \frac{x_i - \overline{x}}{s_x}$$

Scatter Plots for Paper Quality

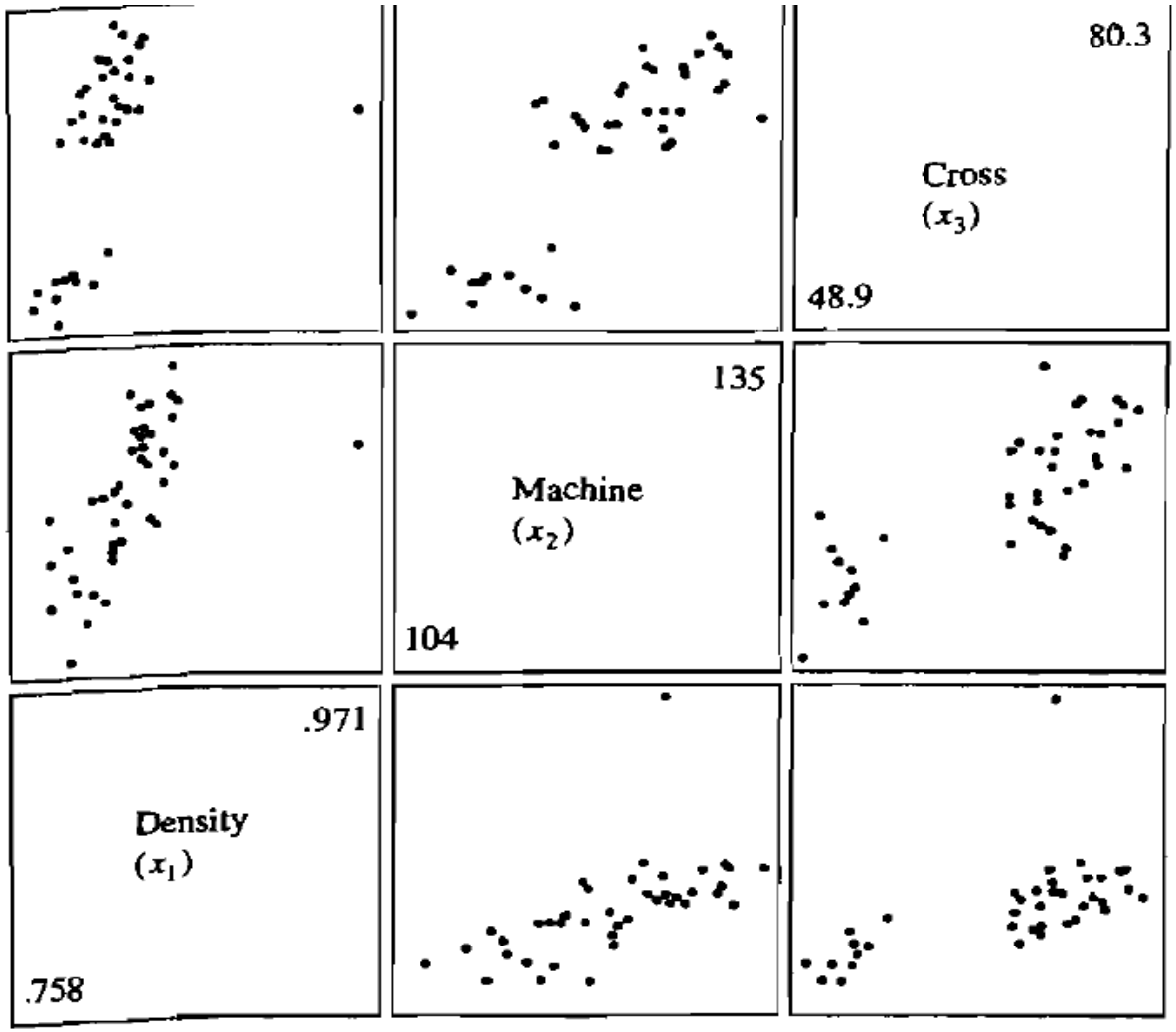
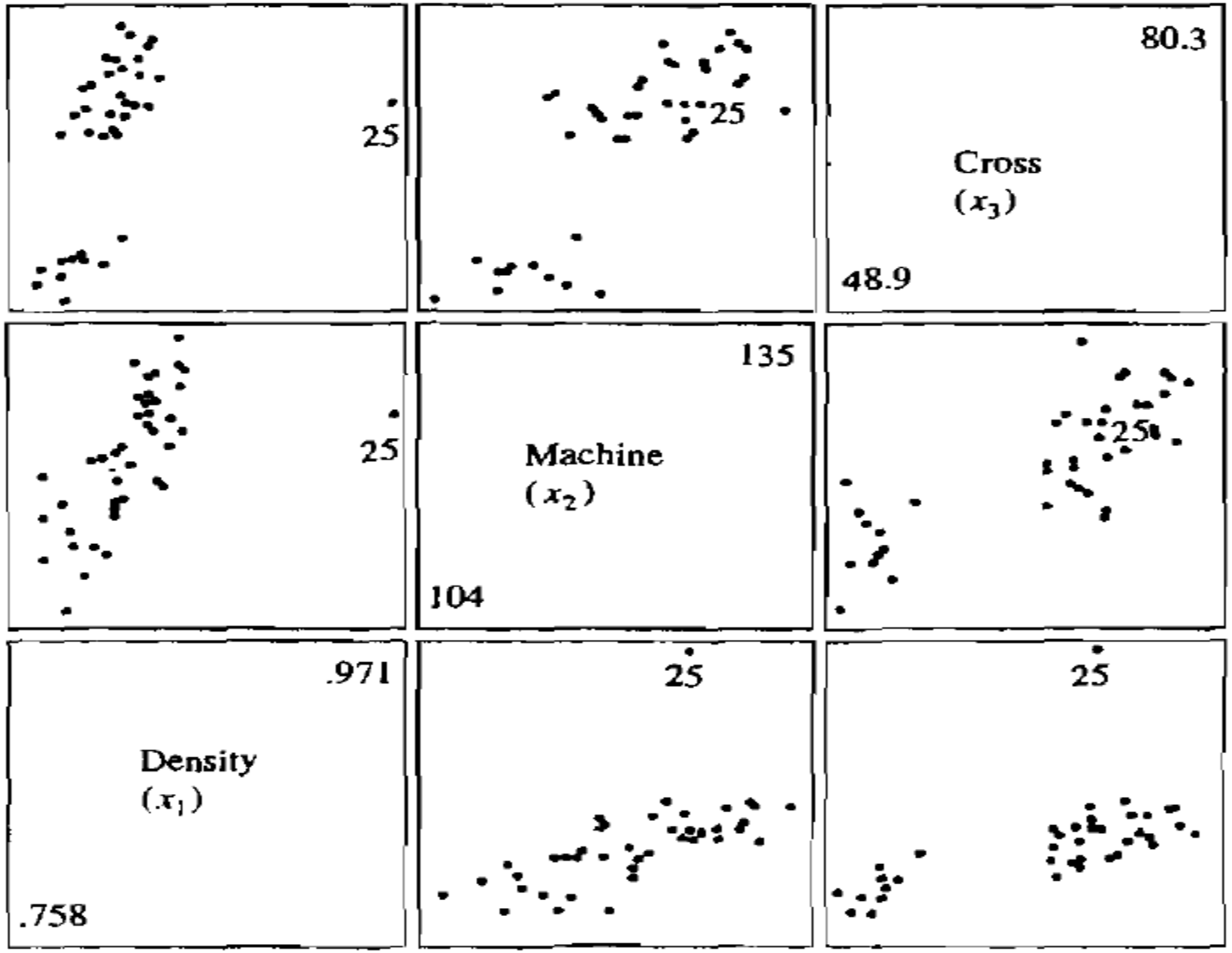


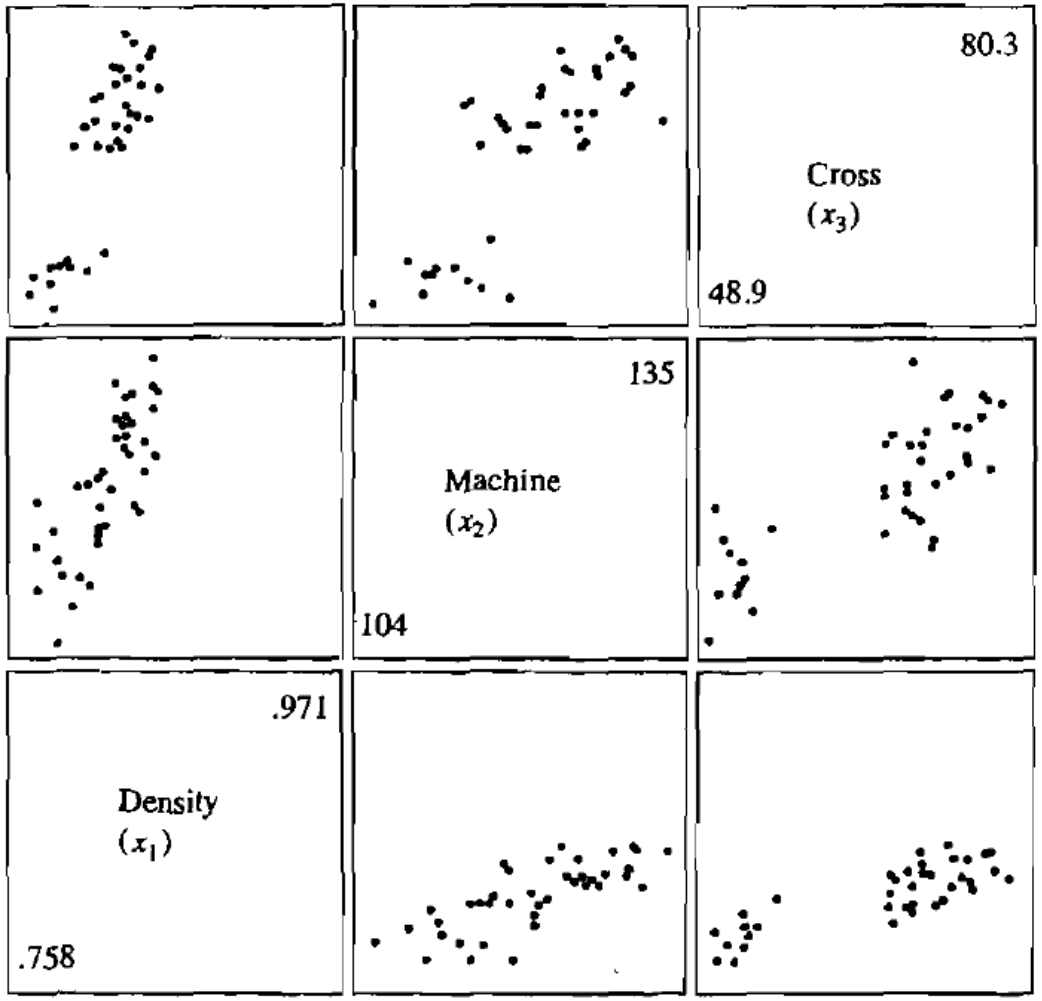
Figure 1.9 Scatter plots for the paper-quality data of Table 1.2.

Scatter Plots – Includes All Variables



(a)

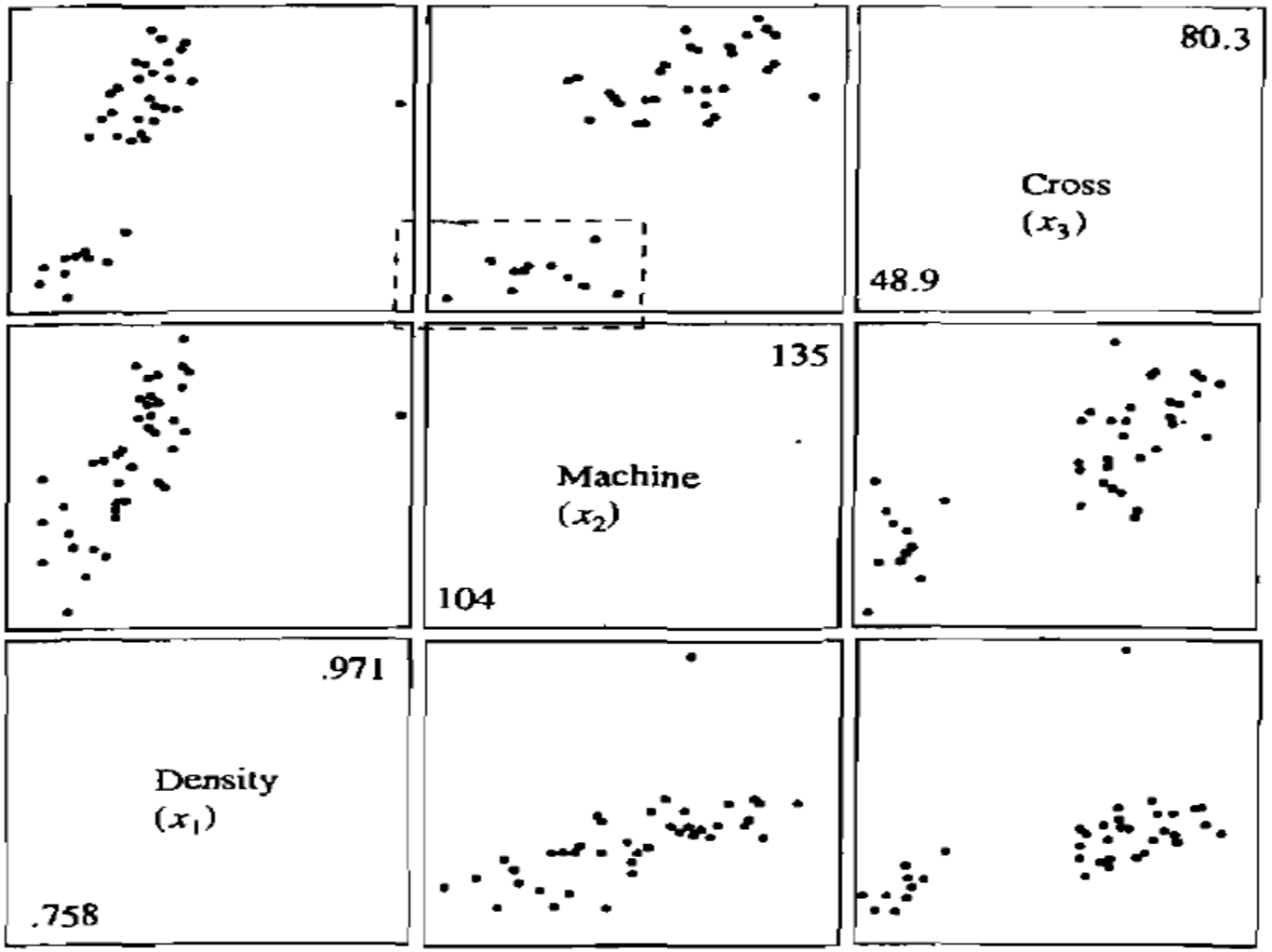
Outliers Selected and Deleted (a) vs. (b)



(b)

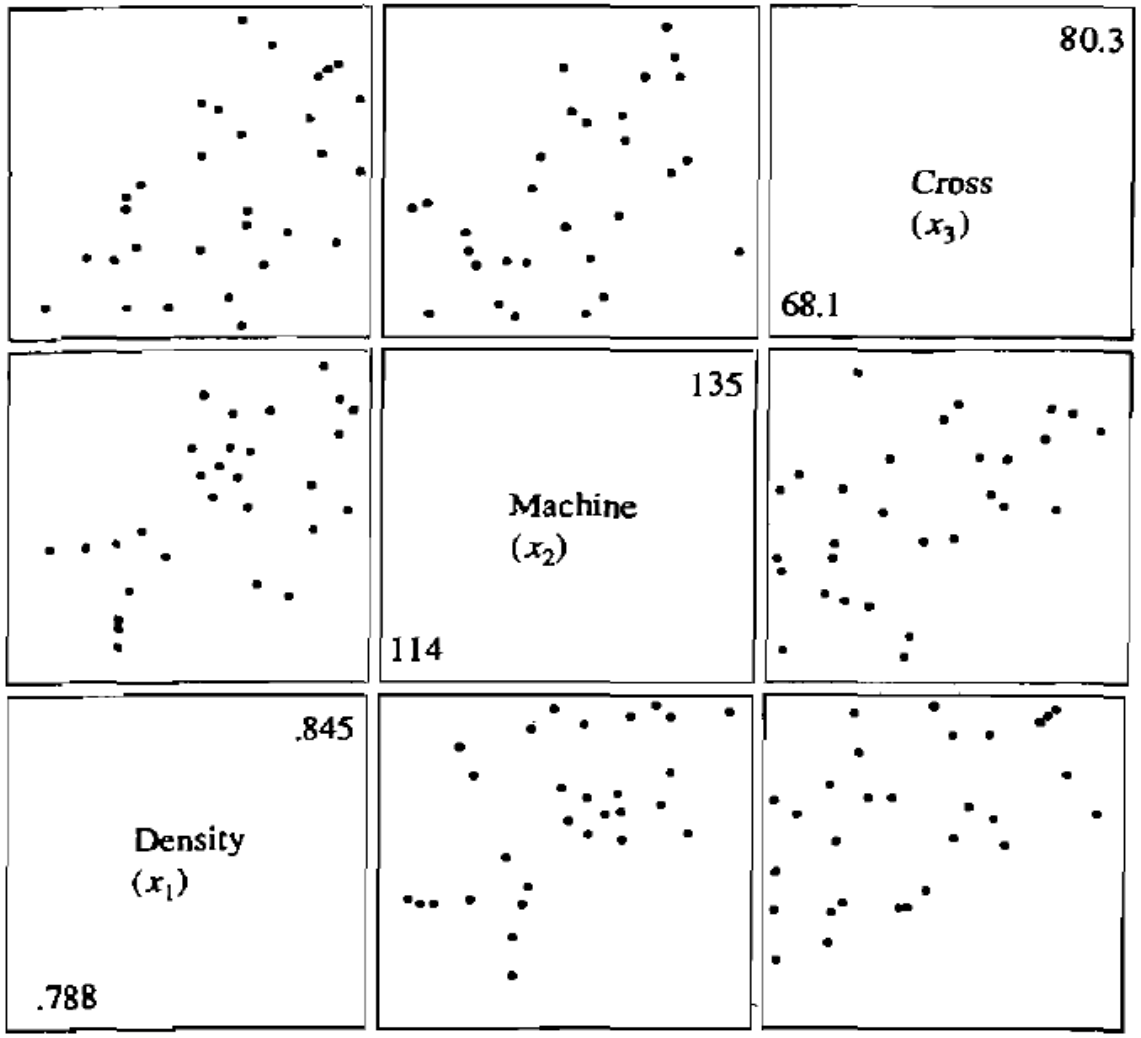
Figure 1.10 Modified scatter plots for the paper-quality data with outlier (25) (a) selected and (b) deleted.

Scatter Plot – Group of Points Selected



(a)

Scatter Plot – Group of Points Deleted



(b)

Figure 1.11 Modified scatter plots with (a) group of points selected and (b) points, including specimen 25, deleted and the scatter plots rescaled.

Brushing – Deleting a Group of Points from a Scatter Plot (previous slide)

from an older roll of paper that was included in order to have enough plies in the cardboard being manufactured. Deleting the outlier and the cases corresponding to the older paper and adjusting the ranges of the remaining observations leads to the scatter plots in Figure 1.11(b).

The operation of highlighting points corresponding to a selected range of one of the variables is called *brushing*. Brushing could begin with a rectangle, as in Figure 1.11(a), but then the brush could be moved to provide a sequence of highlighted points. The process can be stopped at any time to provide a snapshot of the current situation. ■

Scatter Plot – Detect Patterns – Data Could be Divided into Two Groups

Scatter plots like those in Example 1.8 are extremely useful aids in data analysis. Another important new graphical technique uses software that allows the data analyst to view high-dimensional data as slices of various three-dimensional perspectives. This can be done dynamically and continuously until informative views are obtained. A comprehensive discussion of dynamic graphical methods is available in [1]. A strategy for on-line multivariate exploratory graphical analysis, motivated by the need for a routine procedure for searching for structure in multivariate data, is given in [32].

Scatter Plot – Rotation Three-Dimensional Coordinate Axes

Example 1.9 (Rotated plots in three dimensions) Four different measurements of lumber stiffness are given in Table 4.3, page 186. In Example 4.14, specimen (board) 16 and possibly specimen (board) 9 are identified as unusual observations. Figures 1.12(a), (b), and (c) contain perspectives of the stiffness data in the x_1, x_2, x_3 space. These views were obtained by continually rotating and turning the three-dimensional coordinate axes. Spinning the coordinate axes allows one to get a better

(a) Outliers clear. (b) Outliers masked.

Scatter Plot Rotation – Detection of Leverage Points or Outliers

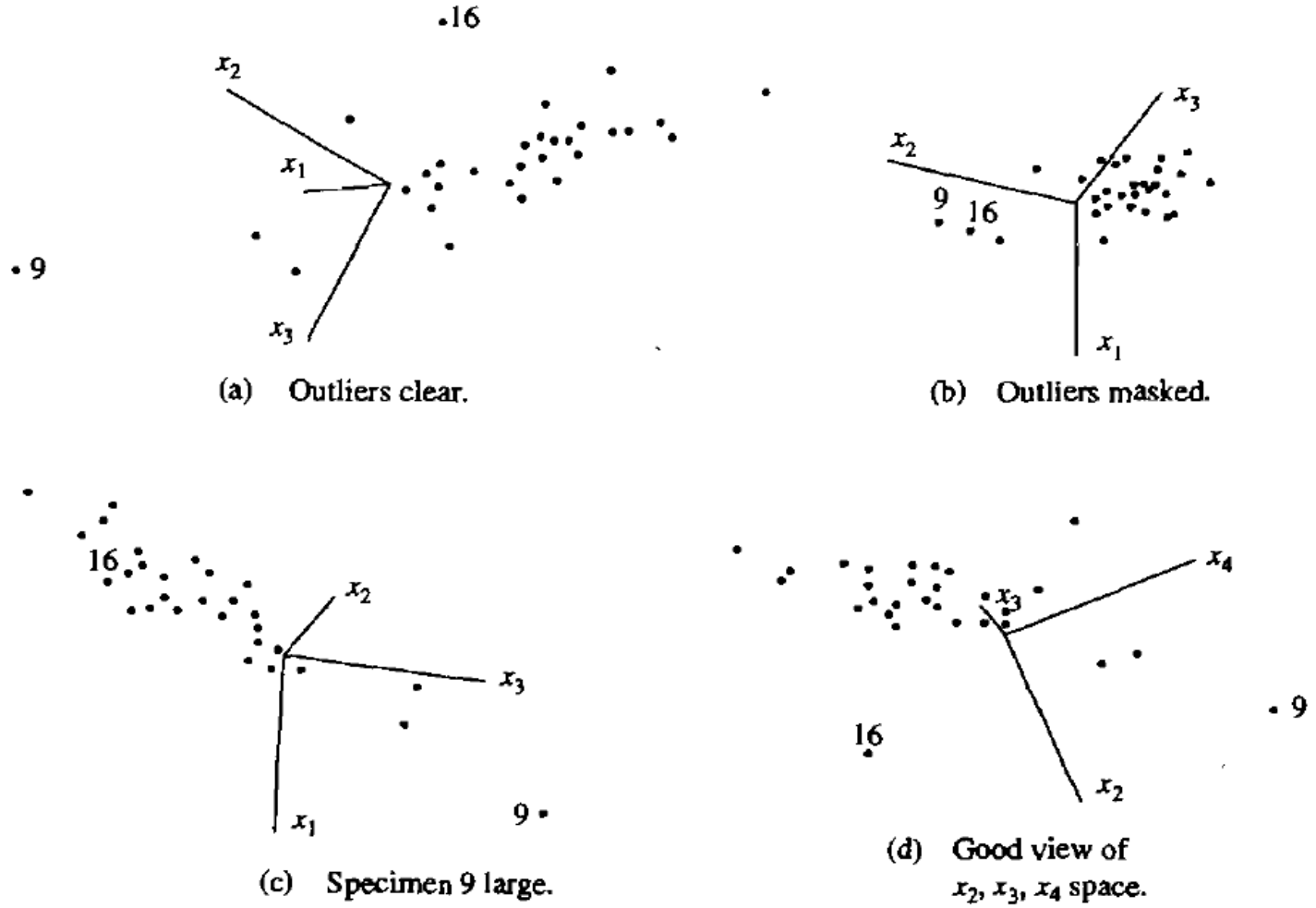


Figure 1.12 Three-dimensional perspectives for the lumber stiffness data.

Scatter Plot – Rotation Three-Dimensional Coordinate Axes – Influential Observations – To Eliminate or Transform Data Set?

understanding of the three-dimensional aspects of the data. Figure 1.12(d) gives one picture of the stiffness data in x_2, x_3, x_4 space. Notice that Figures 1.12(a) and (d) visually confirm specimens 9 and 16 as outliers. Specimen 9 is very large in all three coordinates. A counterclockwiselike rotation of the axes in Figure 1.12(a) produces Figure 1.12(b), and the two unusual observations are masked in this view. A further spinning of the x_2, x_3 axes gives Figure 1.12(c); one of the outliers (16) is now hidden.

Additional insights can sometimes be gleaned from visual inspection of the slowly spinning data. It is this dynamic aspect that statisticians are just beginning to understand and exploit. ■

Plots like those in Figure 1.12 allow one to identify readily observations that do not conform to the rest of the data and that may heavily influence inferences based on standard data-generating models.

Growth Curve Analysis: By Year (unit of analysis is year)

Graphs of Growth Curves

When the height of a young child is measured at each birthday, the points can be plotted and then connected by lines to produce a graph. This is an example of a *growth curve*. In general, repeated measurements of the same characteristic on the same unit or subject can give rise to a growth curve if an increasing, decreasing, or even an increasing followed by a decreasing, pattern is expected.

Example 1.10 (Arrays of growth curves) The Alaska Fish and Game Department monitors grizzly bears with the goal of maintaining a healthy population. Bears are shot with a dart to induce sleep and weighed on a scale hanging from a tripod. Measurements of length are taken with a steel tape. Table 1.4 gives the weights (wt) in kilograms and lengths (lngth) in centimeters of seven female bears at 2, 3, 4, and 5 years of age.

First, for each bear, we plot the weights versus the ages and then connect the weights at successive years by straight lines. This gives an approximation to growth curve for weight. Figure 1.13 shows the growth curves for all seven bears. The noticeable exception to a common pattern is the curve for bear 5. Is this an outlier or just natural variation in the population? In the field, bears are weighed on a scale that

Female Bears Scatter Plot: Weight by Age or Year – Trend Line by Year

First, for each bear, we plot the weights versus the ages and then connect the weights at successive years by straight lines. This gives an approximation to growth curve for weight. Figure 1.13 shows the growth curves for all seven bears. The noticeable exception to a common pattern is the curve for bear 5. Is this an outlier or just natural variation in the population? In the field, bears are weighed on a scale that

Table 1.4 Female Bear Data

Bear	Wt2	Wt3	Wt4	Wt5	Lngth 2	Lngth 3	Lngth 4	Lngth 5
1	48	59	95	82	141	157	168	183
2	59	68	102	102	140	168	174	170
3	61	77	93	107	145	162	172	177
4	54	43	104	104	146	159	176	171
5	100	145	185	247	150	158	168	175
6	68	82	95	118	142	140	178	189
7	68	95	109	111	139	171	176	175

Source: Data courtesy of H. Roberts.

Growth Curve of Female Bears(Each Line Represents One Bear's Trajectory)

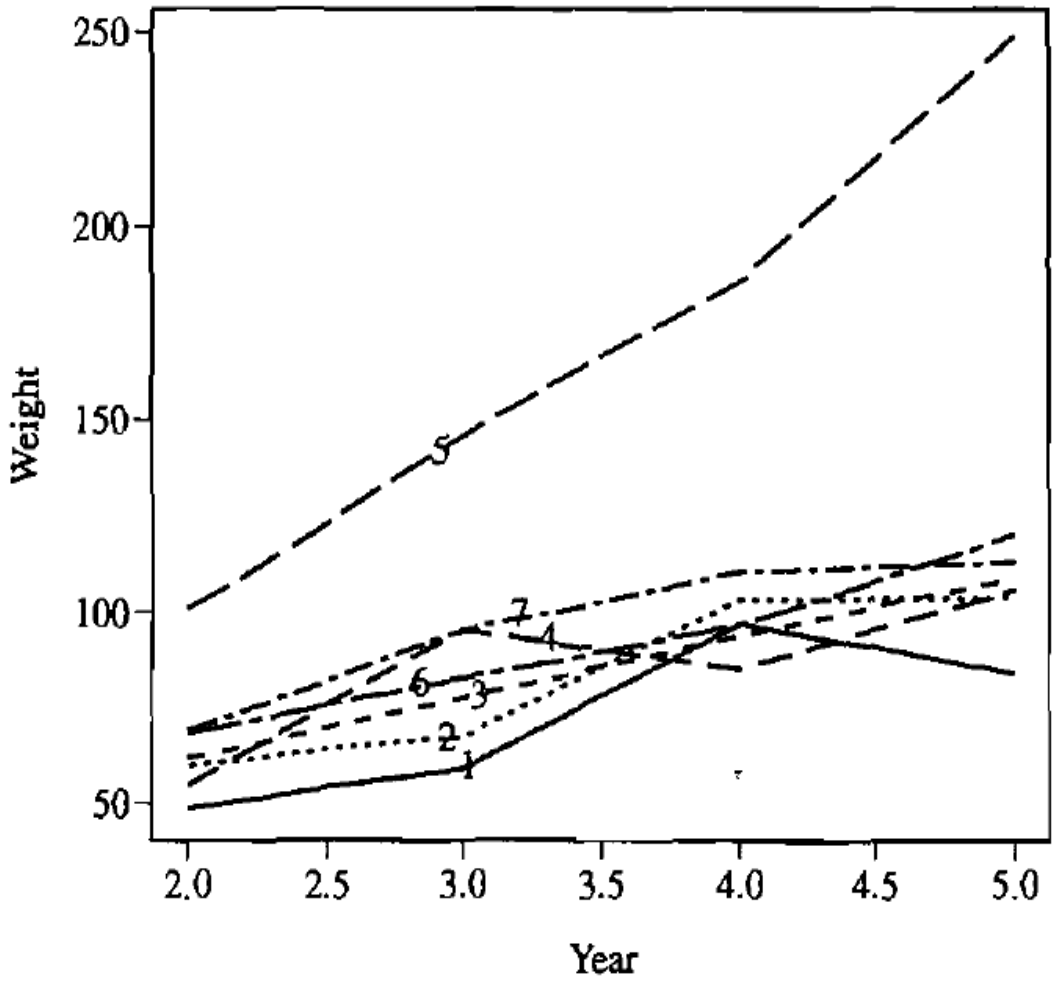


Figure 1.13 Combined growth curves for weight for seven female grizzly bears.

Combined vs. Individual Growth Curve Trend Lines for Each Bear

Because it can be difficult to inspect visually the individual growth curves in a combined plot, the individual curves should be replotted in an array where similarities and differences are easily observed. Figure 1.14 gives the array of seven curves for weight. Some growth curves look linear and others quadratic.

Individual Growth Curves for Each Single Bear in Study (n = 7)

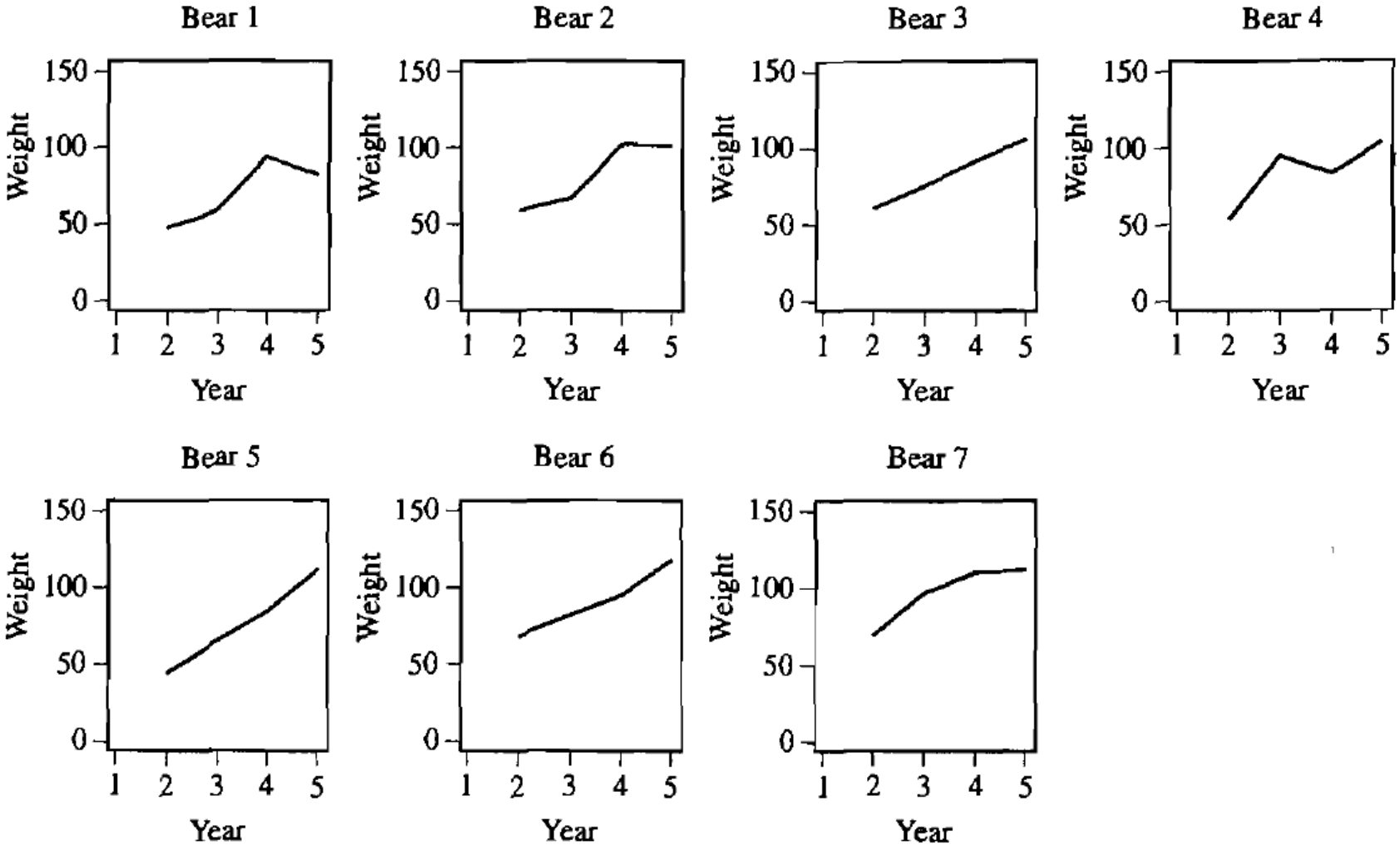
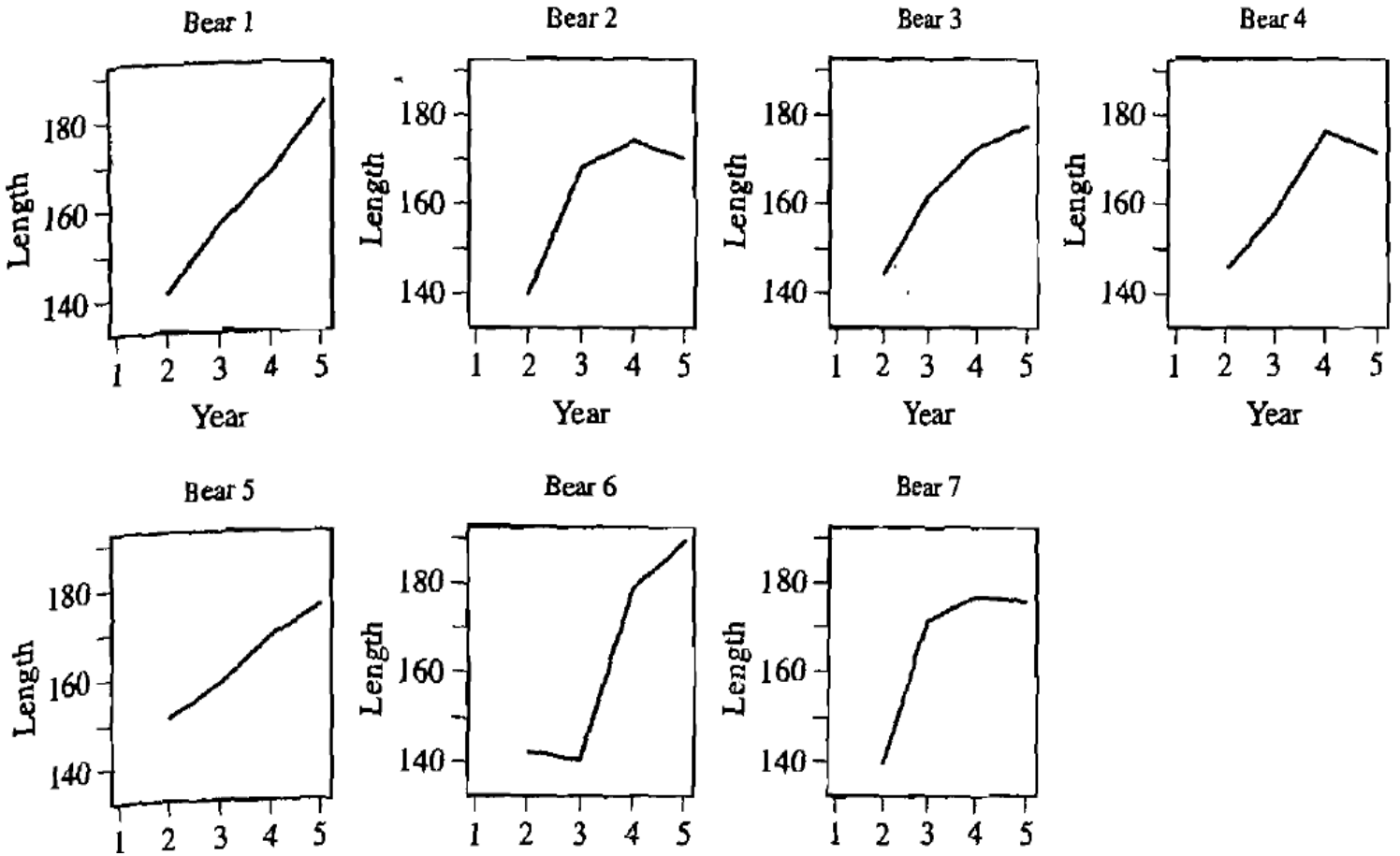


Figure 1.14 Individual growth curves for weight for female grizzly bears.

Growth Curve for Each Female Grizzly Bear's Height or Length (n = 7)

Figure 1.15 gives a growth curve array for length. One bear seemed to get shorter from 2 to 3 years old, but the researcher knows that the steel tape measurement of length can be thrown off by the bear's posture when sedated.



Other Pictorial Displays of Multivariate Data: Stars and Chernoff Faces

We now turn to two popular pictorial representations of multivariate data in two dimensions: stars and Chernoff faces.

Stars

Suppose each data unit consists of nonnegative observations on $p \geq 2$ variables. In two dimensions, we can construct circles of a fixed (reference) radius with p equally spaced rays emanating from the center of the circle. The lengths of the rays represent the values of the variables. The ends of the rays can be connected with straight lines to form a star. Each star represents a multivariate observation, and the stars can be grouped according to their (subjective) similarities.

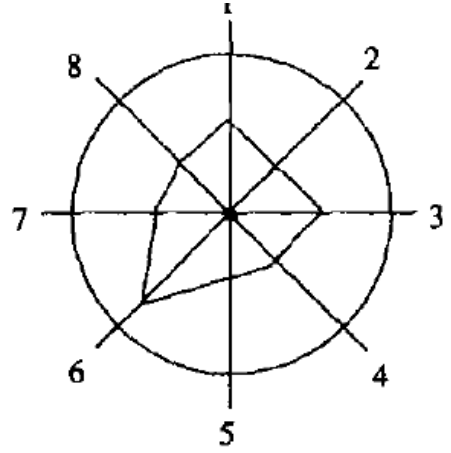
It is often helpful, when constructing the stars, to standardize the observations. In this case some of the observations will be negative. The observations can then be reexpressed so that the center of the circle represents the smallest standardized observation within the entire data set.

Electric Utility Companies Variables – Displayed as Stars

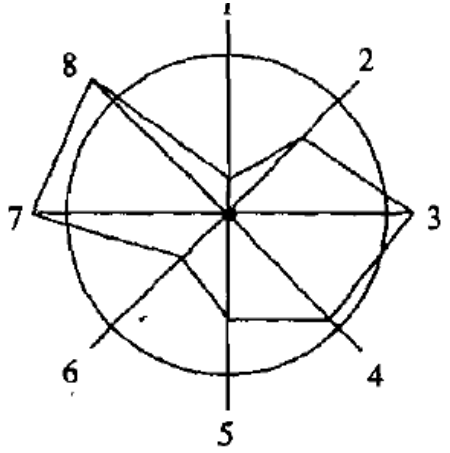
Example 1.11 (Utility data as stars) Stars representing the first 5 of the 22 public utility firms in Table 12.4, page 688, are shown in Figure 1.16. There are eight variables; consequently, the stars are distorted octagons.

Stars and Geo-Spatial Analysis – Service Area Covered by Utility Co.

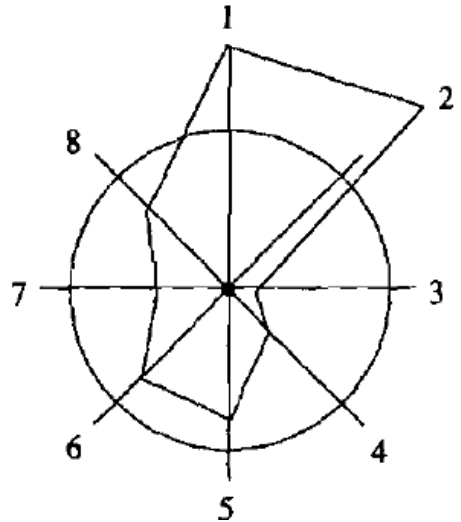
Arizona Public Service (1)



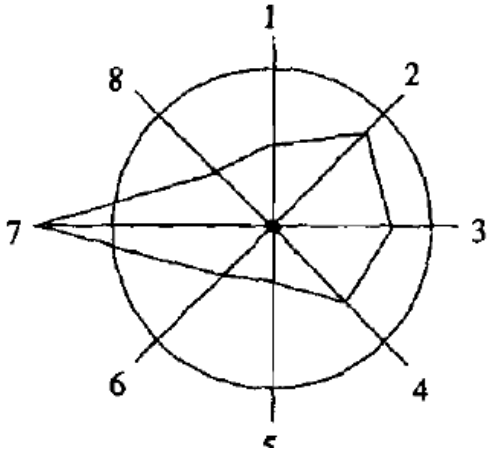
Boston Edison Co. (2)



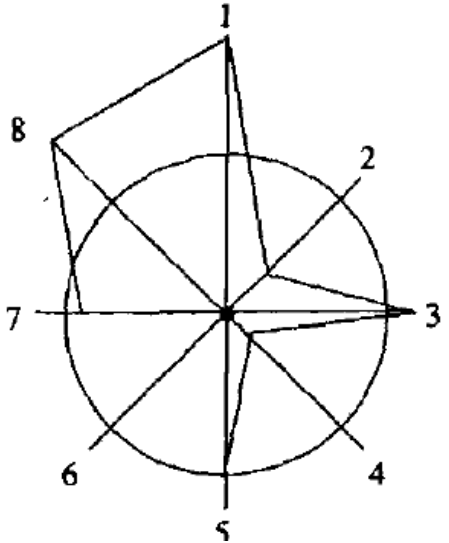
Central Louisiana Electric Co. (3)



Commonwealth Edison Co. (4)



Consolidated Edison Co. (NY) (5)



Electrical Utility Companies: Sales in Kilowatt Hours and Total Fuel Costs in Cents Per Kilowatt Hour

The observations on all variables were standardized. Among the first five utilities, the smallest standardized observation for any variable was −1.6. Treating this value as zero, the variables are plotted on identical scales along eight equiangular rays originating from the center of the circle. The variables are ordered in a clockwise direction, beginning in the 12 o'clock position.

At first glance, none of these utilities appears to be similar to any other. However, because of the way the stars are constructed, each variable gets equal weight in the visual impression. If we concentrate on the variables 6 (sales in kilowatt-hour [kWh] use per year) and 8 (total fuel costs in cents per kWh), then Boston Edison and Consolidated Edison are similar (small variable 6, large variable 8), and Arizona Public Service, Central Louisiana Electric, and Commonwealth Edison are similar (moderate variable 6, moderate variable 8). ■

Chernoff Faces – Emoticon or Emoji??

Chernoff Faces

People react to faces. Chernoff [4] suggested representing p-dimensional observations as a two-dimensional face whose characteristics (face shape, mouth curvature, nose length, eye size, pupil position, and so forth) are determined by the measurements on the p variables.

As originally designed, Chernoff faces can handle up to 18 variables. The assignment of variables to facial features is done by the experimenter, and different choices produce different results. Some iteration is usually necessary before satisfactory representations are achieved.

Chernoff faces appear to be most useful for verifying (1) an initial grouping suggested by subject-matter knowledge and intuition or (2) final groupings produced by clustering algorithms.

Chernoff Face Coding – for Utility Data

Example 1.12 (Utility data as Chernoff faces) From the data in Table 12.4, the 22 public utility companies were represented as Chernoff faces. We have the following correspondences:

Variable		Facial characteristic
X_1: Fixed-charge coverage	$\leftrightarrow$	Half-height of face
X_2: Rate of return on capital	$\leftrightarrow$	Face width
X_3: Cost per kW capacity in place	$\leftrightarrow$	Position of center of mouth
X_4: Annual load factor	$\leftrightarrow$	Slant of eyes
X_5: Peak kWh demand growth from 1974	$\leftrightarrow$	Eccentricity $\left(\frac{\text{height}}{\text{width}}\right)$ of eyes
X_6: Sales (kWh use per year)	$\leftrightarrow$	Half-length of eye
X_7: Percent nuclear	$\leftrightarrow$	Curvature of mouth
X_8: Total fuel costs (cents per kWh)	$\leftrightarrow$	Length of nose

Chernoff Faces – Subjectively Grouping "Similar" Faces into Seven Clusters

The Chernoff faces are shown in Figure 1.17. We have subjectively grouped "similar" faces into seven clusters. If a smaller number of clusters is desired, we might combine clusters 5, 6, and 7 and, perhaps, clusters 2 and 3 to obtain four or five clusters. For our assignment of variables to facial features, the firms group largely according to geographical location. ■

Constructing Chernoff faces is a task that must be done with the aid of a computer. The data are ordinarily standardized within the computer program as part of the process for determining the locations, sizes, and orientations of the facial characteristics. With some training, we can use Chernoff faces to communicate similarities or dissimilarities, as the next example indicates.

Chernoff Faces – Rating System to Track Financial Well-Being of Utility Co.

Example 1.13 (Using Chernoff faces to show changes over time) Figure 1.18 illustrates an additional use of Chernoff faces. (See [24].) In the figure, the faces are used to track the financial well-being of a company over time. As indicated, each facial feature represents a single financial indicator, and the longitudinal changes in these indicators are thus evident at a glance. ■

Chernoff Faces – Clusters (n = 7)

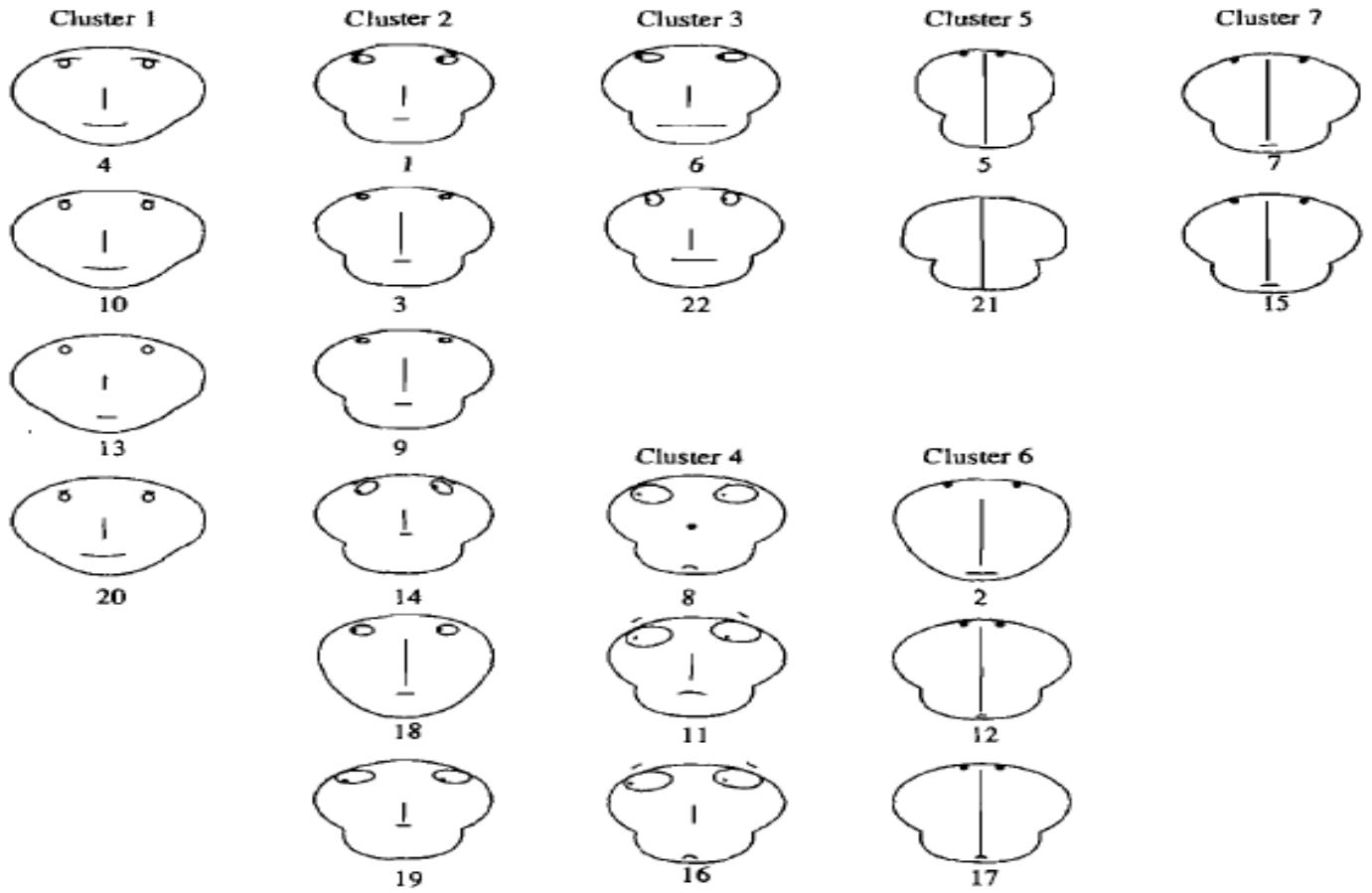


Figure 1.17 Chernoff faces for 22 public utilities.

Chernoff Faces for 22 Public Utilities Plotted from 1975 to 1979

Figure 1.17 Chernoff faces for 22 public utilities.

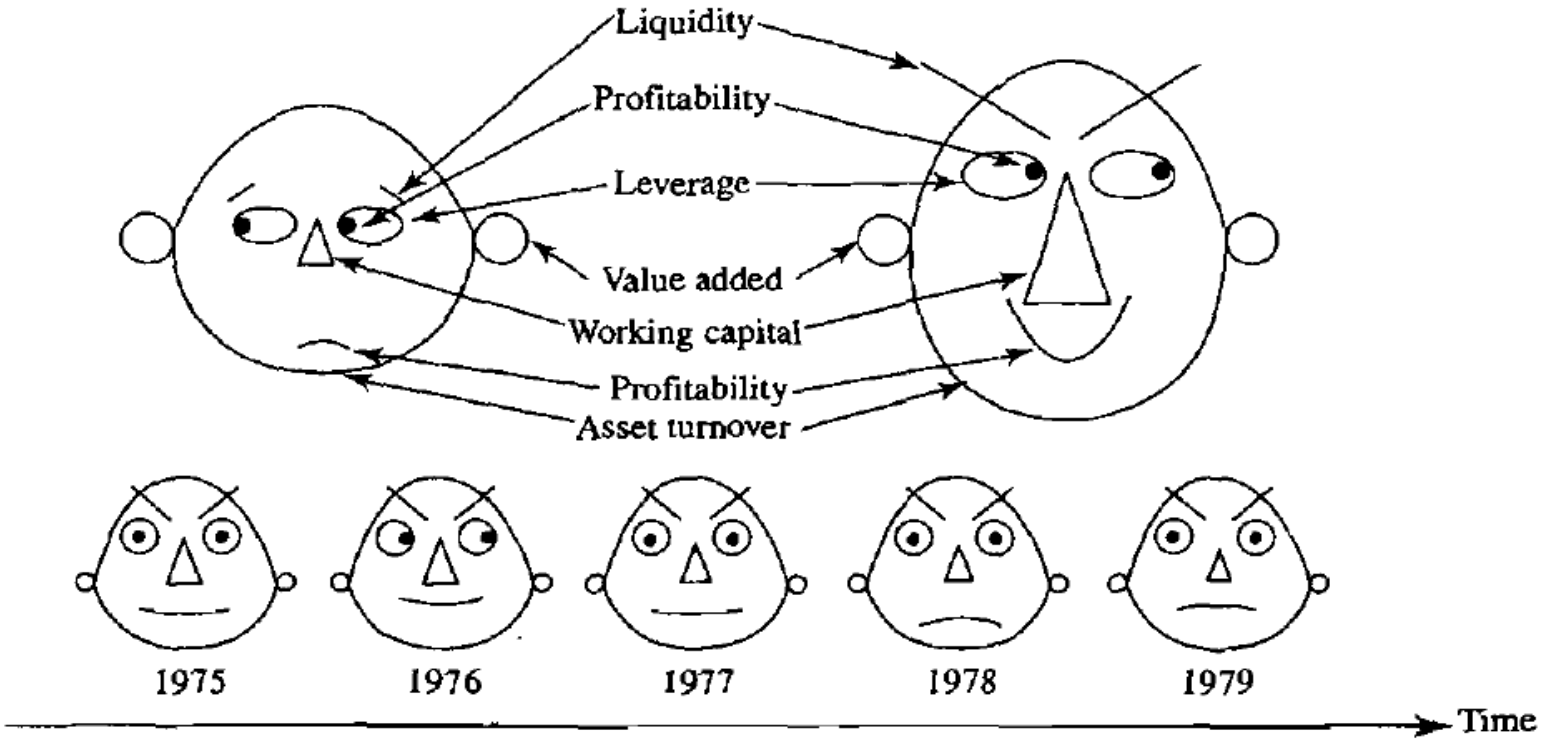


Figure 1.18 Chernoff faces over time.

Euclidean Distance in Multivariate Methods (including Clustering and Factor Analysis)

1.5 Distance

Although they may at first appear formidable, most multivariate techniques are based upon the simple concept of distance. Straight-line, or Euclidean, distance should be familiar. If we consider the point $P = (x_1, x_2)$ in the plane, the straight-line distance, $d(O, P)$, from P to the origin $O = (0, 0)$ is, according to the Pythagorean theorem,

$$d(O, P) = \sqrt{x_1^2 + x_2^2} \tag{1-9}$$

The situation is illustrated in Figure 1.19. In general, if the point P has p coordinates so that $P = (x_1, x_2, \ldots, x_p)$, the straight-line distance from P to the origin $O = (0, 0, \ldots, 0)$ is

$$d(O, P) = \sqrt{x_1^2 + x_2^2 + \cdots + x_p^2} \tag{1-10}$$

(See Chapter 2.) All points $(x_1, x_2, \ldots, x_p)$ that lie a constant squared distance, such as c^2, from the origin satisfy the equation

$$d^2(O, P) = x_1^2 + x_2^2 + \cdots + x_p^2 = c^2 \tag{1-11}$$

Equation of Hypersphere (Circle of p = 2); Points Equidistant from the Origin

Because this is the equation of a hypersphere (a circle if $p = 2$), points equidistant from the origin lie on a hypersphere.

The straight-line distance between two arbitrary points P and Q with coordinates $P = (x_1, x_2, \ldots, x_p)$ and $Q = (y_1, y_2, \ldots, y_p)$ is given by

$$d(P,Q) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_p - y_p)^2} \qquad (1\text{-}12)$$

Straight-line, or Euclidean, distance is unsatisfactory for most statistical purposes. This is because each coordinate contributes equally to the calculation of Euclidean distance. When the coordinates represent measurements that are subject to random fluctuations of differing magnitudes, it is often desirable to weight coordinates subject to a great deal of variability less heavily than those that are not highly variable. This suggests a different measure of distance.

Different Measure of Distance Due to Variability of Determinants (or Variables)

Our purpose now is to develop a "statistical" distance that accounts for differences in variation and, in due course, the presence of correlation. Because our

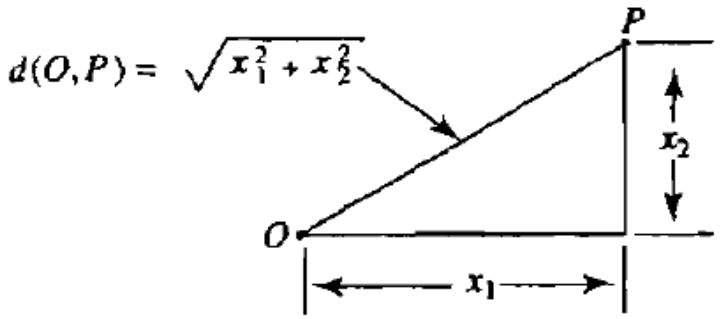


Figure 1.19 Distance given by the Pythagorean theorem.

choice will depend upon the sample variances and covariances, at this point we use the term *statistical distance* to distinguish it from ordinary Euclidean distance. It is statistical distance that is fundamental to multivariate analysis.

To begin, we take as *fixed* the set of observations graphed as the p-dimensional scatter plot. From these, we shall construct a measure of distance from the origin to a point $P = (x_1, x_2, \ldots, x_p)$. In our arguments, the coordinates $(x_1, x_2, \ldots, x_p)$ of P can vary to produce different locations for the point. The data that determine distance will, however, remain fixed.

Scatter Plot with Greater Variability of x_1 direction than in the x_2 direction

To illustrate, suppose we have n pairs of measurements on two variables each having mean zero. Call the variables x_1 and x_2, and assume that the x_1 measurements vary independently of the x_2 measurements.[1] In addition, assume that the variability in the x_1 measurements is larger than the variability in the x_2 measurements. A scatter plot of the data would look something like the one pictured in Figure 1.20.

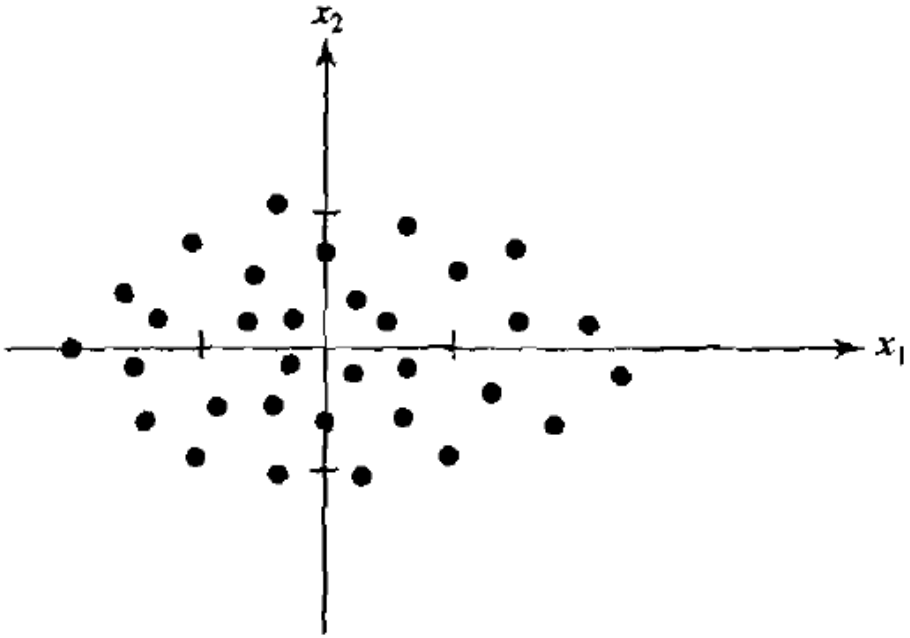


Figure 1.20 A scatter plot with greater variability in the x_1 direction than in the x_2 direction.

Range (Minimum to Maximum Length) of Variable x_1 versus Variable x_2 Differs

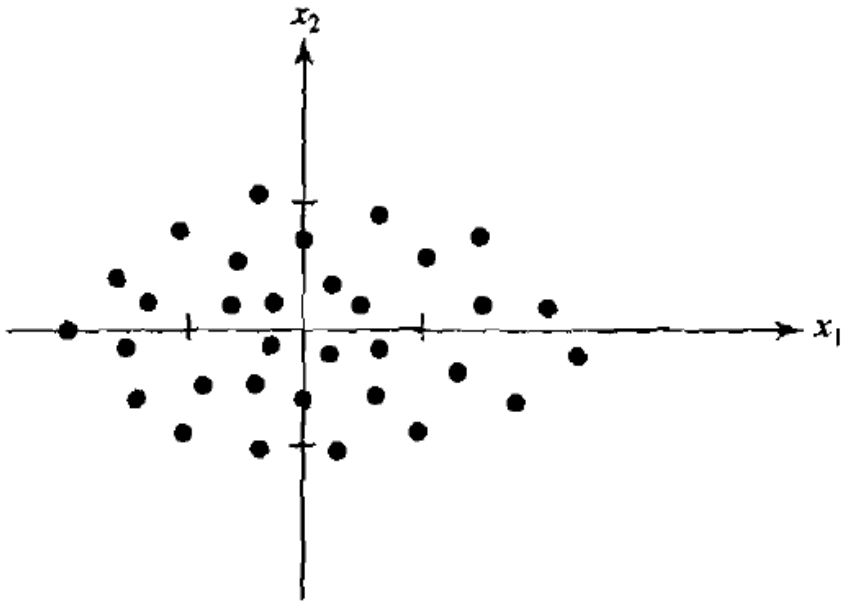


Figure 1.20 A scatter plot with greater variability in the x_1 direction than in the x_2 direction.

Glancing at Figure 1.20, we see that values which are a given deviation from the origin in the x_1 direction are not as "surprising" or "unusual" as are values equidistant from the origin in the x_2 direction. This is because the inherent variability in the x_1 direction is greater than the variability in the x_2 direction. Consequently, large x_1 coordinates (in absolute value) are not as unexpected as large x_2 coordinates. It seems reasonable, then, to weight an x_2 coordinate more heavily than an x_1 coordinate of the same value when computing the "distance" to the origin.

Divide x_1 and x_2 by the Standard Deviation in effect Weighting to Equalize or Standardize the Distance

One way to proceed is to divide each coordinate by the sample standard deviation. Therefore, upon division by the standard deviations, we have the "standardized" coordinates $x_1^* = x_1/\sqrt{s_{11}}$ and $x_2^* = x_2/\sqrt{s_{22}}$. The standardized coordinates are now on an equal footing with one another. After taking the differences in variability into account, we determine distance using the standard Euclidean formula.

Thus, a statistical distance of the point $P = (x_1, x_2)$ from the origin $O = (0, 0)$ can be computed from its standardized coordinates $x_1^* = x_1/\sqrt{s_{11}}$ and $x_2^* = x_2/\sqrt{s_{22}}$ as

$$d(O, P) = \sqrt{(x_1^*)^2 + (x_2^*)^2}$$

$$= \sqrt{\left(\frac{x_1}{\sqrt{s_{11}}}\right)^2 + \left(\frac{x_2}{\sqrt{s_{22}}}\right)^2} = \sqrt{\frac{x_1^2}{s_{11}} + \frac{x_2^2}{s_{22}}} \qquad (1\text{-}13)$$

[1]At this point, "independently" means that the x_2 measurements cannot be predicted with any accuracy from the x_1 measurements, and vice versa.

Using the Formula of an Ellipse to Explain the Scatter Plot Shape

Comparing (1-13) with (1-9), we see that the difference between the two expressions is due to the weights $k_1 = 1/s_{11}$ and $k_2 = 1/s_{22}$ attached to x_1^2 and x_2^2 in (1-13). Note that if the sample variances are the same, $k_1 = k_2$, then x_1^2 and x_2^2 will receive the same weight. In cases where the weights are the same, it is convenient to ignore the common divisor and use the usual Euclidean distance formula. In other words, if the variability in the x_1 direction is the same as the variability in the x_2 direction, and the x_1 values vary independently of the x_2 values, Euclidean distance is appropriate.

Using (1-13), we see that all points which have coordinates (x_1, x_2) and are a constant squared distance c^2 from the origin must satisfy

$$\frac{x_1^2}{s_{11}} + \frac{x_2^2}{s_{22}} = c^2 \qquad (1\text{-}14)$$

Equation (1-14) is the equation of an ellipse centered at the origin whose major and minor axes coincide with the coordinate axes. That is, the statistical distance in (1-13) has an ellipse as the locus of all points a constant distance from the origin. This general case is shown in Figure 1.21.

Ellipse as Correlation Diagram – Using Distance Between Variables to Show Relationship (Closest to 1)

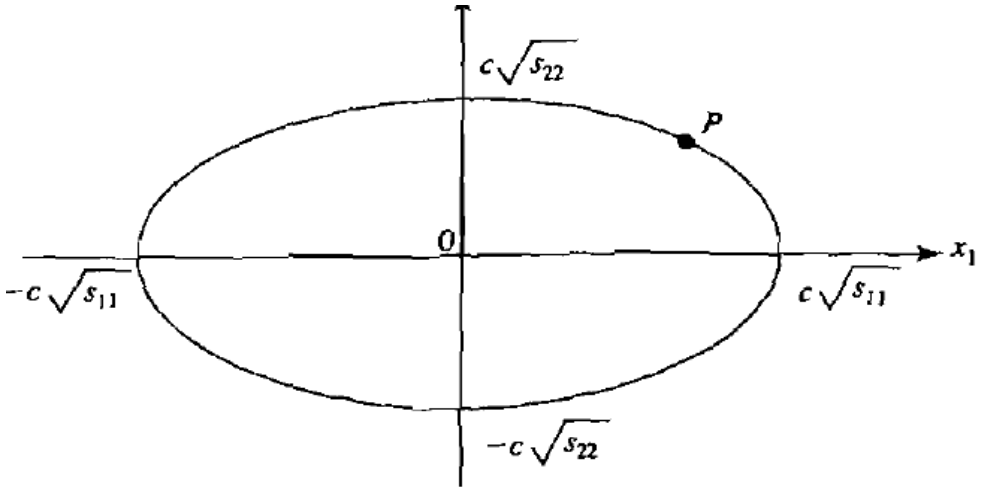


Figure 1.21 The ellipse of constant statistical distance
$d^2(O, P) = x_1^2/s_{11} + x_2^2/s_{22} = c^2$.

Example 1.14 (Calculating a statistical distance) A set of paired measurements (x_1, x_2) on two variables yields $\bar{x}_1 = \bar{x}_2 = 0$, $s_{11} = 4$, and $s_{22} = 1$. Suppose the x_1 measurements are unrelated to the x_2 measurements; that is, measurements within a pair vary independently of one another. Since the sample variances are unequal, we measure the square of the distance of an arbitrary point $P = (x_1, x_2)$ to the origin $O = (0, 0)$ by

$$d^2(O, P) = \frac{x_1^2}{4} + \frac{x_2^2}{1}$$

All Coordinate Points Constant Distant 1 from the Origin Depicted in the Table

All points (x_1, x_2) that are a constant distance 1 from the origin satisfy the equation

$$\frac{x_1^2}{4} + \frac{x_2^2}{1} = 1$$

The coordinates of some points a unit distance from the origin are presented in the following table:

Coordinates: (x_1, x_2)	Distance: $\frac{x_1^2}{4} + \frac{x_2^2}{1} = 1$
$(0,1)$	$\frac{0^2}{4} + \frac{1^2}{1} = 1$
$(0,-1)$	$\frac{0^2}{4} + \frac{(-1)^2}{1} = 1$
$(2,0)$	$\frac{2^2}{4} + \frac{0^2}{1} = 1$
$(1, \sqrt{3}/2)$	$\frac{1^2}{4} + \frac{(\sqrt{3}/2)^2}{1} = 1$

Ellipse Center at (0,0) – All points same distance from the origin

A plot of the equation $x_1^2/4 + x_2^2/1 = 1$ is an ellipse centered at $(0, 0)$ whose major axis lies along the x_1 coordinate axis and whose minor axis lies along the x_2 coordinate axis. The half-lengths of these major and minor axes are $\sqrt{4} = 2$ and $\sqrt{1} = 1$, respectively. The ellipse of unit distance is plotted in Figure 1.22. All points on the ellipse are regarded as being the same statistical distance from the origin—in this case, a distance of 1. ■

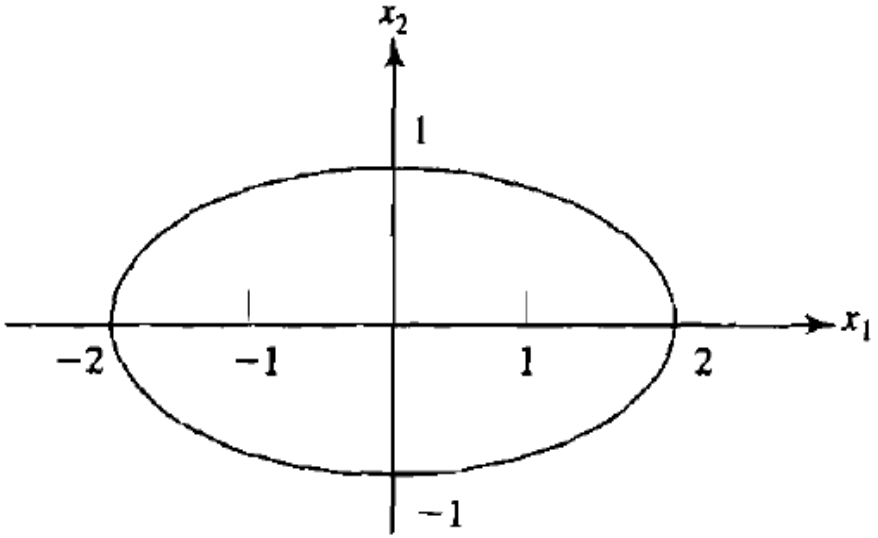


Figure 1.22 Ellipse of unit distance, $\frac{x_1^2}{4} + \frac{x_2^2}{1} = 1$.

Distance of Two Coordinate Variables (P,Q) Equation Below

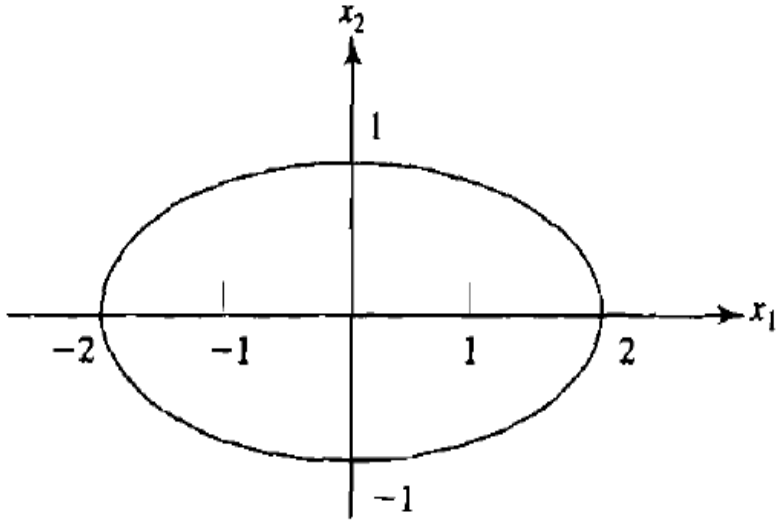


Figure 1.22 Ellipse of unit distance, $\frac{x_1^2}{4} + \frac{x_2^2}{1} = 1$.

The expression in (1-13) can be generalized to accommodate the calculation of statistical distance from an arbitrary point $P = (x_1, x_2)$ to any *fixed* point $Q = (y_1, y_2)$. If we assume that the coordinate variables vary independently of one another, the distance from P to Q is given by

$$d(P,Q) = \sqrt{\frac{(x_1 - y_1)^2}{s_{11}} + \frac{(x_2 - y_2)^2}{s_{22}}} \qquad (1\text{-}15)$$

Distance of Two Coordinate Paired Variables (P,Q) (ad infinitum)Equation

The extension of this statistical distance to more than two dimensions is straightforward. Let the points P and Q have p coordinates such that $P = (x_1, x_2, \ldots, x_p)$ and $Q = (y_1, y_2, \ldots, y_p)$. Suppose Q is a fixed point [it may be the origin $O = (0, 0, \ldots, 0)$] and the coordinate variables vary independently of one another. Let $s_{11}, s_{22}, \ldots, s_{pp}$ be sample variances constructed from n measurements on $x_1, x_2, \ldots, x_p$, respectively. Then the statistical distance from P to Q is

$$d(P,Q) = \sqrt{\frac{(x_1 - y_1)^2}{s_{11}} + \frac{(x_2 - y_2)^2}{s_{22}} + \cdots + \frac{(x_p - y_p)^2}{s_{pp}}} \qquad (1\text{-}16)$$

Considering Euclidian Distance , All points that are P are a Constant Distance from Q like on a Hyperellipsoid

All points P that are a constant squared distance from Q lie on a hyperellipsoid centered at Q whose major and minor axes are parallel to the coordinate axes. We note the following:

1. The distance of P to the origin O is obtained by setting $y_1 = y_2 = \cdots = y_p = 0$ in (1-16).
2. If $s_{11} = s_{22} = \cdots = s_{pp}$, the Euclidean distance formula in (1-12) is appropriate.

The distance in (1-16) still does not include most of the important cases we shall encounter, because of the assumption of independent coordinates. The scatter plot in Figure 1.23 depicts a two-dimensional situation in which the x_1 measurements do not vary independently of the x_2 measurements. In fact, the coordinates of the pairs (x_1, x_2) exhibit a tendency to be large or small together, and the sample correlation coefficient is positive. Moreover, the variability in the x_2 direction is larger than the variability in the x_1 direction.

Distance or d(O,P) as Function of x_1 and x_2 Axes Divided by Respective Variances

What is a meaningful measure of distance when the variability in the x_1 direction is different from the variability in the x_2 direction and the variables x_1 and x_2 are correlated? Actually, we can use what we have already introduced, provided that we look at things in the right way. From Figure 1.23, we see that if we rotate the original coordinate system through the angle θ while keeping the scatter fixed and label the rotated axes $\tilde{x}_1$ and $\tilde{x}_2$, the scatter in terms of the new axes looks very much like that in Figure 1.20. (You may wish to turn the book to place the $\tilde{x}_1$ and $\tilde{x}_2$ axes in their customary positions.) This suggests that we calculate the sample variances using the $\tilde{x}_1$ and $\tilde{x}_2$ coordinates and measure distance as in Equation (1-13). That is, with reference to the $\tilde{x}_1$ and $\tilde{x}_2$ axes, we define the distance from the point $P = (\tilde{x}_1, \tilde{x}_2)$ to the origin $O = (0, 0)$ as

$$d(O, P) = \sqrt{\frac{\tilde{x}_1^2}{\tilde{s}_{11}} + \frac{\tilde{x}_2^2}{\tilde{s}_{22}}} \tag{1-17}$$

where $\tilde{s}_{11}$ and $\tilde{s}_{22}$ denote the sample variances computed with the $\tilde{x}_1$ and $\tilde{x}_2$ measurements.

Divide x_1 and x_2 by the Standard Deviation in effect Weighting to Equalize or Standardize the Distance

$$d(O, P) = \sqrt{\frac{\tilde{x}_1^2}{\tilde{s}_{11}} + \frac{\tilde{x}_2^2}{\tilde{s}_{22}}} \qquad (1\text{-}17)$$

where $\tilde{s}_{11}$ and $\tilde{s}_{22}$ denote the sample variances computed with the $\tilde{x}_1$ and $\tilde{x}_2$ measurements.

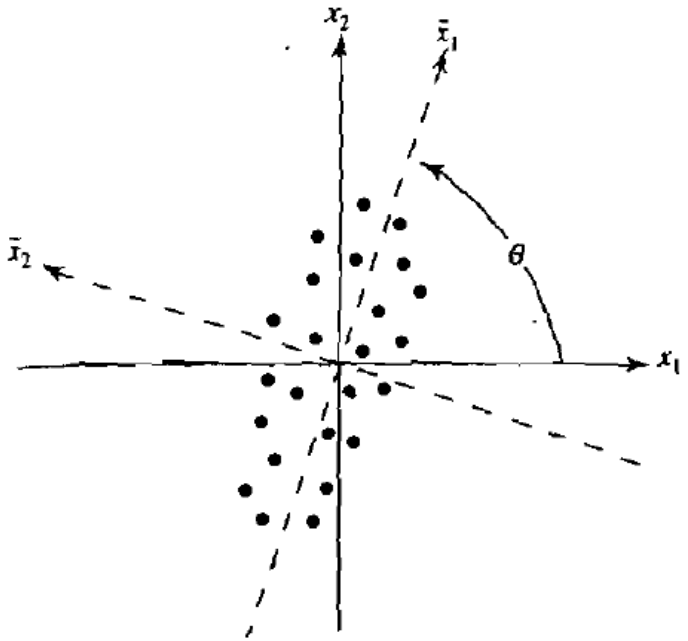


Figure 1.23 A scatter plot for positively correlated measurements and a rotated coordinate system.

Relationship Between Original Coordinates x_1 and x_2 and Rotated Coordinates By Sin and cosine of Theta

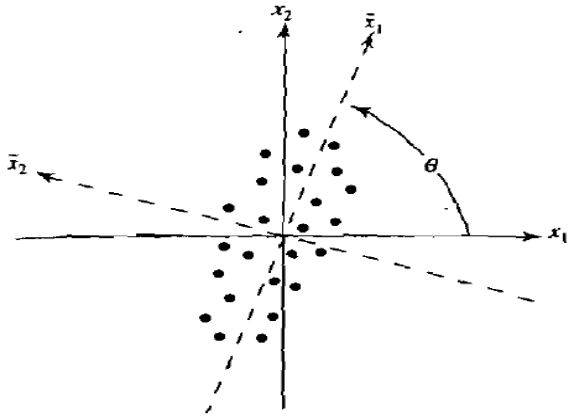


Figure 1.23 A scatter plot for positively correlated measurements and a rotated coordinate system.

The relation between the original coordinates (x_1, x_2) and the rotated coordinates $(\tilde{x}_1, \tilde{x}_2)$ is provided by

$$\begin{aligned} \tilde{x}_1 &= x_1 \cos(\theta) + x_2 \sin(\theta) \\ \tilde{x}_2 &= -x_1 \sin(\theta) + x_2 \cos(\theta) \end{aligned} \tag{1-18}$$

Given the relations in (1-18), we can formally substitute for $\tilde{x}_1$ and $\tilde{x}_2$ in (1-17) and express the distance in terms of the original coordinates.

Distance In Terms of Original Coordinates x_1 and x_2 and Always Positive

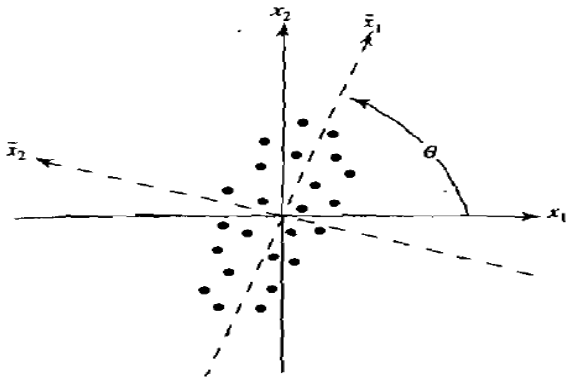


Figure 1.23 A scatter plot for positively correlated measurements and a rotated coordinate system.

After some straightforward algebraic manipulations, the distance from $P = (\tilde{x}_1, \tilde{x}_2)$ to the origin $O = (0, 0)$ can be written in terms of the original coordinates x_1 and x_2 of P as

$$d(O, P) = \sqrt{a_{11}x_1^2 + 2a_{12}x_1x_2 + a_{22}x_2^2} \tag{1-19}$$

where the a's are numbers such that the distance is nonnegative for all possible values of x_1 and x_2. Here a_{11}, a_{12}, and a_{22} are determined by the angle θ, and s_{11}, s_{12}, and s_{22} calculated from the original data.[2] The particular forms for a_{11}, a_{12}, and a_{22} are not important at this point. What is important is the appearance of the cross-product term $2a_{12}x_1x_2$ necessitated by the nonzero correlation r_{12}.

Equation (1-19) can be compared with (1-13). The expression in (1-13) can be regarded as a special case of (1-19) with $a_{11} = 1/s_{11}$, $a_{22} = 1/s_{22}$, and $a_{12} = 0$.

Distance Based on the Array of Coordinates x_1 and x_2 That Are Correlated

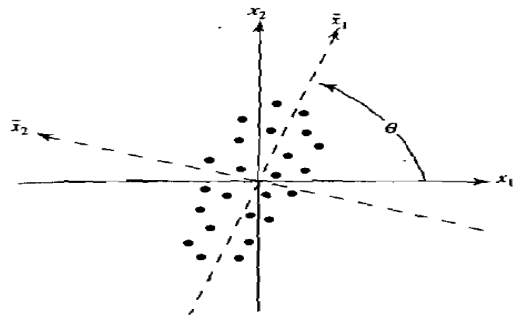


Figure 1.23 A scatter plot for positively correlated measurements and a rotated coordinate system.

Equation (1-19) can be compared with (1-13). The expression in (1-13) can be regarded as a special case of (1-19) with $a_{11} = 1/s_{11}$, $a_{22} = 1/s_{22}$, and $a_{12} = 0$.

In general, the statistical distance of the point $P = (x_1, x_2)$ from the *fixed* point $Q = (y_1, y_2)$ for situations in which the variables are correlated has the general form

$$d(P,Q) = \sqrt{a_{11}(x_1 - y_1)^2 + 2a_{12}(x_1 - y_1)(x_2 - y_2) + a_{22}(x_2 - y_2)^2} \qquad (1\text{-}20)$$

and can always be computed once a_{11}, a_{12}, and a_{22} are known. In addition, the coordinates of all points $P = (x_1, x_2)$ that are a constant squared distance c^2 from Q satisfy

$$a_{11}(x_1 - y_1)^2 + 2a_{12}(x_1 - y_1)(x_2 - y_2) + a_{22}(x_2 - y_2)^2 = c^2 \qquad (1\text{-}21)$$

Distance is the Equation of An Ellipse Centered at Q (Values for Each Coefficient in the Matrix Below)

By definition, this is the equation of an ellipse centered at Q. The graph of such an equation is displayed in Figure 1.24. The major (long) and minor (short) axes are indicated. They are parallel to the $\tilde{x}_1$ and $\tilde{x}_2$ axes. For the choice of a_{11}, a_{12}, and a_{22} in footnote 2, the $\tilde{x}_1$ and $\tilde{x}_2$ axes are at an angle θ with respect to the x_1 and x_2 axes.

The generalization of the distance formulas of (1-19) and (1-20) to p dimensions is straightforward. Let $P = (x_1, x_2, \ldots, x_p)$ be a point whose coordinates represent variables that are correlated and subject to inherent variability. Let

[2]Specifically,

$$a_{11} = \frac{\cos^2(\theta)}{\cos^2(\theta)s_{11} + 2\sin(\theta)\cos(\theta)s_{12} + \sin^2(\theta)s_{22}} + \frac{\sin^2(\theta)}{\cos^2(\theta)s_{22} - 2\sin(\theta)\cos(\theta)s_{12} + \sin^2(\theta)s_{11}}$$

$$a_{22} = \frac{\sin^2(\theta)}{\cos^2(\theta)s_{11} + 2\sin(\theta)\cos(\theta)s_{12} + \sin^2(\theta)s_{22}} + \frac{\cos^2(\theta)}{\cos^2(\theta)s_{22} - 2\sin(\theta)\cos(\theta)s_{12} + \sin^2(\theta)s_{11}}$$

and

$$a_{12} = \frac{\cos(\theta)\sin(\theta)}{\cos^2(\theta)s_{11} + 2\sin(\theta)\cos(\theta)s_{12} + \sin^2(\theta)s_{22}} - \frac{\sin(\theta)\cos(\theta)}{\cos^2(\theta)s_{22} - 2\sin(\theta)\cos(\theta)s_{12} + \sin^2(\theta)s_{11}}$$

Ellipse of Points – a Constant Distance from the Point Q

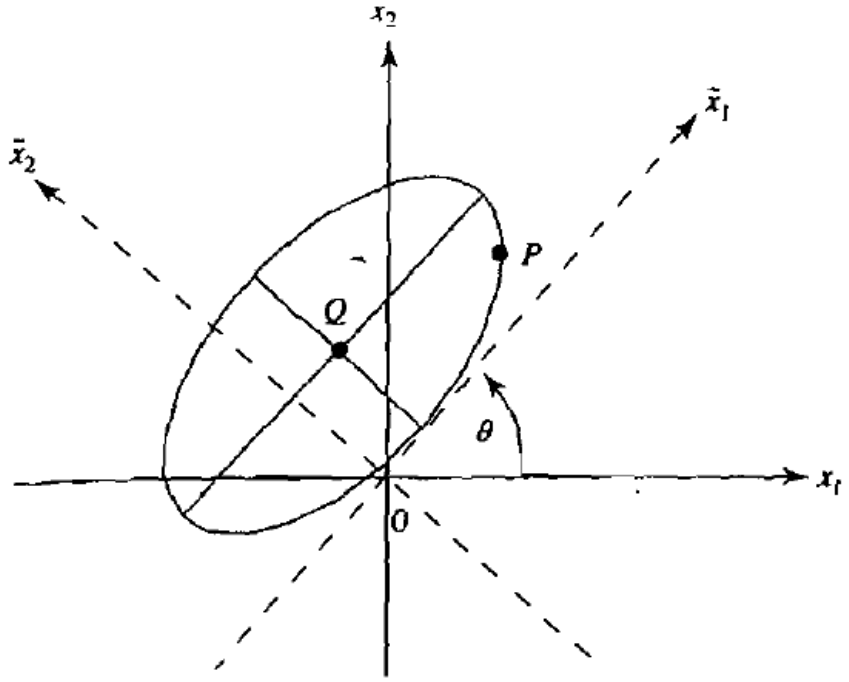


Figure 1.24 Ellipse of points a constant distance from the point Q.

$O = (0, 0, \ldots, 0)$ denote the origin, and let $Q = (y_1, y_2, \ldots, y_p)$ be a specified *fixed* point. Then the distances from P to O and from P to Q have the general forms

$$d(O,P) = \sqrt{a_{11}x_1^2 + a_{22}x_2^2 + \cdots + a_{pp}x_p^2 + 2a_{12}x_1x_2 + 2a_{13}x_1x_3 + \cdots + 2a_{p-1,p}x_{p-1}x_p} \tag{1-22}$$

Distance Determined by Coefficients or Weights Shown in the Array Below

$$d(O,P) = \sqrt{a_{11}x_1^2 + a_{22}x_2^2 + \cdots + a_{pp}x_p^2 + 2a_{12}x_1x_2 + 2a_{13}x_1x_3 + \cdots + 2a_{p-1,p}x_{p-1}x_p} \tag{1-22}$$

and

$$P,Q) = \sqrt{\begin{aligned}[a_{11}(x_1 - y_1)^2 + a_{22}(x_2 - y_2)^2 + \cdots + a_{pp}(x_p - y_p)^2 + 2a_{12}(x_1 - y_1)(x_2 - y_2) \\ + 2a_{13}(x_1 - y_1)(x_3 - y_3) + \cdots + 2a_{p-1,p}(x_{p-1} - y_{p-1})(x_p - y_p)]\end{aligned}} \tag{1-23}$$

where the a's are numbers such that the distances are always nonnegative.[3]

We note that the distances in (1-22) and (1-23) are completely determined by the coefficients (weights) a_{ik}, $i = 1, 2, \ldots, p$, $k = 1, 2, \ldots, p$. These coefficients can be set out in the rectangular array

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{12} & a_{22} & \cdots & a_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1p} & a_{2p} & \cdots & a_{pp} \end{bmatrix} \tag{1-24}$$

Array: A_{ik} Displayed Twice Since Multiplied by 2 in the Distance Formula

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{12} & a_{22} & \cdots & a_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1p} & a_{2p} & \cdots & a_{pp} \end{bmatrix} \tag{1-24}$$

where the a_{ik}'s with $i \neq k$ are displayed twice, since they are multiplied by 2 in the distance formulas. Consequently, the entries in this array specify the distance functions. The a_{ik}'s cannot be arbitrary numbers; they must be such that the computed distance is nonnegative for every pair of points. (See Exercise 1.10.)

Contours of constant distances computed from (1-22) and (1-23) are hyperellipsoids. A hyperellipsoid resembles a football when $p = 3$; it is impossible to visualize in more than three dimensions.

[3]The algebraic expressions for the *squares* of the distances in (1-22) and (1-23) are known as *quadratic forms* and, in particular, *positive definite quadratic forms*. It is possible to display these quadratic forms in a simpler manner using matrix algebra; we shall do so in Section 2.3 of Chapter 2.

Euclidean vs. Statistical Distance: Statistical Distance Takes into Account Variability of the Points in the Cluster

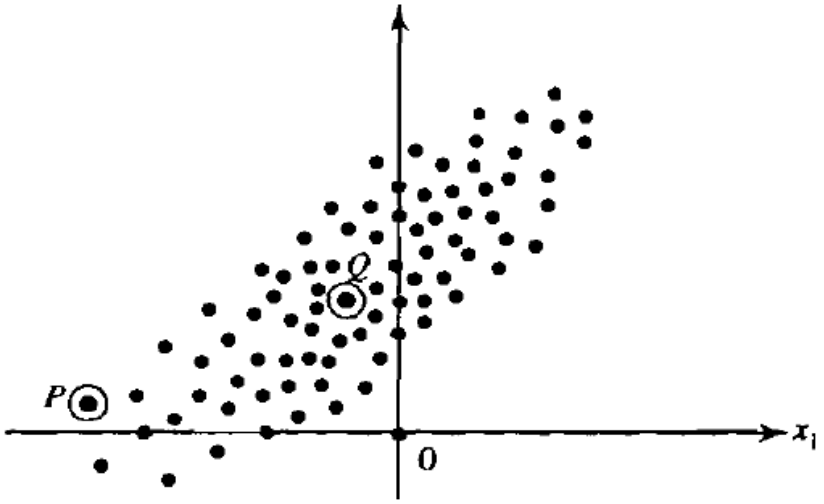


Figure 1.25 A cluster of points relative to a point P and the origin.

The need to consider statistical rather than Euclidean distance is illustrated heuristically in Figure 1.25. Figure 1.25 depicts a cluster of points whose center of gravity (sample mean) is indicated by the point Q. Consider the Euclidean distances from the point Q to the point P and the origin O. The Euclidean distance from Q to P is larger than the Euclidean distance from Q to O. However, P appears to be more like the points in the cluster than does the origin. If we take into account the variability of the points in the cluster and measure distance by the statistical distance in (1-20), then Q will be closer to P than to O. This result seems reasonable, given the nature of the scatter.

Other Distances not Elliptical Shaped Must Satisfy Properties

useful to consider distances that are not related to circles or ellipses. Any distance measure $d(P, Q)$ between two points P and Q is valid provided that it satisfies the following properties, where R is any other intermediate point:

$$
\begin{aligned}
&d(P, Q) = d(Q, P) \\
&d(P, Q) > 0 \text{ if } P \neq Q \\
&d(P, Q) = 0 \text{ if } P = Q \\
&d(P, Q) \leq d(P, R) + d(R, Q) \qquad \text{(triangle inequality)}
\end{aligned}
\tag{1-25}
$$

1.6 Final Comments

We have attempted to motivate the study of multivariate analysis and to provide you with some rudimentary, but important, methods for organizing, summarizing, and displaying data. In addition, a general concept of distance has been introduced that will be used repeatedly in later chapters.

Examples
Exercise 1.1

Exercises

1.1. Consider the seven pairs of measurements (x_1, x_2) plotted in Figure 1.1:

x_1	3	4	2	6	8	2	5
x_2	5	5.5	4	7	10	5	7.5

Calculate the sample means $\bar{x}_1$ and $\bar{x}_2$, the sample variances s_{11} and s_{22}, and the sample covariance s_{12}.

Exercises 1.1 Solution

Chapter 1

1.1 $\bar{x}_1 = 4.29$ $\bar{x}_2 = 6.29$

$s_{11} = 4.20$ $s_{22} = 3.56$ $s_{12} = 3.70$

Exercise 1.2

1.2. A morning newspaper lists the following used-car prices for a foreign compact with age x_1 measured in years and selling price x_2 measured in thousands of dollars:

x_1	1	2	3	3	4	5	6	8	9	11
x_2	18.95	19.00	17.95	15.54	14.00	12.95	8.94	7.49	6.00	3.99

(a) Construct a scatter plot of the data and marginal dot diagrams.

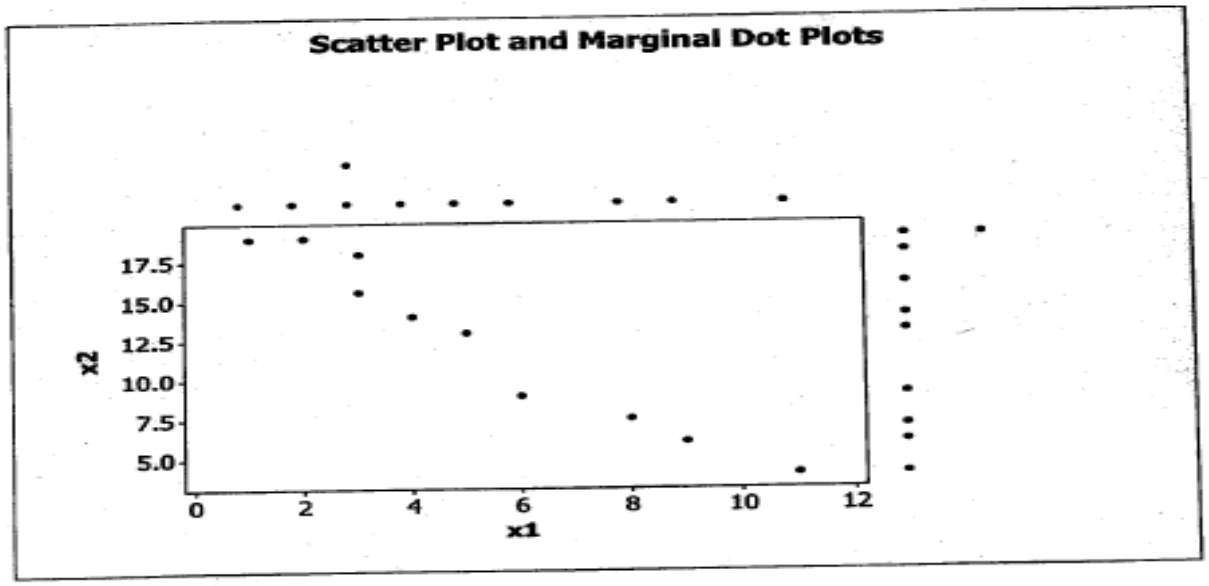
(b) Infer the sign of the sample covariance s_{12} from the scatter plot.

(c) Compute the sample means $\bar{x}_1$ and $\bar{x}_2$ and the sample variances s_{11} and s_{22}. Compute the sample covariance s_{12} and the sample correlation coefficient r_{12}. Interpret these quantities.

(d) Display the sample mean array $\bar{\mathbf{x}}$, the sample variance-covariance array $\mathbf{S}_n$, and the sample correlation array $\mathbf{R}$ using (1-8).

Solution 1.2

1.2 a)



b) s_{12} is negative

c)

$\bar{x}_1 = 5.20 \qquad \bar{x}_2 = 12.48 \qquad s_{11} = 3.09 \qquad s_{22} = 5.27$

$s_{12} = -15.94 \qquad r_{12} = -.98$

Large x_1 occurs with small x_2 and vice versa.

d)

$$\bar{x} = \begin{bmatrix} 5.20 \\ 12.48 \end{bmatrix} \qquad S_n = \begin{bmatrix} 3.09 & -15.94 \\ -15.94 & 5.27 \end{bmatrix} \qquad R = \begin{bmatrix} 1 & -.98 \\ -.98 & 1 \end{bmatrix}$$

Exercise 1.3

1.3. The following are five measurements on the variables x_1, x_2, and x_3:

x_1	9	2	6	5	8
x_2	12	8	6	4	10
x_3	3	4	0	2	1

Find the arrays $\bar{\mathbf{x}}$, $\mathbf{S}_n$, and $\mathbf{R}$.

Solution 1.3

1.3

$$\bar{x} = \begin{bmatrix} 6 \\ 8 \\ 2 \end{bmatrix} \qquad S_n = \begin{bmatrix} 6 & 4 & -1.4 \\ & 8 & 1.2 \\ \text{(symmetric)} & & 2 \end{bmatrix} \qquad R = \begin{bmatrix} 1 & .577 & -.404 \\ & 1 & .300 \\ \text{(symmetric)} & & 1 \end{bmatrix}$$

Exercise 1.4

1.4. The world's 10 largest companies yield the following data:

The World's 10 Largest Companies[1]

Company	x_1 = sales (billions)	x_2 = profits (billions)	x_3 = assets (billions)
Citigroup	108.28	17.05	1,484.10
General Electric	152.36	16.59	750.33
American Intl Group	95.04	10.91	766.42
Bank of America	65.45	14.14	1,110.46
HSBC Group	62.97	9.52	1,031.29
ExxonMobil	263.99	25.33	195.26
Royal Dutch/Shell	265.19	18.54	193.83
BP	285.06	15.73	191.11
ING Group	92.01	8.10	1,175.16
Toyota Motor	165.68	11.13	211.15

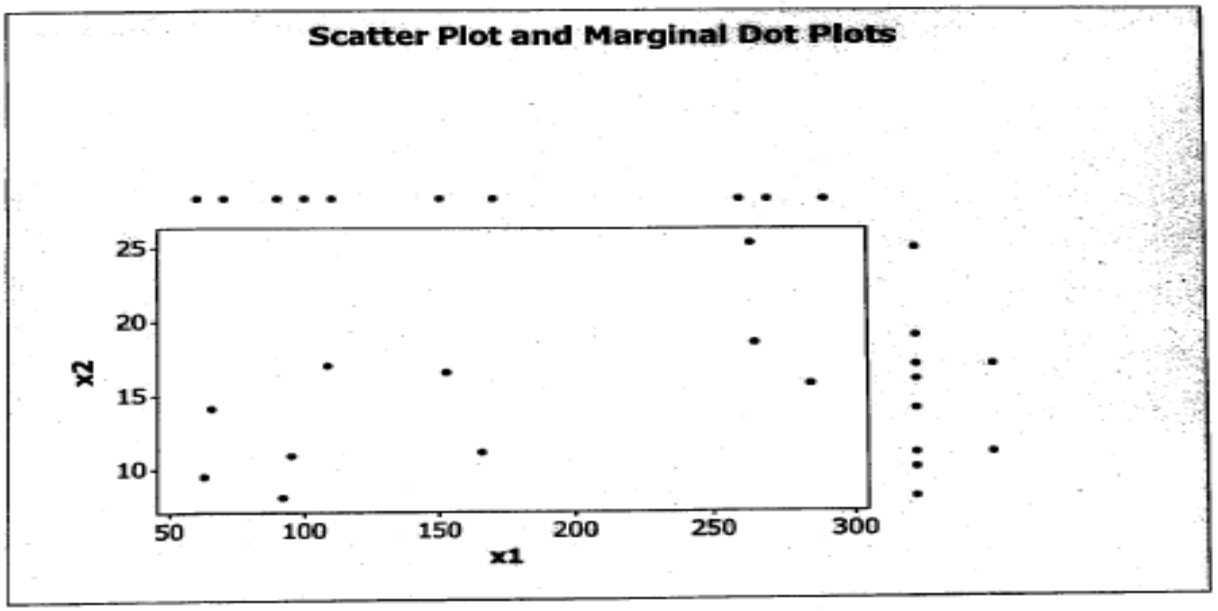
[1]From www.Forbes.com partially based on *Forbes*' The Forbes Global 2000, April 18, 2005.

(a) Plot the scatter diagram and marginal dot diagrams for variables x_1 and x_2. Comment on the appearance of the diagrams.

(b) Compute $\bar{x}_1$, $\bar{x}_2$, s_{11}, s_{22}, s_{12}, and r_{12}. Interpret r_{12}.

Solution 1.4

1.4 a) There is a positive correlation between x_1 and x_2. Since sample size is small, hard to be definitive about nature of marginal distributions. However, marginal distribution of x_1 appears to be skewed to the right. The marginal distribution of x_2 seems reasonably symmetric.



b)

$\bar{x}_1 = 155.60$ $\bar{x}_2 = 14.70$ $s_{11} = 82.03$ $s_{22} = 4.85$

$s_{12} = 273.26$ $r_{12} = .69$

Large profits (x_2) tend to be associated with large sales (x_1); small profits with small sales.

Exercise 1.5

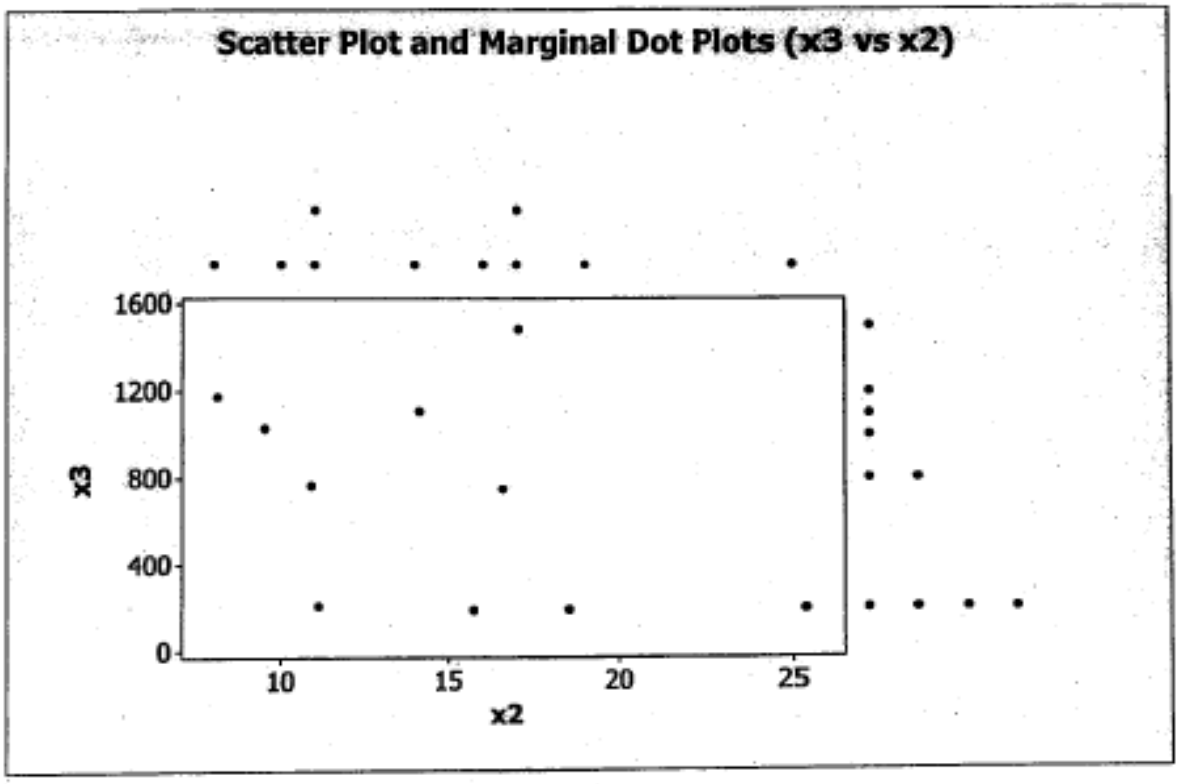
1.5. Use the data in Exercise 1.4.

(a) Plot the scatter diagrams and dot diagrams for (x_2, x_3) and (x_1, x_3). Comment on the patterns.

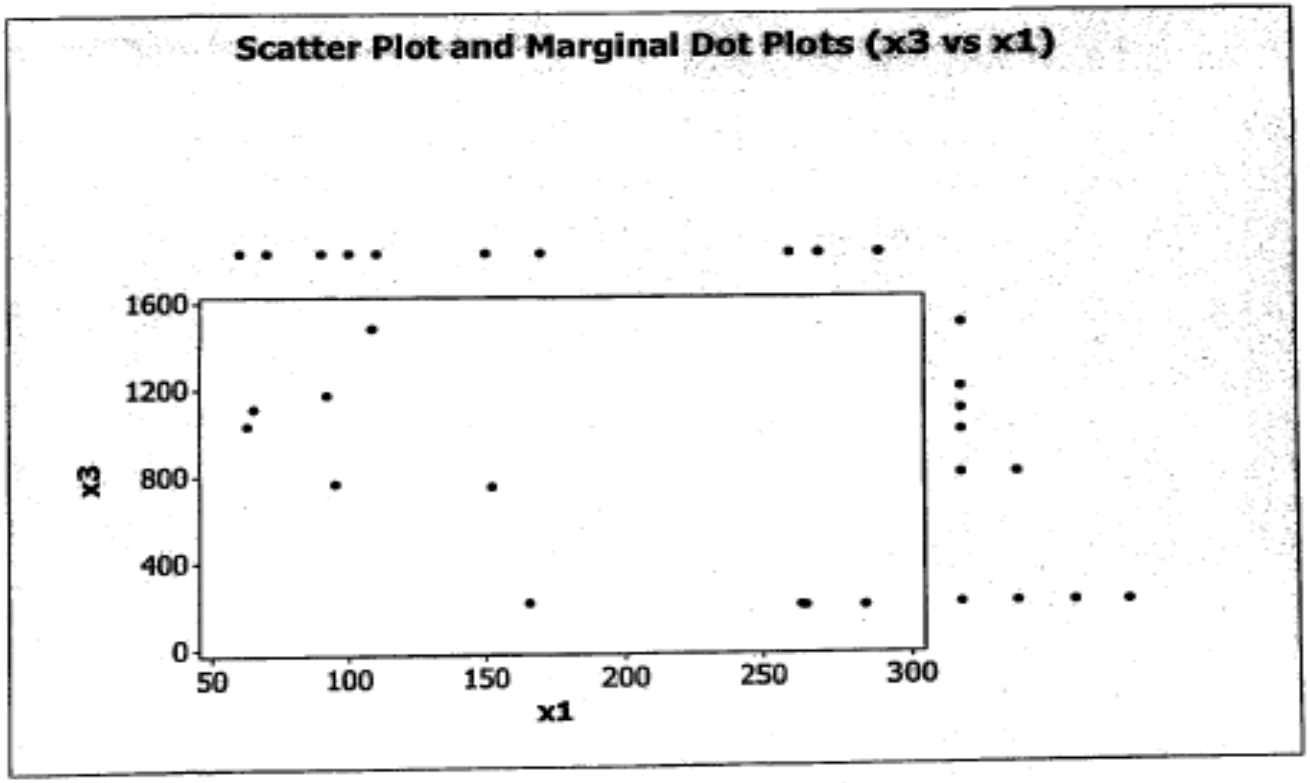
(b) Compute the $\bar{\mathbf{x}}$, $\mathbf{S}_n$, and $\mathbf{R}$ arrays for (x_1, x_2, x_3).

Solution 1.5

1.5 a) There is negative correlation between x_2 and x_3 and negative correlation between x_1 and x_3. The marginal distribution of x_1 appears to be skewed to the right. The marginal distribution of x_2 seems reasonably symmetric. The marginal distribution of x_3 also appears to be skewed to the right.



Solution 1.5


Scatter Plot and Marginal Dot Plots (x3 vs x1)
x3
1600
1200
800
400
0
50
100
150
200
250
300
x1

Solution 1.5

1.5 b)

$$\bar{x} = \begin{bmatrix} 155.60 \\ 14.70 \\ 710.91 \end{bmatrix} \qquad S_n = \begin{bmatrix} 82.03 & 273.26 & -32018.36 \\ 273.26 & 4.85 & -948.45 \\ -32018.36 & -948.45 & 461.90 \end{bmatrix}$$

$$R = \begin{bmatrix} 1 & .69 & -.85 \\ .69 & 1 & -.42 \\ -.85 & -.42 & 1 \end{bmatrix}$$

Exercise 1.6

1.6. The data in Table 1.5 are 42 measurements on air-pollution variables recorded at 12:00 noon in the Los Angeles area on different days. (See also the air-pollution data on the web at www.prenhall.com/statistics.)

(a) Plot the marginal dot diagrams for all the variables.

(b) Construct the $\bar{x}$, S_n, and **R** arrays, and interpret the entries in **R**.

Table 1.5 Air-Pollution Data

Wind (x_1)	Solar radiation (x_2)	CO (x_3)	NO (x_4)	NO_2 (x_5)	O_3 (x_6)	HC (x_7)
8	98	7	2	12	8	2
7	107	4	3	9	5	3
7	103	4	3	5	6	3
10	88	5	2	8	15	4
6	91	4	2	8	10	3
8	90	5	2	12	12	4
9	84	7	4	12	15	5
5	72	6	4	21	14	4
7	82	5	1	11	11	3
8	64	5	2	13	9	4
6	71	5	4	10	3	3
6	91	4	2	12	7	3
7	72	7	4	18	10	3
10	70	4	2	11	7	3
10	72	4	1	8	10	3
9	77	4	1	9	10	3
8	76	4	1	7	7	3
8	71	5	3	16	4	4
9	67	4	2	13	2	3
9	69	3	3	9	5	3
10	62	5	3	14	4	4
9	88	4	2	7	6	3
8	80	4	2	13	11	4
5	30	3	3	5	2	3
6	83	5	1	10	23	4
8	84	3	2	7	6	3
6	78	4	2	11	11	3
8	79	2	1	7	10	3

Exercise 1.6

6	62	4	3	9	8	3
10	37	3	1	7	2	3
8	71	4	1	10	7	3
7	52	4	1	12	8	4
5	48	6	5	8	4	3
6	75	4	1	10	24	3
10	35	4	1	6	9	2
8	85	4	1	9	10	2
5	86	3	1	6	12	2
5	86	7	2	13	18	2
7	79	7	4	9	25	3
7	79	5	2	8	6	2
6	68	6	2	11	14	3
8	40	4	3	6	5	2

Source: Data courtesy of Professor G. C. Tiao.

Solution 1.6

1.6 a) Histograms

X_1

MIDDLE OF INTERVAL	NUMBER OF OBSERVATIONS	
5.	5	*****
6.	8	********
7.	7	*******
8.	11	***********
9.	5	*****
10.	6	******

X_2

MIDDLE OF INTERVAL	NUMBER OF OBSERVATIONS	
30.	1	*
40.	3	***
50.	2	**
60.	3	***
70.	10	**********
80.	12	************
90.	8	********
100.	2	**
110.	1	*

X_3

MIDDLE OF INTERVAL	NUMBER OF OBSERVATIONS	
2.	1	*
3.	5	*****
4.	19	*******************
5.	9	*********
6.	3	***
7.	5	*****

X_4

MIDDLE OF INTERVAL	NUMBER OF OBSERVATIONS	
1.	13	*************
2.	15	***************
3.	8	********
4.	5	*****
5.	1	*

X_5

MIDDLE OF INTERVAL	NUMBER OF OBSERVATIONS	
5.	2	**
6.	3	***
7.	5	*****
8.	5	*****
9.	6	******
10.	4	****
11.	4	****
12.	5	*****
13.	4	****
14.	1	*
15.	0	
16.	1	*
17.	0	
18.	1	*
19.	0	
20.	0	
21.	1	*

X_6

MIDDLE OF INTERVAL	NUMBER OF OBSERVATIONS	
2.	3	***
4.	4	****
6.	7	*******
8.	7	*******
10.	8	********
12.	5	*****
14.	2	**
16.	2	**
18.	1	*
20.	0	
22.	0	
24.	2	**
26.	1	*

X_7

MIDDLE OF INTERVAL	NUMBER OF OBSERVATIONS	
2.	7	*******
3.	25	*************************
4.	9	*********
5.	1	*

Solution 1.6

1.6 **b)**

$$\bar{\underline{x}} = \begin{bmatrix} 7.5 \\ 73.857 \\ 4.548 \\ 2.191 \\ 10.048 \\ 9.405 \\ 3.095 \end{bmatrix} \quad S_n = \begin{bmatrix} 2.440 & -2.714 & -.369 & -.452 & -.571 & -2.179 & .167 \\ & 293.360 & 3.816 & -1.354 & 6.602 & 30.058 & .609 \\ & & 1.486 & .658 & 2.260 & 2.755 & .138 \\ & & & 1.154 & 1.062 & -.791 & .172 \\ & & & & 11.093 & 3.052 & 1.019 \\ & & & & & 30.241 & .580 \\ & & & & & & .467 \end{bmatrix}$$

(symmetric)

$$R = \begin{bmatrix} 1 & -.101 & -.194 & -.270 & -.110 & -.254 & .156 \\ & 1 & .183 & -.074 & .116 & .319 & .052 \\ & & 1 & .502 & .557 & .411 & .166 \\ & & & 1 & .297 & -.134 & .235 \\ & & & & 1 & .167 & .448 \\ & & & & & 1 & .154 \\ & & & & & & 1 \end{bmatrix}$$

(symmetric)

The pair x_3, x_4 exhibits a small to moderate positive correlation and so does the pair x_3, x_5. Most of the entries are small.

1.7

Exercise 1.7

1.7. You are given the following $n = 3$ observations on $p = 2$ variables:

Variable 1: $x_{11} = 2 \quad x_{21} = 3 \quad x_{31} = 4$

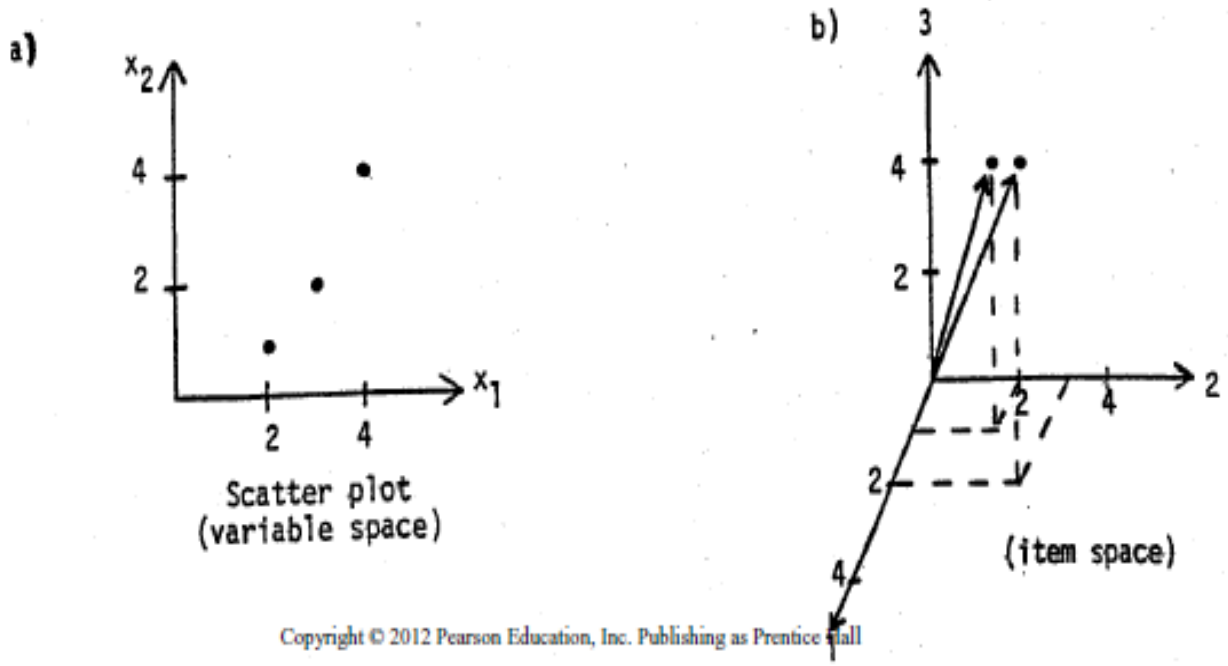
Variable 2: $x_{12} = 1 \quad x_{22} = 2 \quad x_{32} = 4$

(a) Plot the pairs of observations in the two-dimensional "variable space." That is, construct a two-dimensional scatter plot of the data.

(b) Plot the data as two points in the three-dimensional "item space."

Solution 1.7

1.7



Exercise 1.8

1.8. Evaluate the distance of the point $P = (-1, -1)$ to the point $Q = (1, 0)$ using the Euclidean distance formula in (1-12) with $p = 2$ and using the statistical distance in (1-20) with $a_{11} = 1/3$, $a_{22} = 4/27$, and $a_{12} = 1/9$. Sketch the locus of points that are a constant squared statistical distance 1 from the point Q.

Solution 1.8

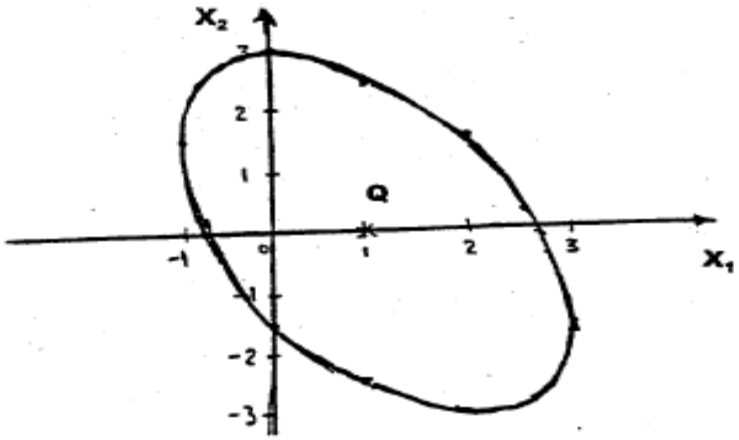
1.8 Using (1-12) $d(P,Q) = \sqrt{(-1-1)^2+(-1-0)^2} = \sqrt{5} = 2.236$

Using (1-20) $d(P,Q) = \sqrt{\frac{1}{3}(-1-1)^2+2(\frac{1}{9})(-1-1)(-1-0)+\frac{4}{27}(-1-0)^2} = \sqrt{\frac{52}{27}} = 1.388$

Using (1-20) the locus of points a constant squared distance 1 from Q = (1,0) is given by the expression $\frac{1}{3}(x_1-1)^2+\frac{2}{9}(x_1-1)x_2+\frac{4}{27}x_2^2 = 1$. To sketch the locus of points defined by this equation, we first obtain the coordinates of some points satisfying the equation:

(-1,1.5), (0,-1.5), (0,3), (1,-2.6), (1,2.6), (2,-3), (2,1.5), (3,-1.5)

The resulting ellipse is:



Exercise 1.9

1.9. Consider the following eight pairs of measurements on two variables x_1 and x_2:

x_1	−6	−3	−2	1	2	5	6	8
x_2	−2	−3	1	−1	2	1	5	3

(a) Plot the data as a scatter diagram, and compute s_{11}, s_{22}, and s_{12}.

(b) Using (1-18), calculate the corresponding measurements on variables $\tilde{x}_1$ and $\tilde{x}_2$, assuming that the original coordinate axes are rotated through an angle of $\theta = 26°$ [given $\cos(26°) = .899$ and $\sin(26°) = .438$].

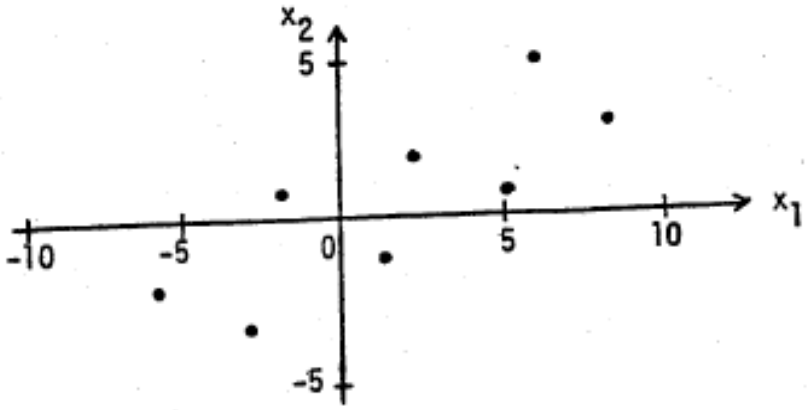
(c) Using the $\tilde{x}_1$ and $\tilde{x}_2$ measurements from (b), compute the sample variances $\tilde{s}_{11}$ and $\tilde{s}_{22}$.

(d) Consider the *new* pair of measurements $(x_1, x_2) = (4, -2)$. Transform these to measurements on $\tilde{x}_1$ and $\tilde{x}_2$ using (1-18), and calculate the distance $d(O, P)$ of the new point $P = (\tilde{x}_1, \tilde{x}_2)$ from the origin $O = (0, 0)$ using (1-17).
Note: You will need $\tilde{s}_{11}$ and $\tilde{s}_{22}$ from (c).

(e) Calculate the distance from $P = (4, -2)$ to the origin $O = (0, 0)$ using (1-19) and the expressions for a_{11}, a_{22}, and a_{12} in footnote 2.
Note: You will need s_{11}, s_{22}, and s_{12} from (a).
Compare the distance calculated here with the distance calculated using the $\tilde{x}_1$ and $\tilde{x}_2$ values in (d). (Within rounding error, the numbers should be the same.)

Solution 1.9

1.9 **a)** $s_{11} = 20.48$ $s_{22} = 6.19$ $s_{12} = 9.09$



Solution 1.9

1.9 b)

$\tilde{x}_1$	-6.20	-4.10	-1.23	.37	2.73	4.83	7.70	8.43
$\tilde{x}_2$	1.27	-1.10	1.87	-1.37	.73	-1.63	1.33	-1.40

c)

$\tilde{s}_{11} = 24.90$ $\tilde{s}_{22} = 1.77$ (Note $\tilde{s}_{12} = .00$)

d)

$(\bar{\tilde{x}}_1, \bar{\tilde{x}}_2) = (2.72, -3.55)$

$d(0,P) = 2.72$ using (1-17).

e) $d(0,P) = 2.72$ using (1-19).

Exercise 1.10

1.10. Are the following distance functions valid for distance from the origin? Explain.

(a) $x_1^2 + 4x_2^2 + x_1x_2 = (\text{distance})^2$

(b) $x_1^2 - 2x_2^2 = (\text{distance})^2$

Solution 1.10

1.10 a) This equation is of the form (1-19) with $a_{11} = 1$, $a_{12} = \frac{1}{2}$ and $a_{22} = 4$. Therefore this is a distance for correlated variables if it is non-negative for all values of x_1, x_2. But this follows easily if we write

$$x_1^2 + 4x_2^2 + x_1x_2 = (x_1 + \tfrac{1}{2}x_2)^2 + \tfrac{15}{4}x_2^2 \geq 0.$$

b) In order for this expression to be a distance it has to be non-negative for all values x_1, x_2. Since, for $(x_1,x_2) = (0,1)$ we have $x_1^2 - 2x_2^2 = -2$, we conclude that this is not a valid distance function.

Exercise 1.11

1.11. Verify that distance defined by (1-20) with $a_{11} = 4$, $a_{22} = 1$, and $a_{12} = -1$ satisfies the first three conditions in (1-25). (The triangle inequality is more difficult to verify.)

Solution 1.11

1.11

$$d(P,Q) = \sqrt{4(x_1-y_1)^2 + 2(-1)(x_1-y_1)(x_2-y_2) + (x_2-y_2)^2}$$

$$= \sqrt{4(y_1-x_1)^2 + 2(-1)(y_1-x_1)(y_2-x_2) + (x_2-y_2)^2} = d(Q,P)$$

Next, $4(x_1-y_1)^2 - 2(x_1-y_1)(x_2-y_2) + (x_2-y_2)^2 =$

$$= (x_1-y_1-x_2+y_2)^2 + 3(x_1-y_1)^2 \geq 0 \quad \text{so} \quad d(P,Q) \geq 0.$$

The second term is zero in this last expression only if $x_1 = y_1$ and then the first is zero only if $x_2 = y_2$.

Exercise 1.12

1.12. Define the distance from the point $P = (x_1, x_2)$ to the origin $O = (0, 0)$ as

$$d(O, P) = \max(|x_1|, |x_2|)$$

(a) Compute the distance from $P = (-3, 4)$ to the origin.

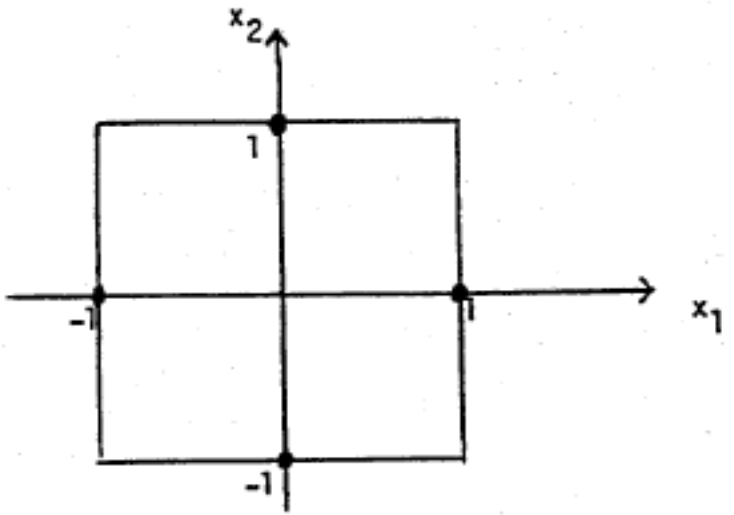
(b) Plot the locus of points whose squared distance from the origin is 1.

(c) Generalize the foregoing distance expression to points in p dimensions.

Solution 1.12

1.12 a) If P = (-3,4) then $d(0,P) = \max(|-3|, |4|) = 4$

b) The locus of points whose squared distance from (0,0) is 1 is



c) The generalization to p-dimensions is given by $d(0,P) = \max(|x_1|, |x_2|, \ldots, |x_p|)$.

Exercise 1.14

1.14. Table 1.6 contains some of the raw data discussed in Section 1.2. (See also the multiple-sclerosis data on the web at www.prenhall.com/statistics.) Two different visual stimuli ($S1$ and $S2$) produced responses in both the left eye (L) and the right eye (R) of subjects in the study groups. The values recorded in the table include x_1 (subject's age); x_2 (total response of both eyes to stimulus $S1$, that is, $S1L + S1R$); x_3 (difference between responses of eyes to stimulus $S1$, $|S1L - S1R|$); and so forth.

(a) Plot the two-dimensional scatter diagram for the variables x_2 and x_4 for the multiple-sclerosis group. Comment on the appearance of the diagram.

(b) Compute the $\bar{\mathbf{x}}$, $\mathbf{S}_n$, and $\mathbf{R}$ arrays for the non-multiple-sclerosis and multiple-sclerosis groups separately.

Exercise 1.14

Table 1.6 Multiple-Sclerosis Data

Non-Multiple-Sclerosis Group Data

Subject number	x_1 (Age)	x_2 $(S1L + S1R)$	x_3 $\lvert S1L - S1R \rvert$	x_4 $(S2L + S2R)$	x_5 $\lvert S2L - S2R \rvert$
1	18	152.0	1.6	198.4	.0
2	19	138.0	.4	180.8	1.6
3	20	144.0	.0	186.4	.8
4	20	143.6	3.2	194.8	.0
5	20	148.8	.0	217.6	.0
⋮	⋮	⋮	⋮	⋮	⋮
65	67	154.4	2.4	205.2	6.0
66	69	171.2	1.6	210.4	.8
67	73	157.2	.4	204.8	.0
68	74	175.2	5.6	235.6	.4
69	79	155.0	1.4	204.4	.0

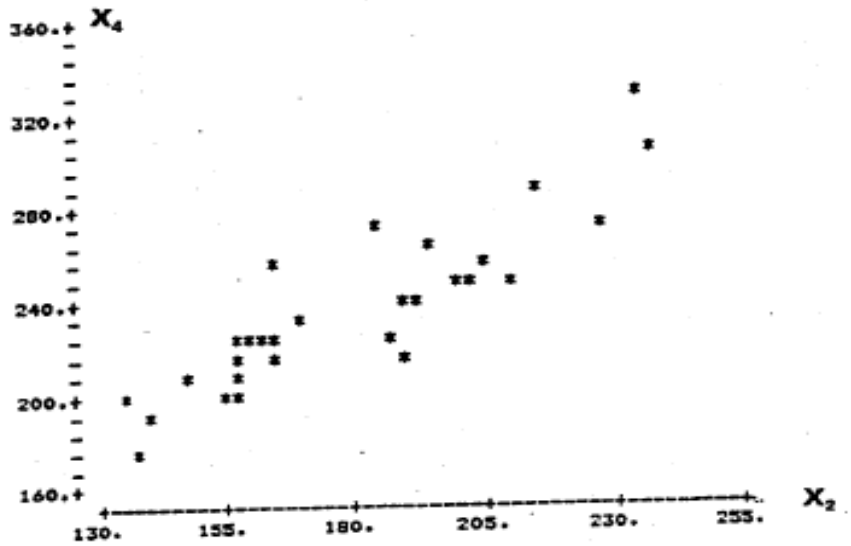
Multiple-Sclerosis Group Data

Subject number	x_1	x_2	x_3	x_4	x_5
1	23	148.0	.8	205.4	.6
2	25	195.2	3.2	262.8	.4
3	25	158.0	8.0	209.8	12.2
4	28	134.4	.0	198.4	3.2
5	29	190.2	14.2	243.8	10.6
⋮	⋮	⋮	⋮	⋮	⋮
25	57	165.6	16.8	229.2	15.6
26	58	238.4	8.0	304.4	6.0
27	58	164.0	.8	216.8	.8
28	58	169.8	.0	219.2	1.6
29	59	199.8	4.6	250.2	1.0

Source: Data courtesy of Dr. G. G. Celesia.

Solution 1.14

1.14 a)



Strong positive correlation. No obvious "unusual" observations.

b) Multiple-sclerosis group.

$$\bar{\underline{x}} = \begin{pmatrix} 42.07 \\ 179.64 \\ 12.31 \\ 236.62 \\ 13.16 \end{pmatrix}$$

Solution 1.14

b) Multiple-sclerosis group.

$$\bar{\underset{\sim}{x}} = \begin{pmatrix} 42.07 \\ 179.64 \\ 12.31 \\ 236.62 \\ 13.16 \end{pmatrix}$$

$$S_n = \begin{pmatrix} 116.91 & 61.78 & -20.10 & 61.13 & -27.65 \\ & 812.72 & 218.35 & 865.32 & 90.48 \\ & & 305.94 & 221.93 & 286.60 \\ & & & 1146.38 & 82.53 \\ & (\text{symmetric}) & & & 337.80 \end{pmatrix}$$

$$R = \begin{pmatrix} 1 & .200 & -.106 & .167 & -.139 \\ & 1 & .438 & .896 & .173 \\ & & 1 & .375 & .892 \\ & & & 1 & .133 \\ & (\text{symmetric}) & & & 1 \end{pmatrix}$$

Solution 1.14

Non multiple-sclerosis group.

$$\bar{x} = \begin{pmatrix} 37.99 \\ 147.21 \\ 1.56 \\ 195.57 \\ 1.62 \end{pmatrix}$$

$$S_n = \begin{pmatrix} 273.61 & 95.08 & 5.28 & 101.67 & 3.20 \\ & 110.13 & 1.84 & 103.28 & 2.15 \\ & & 1.78 & 2.22 & .49 \\ & & & 183.04 & 2.35 \\ & (\text{symmetric}) & & & 2.32 \end{pmatrix}$$

$$R = \begin{pmatrix} 1 & .548 & .239 & .454 & .127 \\ & 1 & .132 & .727 & .134 \\ & & 1 & .123 & .244 \\ & & & 1 & .114 \\ & (\text{symmetric}) & & & 1 \end{pmatrix}$$

Exercise 1.15

1.15. Some of the 98 measurements described in Section 1.2 are listed in Table 1.7 (See also the radiotherapy data on the web at www.prenhall.com/statistics.) The data consist of average ratings over the course of treatment for patients undergoing radiotherapy. Variables measured include x_1 (number of symptoms, such as sore throat or nausea); x_2 (amount of activity, on a 1–5 scale); x_3 (amount of sleep, on a 1–5 scale); x_4 (amount of food consumed, on a 1–3 scale); x_5 (appetite, on a 1–5 scale); and x_6 (skin reaction, on a 0–3 scale).

(a) Construct the two-dimensional scatter plot for variables x_2 and x_3 and the marginal dot diagrams (or histograms). Do there appear to be any errors in the x_3 data?

(b) Compute the $\bar{\mathbf{x}}$, $\mathbf{S}_n$, and $\mathbf{R}$ arrays. Interpret the pairwise correlations.

Exercise 1.15

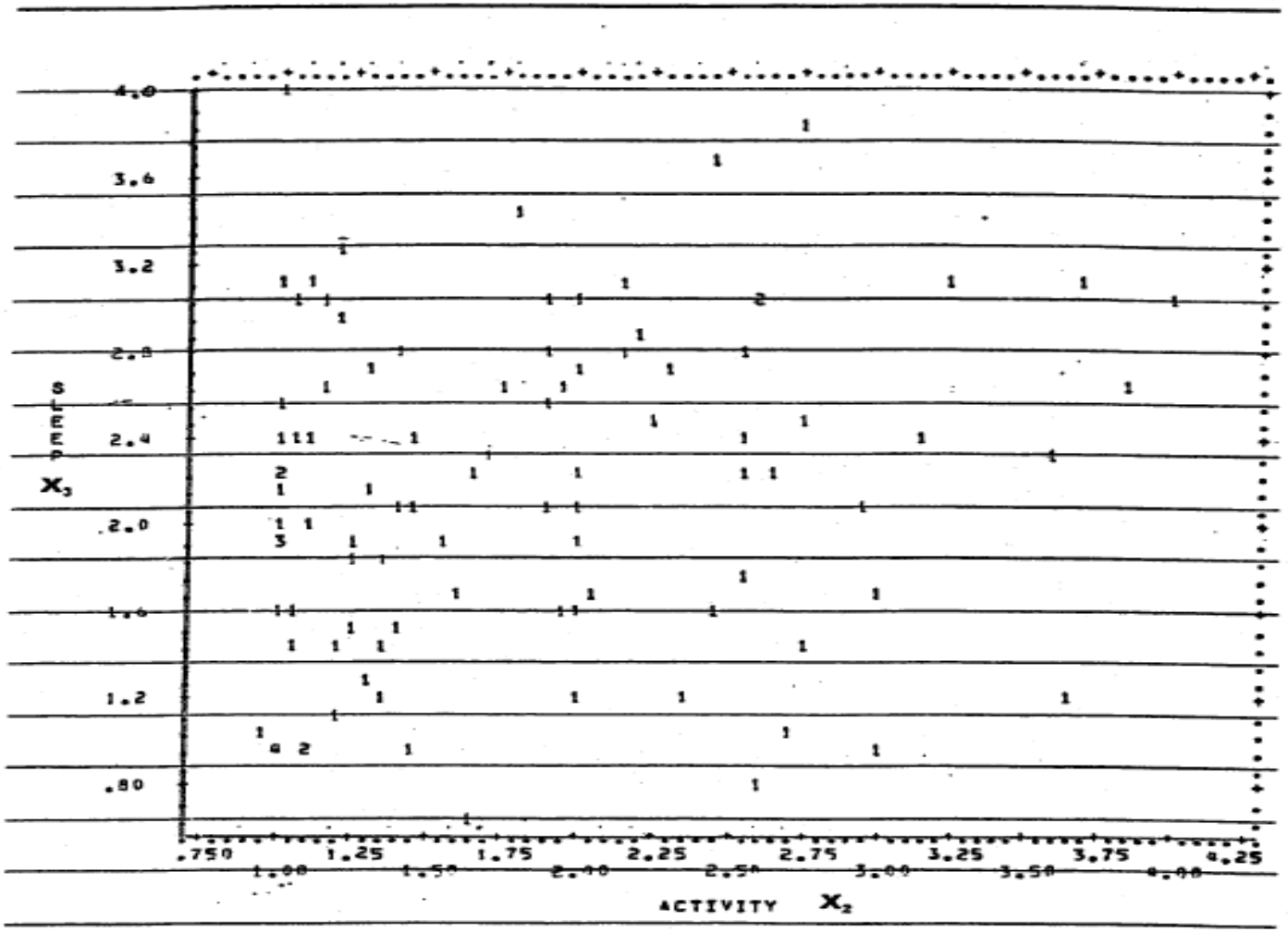
Table 1.7 Radiotherapy Data

x_1 Symptoms	x_2 Activity	x_3 Sleep	x_4 Eat	x_5 Appetite	x_6 Skin reaction
.889	1.389	1.555	2.222	1.945	1.000
2.813	1.437	.999	2.312	2.312	2.000
1.454	1.091	2.364	2.455	2.909	3.000
.294	.941	1.059	2.000	1.000	1.000
2.727	2.545	2.819	2.727	4.091	.000
⋮	⋮	⋮	⋮	⋮	⋮
4.100	1.900	2.800	2.000	2.600	2.000
.125	1.062	1.437	1.875	1.563	.000
6.231	2.769	1.462	2.385	4.000	2.000
3.000	1.455	2.090	2.273	3.272	2.000
.889	1.000	1.000	2.000	1.000	2.000

Source: Data courtesy of Mrs. Annette Tealey, R.N. Values of x_2 and x_3 less than 1.0 are due to errors in the data-collection process. Rows containing values of x_2 and x_3 less than 1.0 may be omitted.

Solution 1.15

1.15 a) Scatterplot of x_2 and x_3.



Solution 1.15

b)

$$\bar{\underset{\sim}{x}} = \begin{pmatrix} 3.54 \\ 1.81 \\ 2.14 \\ 2.21 \\ 2.58 \\ 1.27 \end{pmatrix}$$

Solution 1.15

1.15

$$S_n = \begin{pmatrix} 4.61 & .92 & .58 & .27 & 1.06 & .15 \\ & .61 & .11 & .12 & .39 & -.02 \\ & & .57 & .09 & .34 & .11 \\ & & & .11 & .21 & .02 \\ & & & & .85 & -.01 \\ & (\text{symmetric}) & & & & .85 \end{pmatrix}$$

$$R = \begin{pmatrix} 1 & .551 & .362 & .386 & .537 & .077 \\ & 1 & .187 & .455 & .535 & -.035 \\ & & 1 & .346 & .496 & .156 \\ & & & 1 & .704 & .071 \\ & & & & 1 & -.010 \\ & (\text{symmetric}) & & & & 1 \end{pmatrix}$$

The largest correlation is between appetite and amount of food eaten. Both activity and appetite have moderate positive correlations with symptoms. Also, appetite and activity have a moderate positive correlation.

Exercise 1.16

1.16. At the start of a study to determine whether exercise or dietary supplements would slow bone loss in older women, an investigator measured the mineral content of bones by photon absorptiometry. Measurements were recorded for three bones on the dominant and nondominant sides and are shown in Table 1.8. (See also the mineral-content data on the web at www.prenhall.com/statistics.)

Compute the $\bar{\mathbf{x}}$, $\mathbf{S}_n$, and $\mathbf{R}$ arrays. Interpret the pairwise correlations.

Exercise 1.16

Table 1.8 Mineral Content in Bones

Subject number	Dominant radius	Radius	Dominant humerus	Humerus	Dominant ulna	Ulna
1	1.103	1.052	2.139	2.238	.873	.872
2	.842	.859	1.873	1.741	.590	.744
3	.925	.873	1.887	1.809	.767	.713
4	.857	.744	1.739	1.547	.706	.674
5	.795	.809	1.734	1.715	.549	.654
6	.787	.779	1.509	1.474	.782	.571
7	.933	.880	1.695	1.656	.737	.803
8	.799	.851	1.740	1.777	.618	.682
9	.945	.876	1.811	1.759	.853	.777
10	.921	.906	1.954	2.009	.823	.765
11	.792	.825	1.624	1.657	.686	.668
12	.815	.751	2.204	1.846	.678	.546
13	.755	.724	1.508	1.458	.662	.595
14	.880	.866	1.786	1.811	.810	.819
15	.900	.838	1.902	1.606	.723	.677
16	.764	.757	1.743	1.794	.586	.541
17	.733	.748	1.863	1.869	.672	.752
18	.932	.898	2.028	2.032	.836	.805
19	.856	.786	1.390	1.324	.578	.610
20	.890	.950	2.187	2.087	.758	.718
21	.688	.532	1.650	1.378	.533	.482
22	.940	.850	2.334	2.225	.757	.731
23	.493	.616	1.037	1.268	.546	.615
24	.835	.752	1.509	1.422	.618	.664
25	.915	.936	1.971	1.869	.869	.868

Source: Data courtesy of Everett Smith.

Solution 1.16

1.16

There are significant positive correlations among all variables. The lowest correlation is 0.4420 between Dominant humerus and Ulna, and the highest correlation is 0.89365 bewteen Dominant hemerus and Hemerus.

$$\bar{x} = \begin{pmatrix} 0.8438 \\ 0.8183 \\ 1.7927 \\ 1.7348 \\ 0.7044 \\ 0.6938 \end{pmatrix}, \quad R = \begin{pmatrix} 1.00000 & 0.85181 & 0.69146 & 0.66826 & 0.74369 & 0.67789 \\ 0.85181 & 1.00000 & 0.61192 & 0.74909 & 0.74218 & 0.80980 \\ 0.69146 & 0.61192 & 1.00000 & 0.89365 & 0.55222 & 0.44020 \\ 0.66826 & 0.74909 & 0.89365 & 1.00000 & 0.62555 & 0.61882 \\ 0.74369 & 0.74218 & 0.55222 & 0.62555 & 1.00000 & 0.72889 \\ 0.67789 & 0.80980 & 0.44020 & 0.61882 & 0.72889 & 1.00000 \end{pmatrix},$$

$$S_n = \begin{pmatrix} 0.0124815 & 0.0099633 & 0.0214560 & 0.0192822 & 0.0087559 & 0.0076395 \\ 0.0099633 & 0.0109612 & 0.0177938 & 0.0202555 & 0.0081886 & 0.0085522 \\ 0.0214560 & 0.0177938 & 0.0771429 & 0.0641052 & 0.0161635 & 0.0123332 \\ 0.0192822 & 0.0202555 & 0.0641052 & 0.0667051 & 0.0170261 & 0.0161219 \\ 0.0087559 & 0.0081886 & 0.0161635 & 0.0170261 & 0.0111057 & 0.0077483 \\ 0.0076395 & 0.0085522 & 0.0123332 & 0.0161219 & 0.0077483 & 0.0101752 \end{pmatrix}.$$

Exercise 1.17

1.17. Some of the data described in Section 1.2 are listed in Table 1.9. (See also the national-track-records data on the web at www.prenhall.com/statistics.) The national track records for women in 54 countries can be examined for the relationships among the running events. Compute the $\bar{\mathbf{x}}$, $\mathbf{S}_n$, and $\mathbf{R}$ arrays. Notice the magnitudes of the correlation coefficients as you go from the shorter (100-meter) to the longer (marathon) running distances. Interpret these pairwise correlations.

Exercise 1.17

Table 1.9 National Track Records for Women

Country	100 m (s)	200 m (s)	400 m (s)	800 m (min)	1500 m (min)	3000 m (min)	Marathon (min)
Argentina	11.57	22.94	52.50	2.05	4.25	9.19	150.32
Australia	11.12	22.23	48.63	1.98	4.02	8.63	143.51
Austria	11.15	22.70	50.62	1.94	4.05	8.78	154.35
Belgium	11.14	22.48	51.45	1.97	4.08	8.82	143.05
Bermuda	11.46	23.05	53.30	2.07	4.29	9.81	174.18
Brazil	11.17	22.60	50.62	1.97	4.17	9.04	147.41
Canada	10.98	22.62	49.91	1.97	4.00	8.54	148.36
Chile	11.65	23.84	53.68	2.00	4.22	9.26	152.23
China	10.79	22.01	49.81	1.93	3.84	8.10	139.39
Columbia	11.31	22.92	49.64	2.04	4.34	9.37	155.19
Cook Islands	12.52	25.91	61.65	2.28	4.82	11.10	212.33
Costa Rica	11.72	23.92	52.57	2.10	4.52	9.84	164.33
Czech Republic	11.09	21.97	47.99	1.89	4.03	8.87	145.19
Denmark	11.42	23.36	52.92	2.02	4.12	8.71	149.34
Dominican Republic	11.63	23.91	53.02	2.09	4.54	9.89	166.46
Finland	11.13	22.39	50.14	2.01	4.10	8.69	148.00
France	10.73	21.99	48.25	1.94	4.03	8.64	148.27
Germany	10.81	21.71	47.60	1.92	3.96	8.51	141.45
Great Britain	11.10	22.10	49.43	1.94	3.97	8.37	135.25
Greece	10.83	22.67	50.56	2.00	4.09	8.96	153.40
Guatemala	11.92	24.50	55.64	2.15	4.48	9.71	171.33
Hungary	11.41	23.06	51.50	1.99	4.02	8.55	148.50
India	11.56	23.86	55.08	2.10	4.36	9.50	154.29
Indonesia	11.38	22.82	51.05	2.00	4.10	9.11	158.10
Ireland	11.43	23.02	51.07	2.01	3.98	8.36	142.23
Israel	11.45	23.15	52.06	2.07	4.24	9.33	156.36
Italy	11.14	22.60	51.31	1.96	3.98	8.59	143.47
Japan	11.36	23.33	51.93	2.01	4.16	8.74	139.41
Kenya	11.62	23.37	51.56	1.97	3.96	8.39	138.47
Korea, South	11.49	23.80	53.67	2.09	4.24	9.01	146.12
Korea, North	11.80	25.10	56.23	1.97	4.25	8.96	145.31
Luxembourg	11.76	23.96	56.07	2.07	4.35	9.21	149.23
Malaysia	11.50	23.37	52.56	2.12	4.39	9.31	169.28
Mauritius	11.72	23.83	54.62	2.06	4.33	9.24	167.09
Mexico	11.09	23.13	48.89	2.02	4.19	8.89	144.06

Exercise 1.17

Netherlands	11.08	22.81	51.35	1.93	4.06	8.57	143.43
New Zealand	11.32	23.13	51.60	1.97	4.10	8.76	146.46
Norway	11.41	23.31	52.45	2.03	4.01	8.53	141.06
Papua New Guinea	11.96	24.68	55.18	2.24	4.62	10.21	221.14
Philippines	11.28	23.35	54.75	2.12	4.41	9.81	165.48
Poland	10.93	22.13	49.28	1.95	3.99	8.53	144.18
Portugal	11.30	22.88	51.92	1.98	3.96	8.50	143.29
Romania	11.30	22.35	49.88	1.92	3.90	8.36	142.50
Russia	10.77	21.87	49.11	1.91	3.87	8.38	141.31
Samoa	12.38	25.45	56.32	2.29	5.42	13.12	191.58

(continues)

Country	100 m (s)	200 m (s)	400 m (s)	800 m (min)	1500 m (min)	3000 m (min)	Marathon (min)
Singapore	12.13	24.54	55.08	2.12	4.52	9.94	154.41
Spain	11.06	22.38	49.67	1.96	4.01	8.48	146.51
Sweden	11.16	22.82	51.69	1.99	4.09	8.81	150.39
Switzerland	11.34	22.88	51.32	1.98	3.97	8.60	145.51
Taiwan	11.22	22.56	52.74	2.08	4.38	9.63	159.53
Thailand	11.33	23.30	52.60	2.06	4.38	10.07	162.39
Turkey	11.25	22.71	53.15	2.01	3.92	8.53	151.43
U.S.A.	10.49	21.34	48.83	1.94	3.95	8.43	141.16

Source: *IAAF/ATFS Track and Field Handbook for Helsinki* 2005 (courtesy of Ottavio Castellini).

Solution 1.17

1.17

There are large positive correlations among all variables. Particularly large correlations occur between running events that are "similar", for example, the 100m and 200m dashes, and the 1500m and 3000m runs.

$$\mathbf{x} = \begin{bmatrix} 11.36 \\ 23.12 \\ 51.99 \\ 2.02 \\ 4.19 \\ 9.08 \\ 153.62 \end{bmatrix} \quad \mathbf{S}_n = \begin{bmatrix} .152 & .338 & .875 & .027 & .082 & .230 & 4.254 \\ .338 & .847 & 2.152 & .065 & .199 & .544 & 10.193 \\ .875 & 2.152 & 6.621 & .178 & .500 & 1.400 & 28.368 \\ .027 & .065 & .178 & .007 & .021 & .060 & 1.197 \\ .082 & .199 & .500 & .021 & .073 & .212 & 3.474 \\ .230 & .544 & 1.400 & .060 & .212 & .652 & 10.508 \\ 4.254 & 10.193 & 28.368 & 1.197 & 3.474 & 10.508 & 265.265 \end{bmatrix}$$

$$\mathbf{R} = \begin{bmatrix} 1.000 & .941 & .871 & .809 & .782 & .728 & .669 \\ .941 & 1.000 & .909 & .820 & .801 & .732 & .680 \\ .871 & .909 & 1.000 & .806 & .720 & .674 & .677 \\ .809 & .820 & .806 & 1.000 & .905 & .867 & .854 \\ .782 & .801 & .720 & .905 & 1.000 & .973 & .791 \\ .728 & .732 & .674 & .867 & .973 & 1.000 & .799 \\ .669 & .680 & .677 & .854 & .791 & .799 & 1.000 \end{bmatrix}$$

Exercise 1.18

1.18. Convert the national track records for women in Table 1.9 to speeds measured in meters per second. For example, the record speed for the 100-m dash for Argentinian women is 100 m/11.57 sec = 8.643 m/sec. Notice that the records for the 800-m, 1500-m, 3000-m and marathon runs are measured in minutes. The marathon is 26.2 miles, or 42,195 meters, long. Compute the $\bar{\mathbf{x}}$, $\mathbf{S}_n$, and $\mathbf{R}$ arrays. Notice the magnitudes of the correlation coefficients as you go from the shorter (100 m) to the longer (marathon) running distances. Interpret these pairwise correlations. Compare your results with the results you obtained in Exercise 1.17.

Solution 1.18

1.18

There are positive correlations among all variables. Notice the correlations decrease as the distances between pairs of running events increase (see the first column of the correlation matrix **R**). The correlation matrix for running events measured in meters per second is very similar to the correlation matrix for the running event times given in Exercise 1.17.

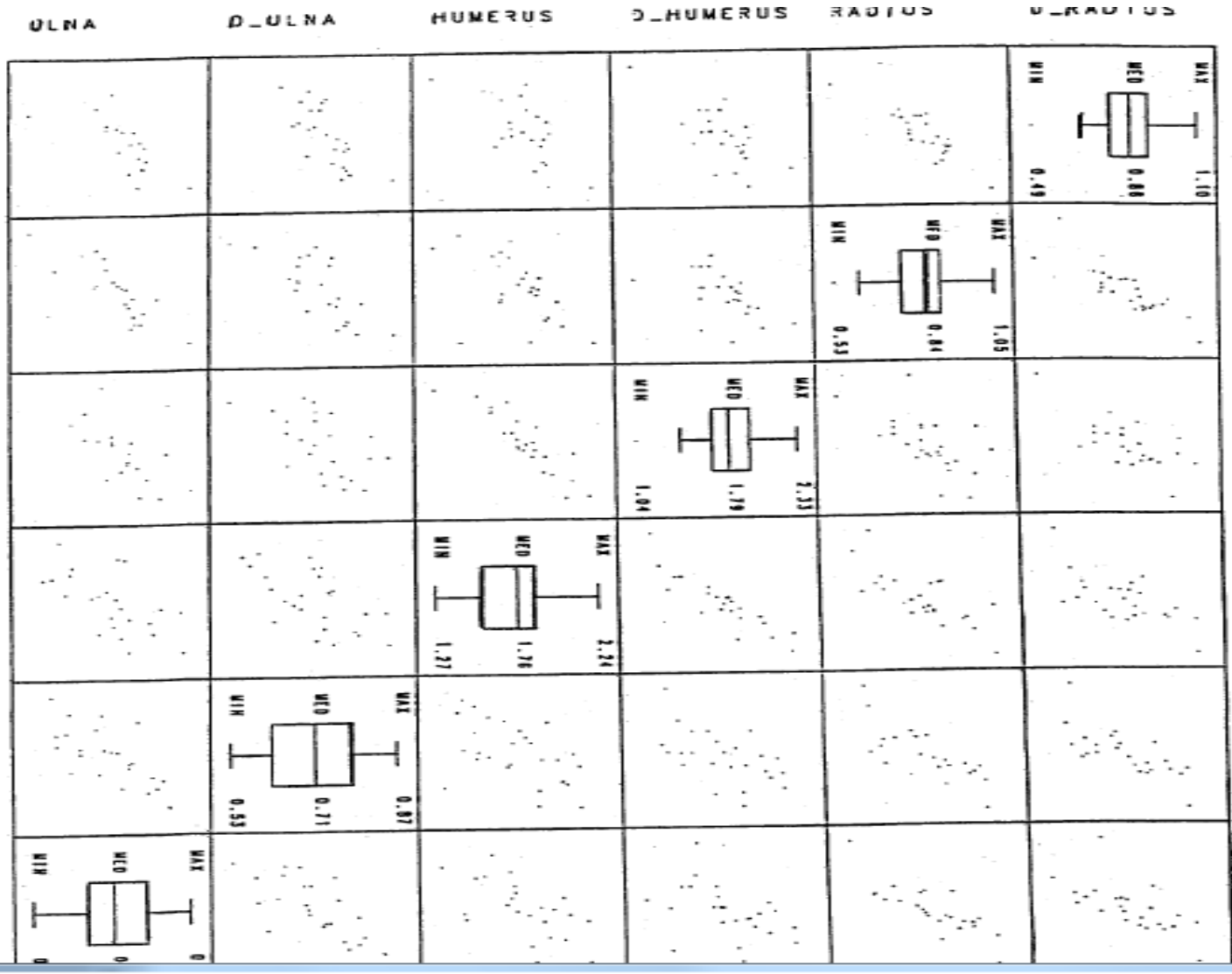
$$\bar{\mathbf{x}} = \begin{bmatrix} 8.81 \\ 8.66 \\ 7.71 \\ 6.60 \\ 5.99 \\ 5.54 \\ 4.62 \end{bmatrix} \qquad \mathbf{S}_n = \begin{bmatrix} .091 & .096 & .097 & .065 & .082 & .092 & .081 \\ .096 & .115 & .114 & .075 & .096 & .105 & .093 \\ .097 & .114 & .138 & .081 & .095 & .108 & .102 \\ .065 & .075 & .081 & .074 & .086 & .100 & .094 \\ .082 & .096 & .095 & .086 & .124 & .144 & .118 \\ .092 & .105 & .108 & .100 & .144 & .177 & .147 \\ .081 & .093 & .102 & .094 & .118 & .147 & .167 \end{bmatrix}$$

$$\mathbf{R} = \begin{bmatrix} 1.000 & .938 & .866 & .797 & .776 & .729 & .660 \\ .938 & 1.000 & .906 & .816 & .806 & .741 & .675 \\ .866 & .906 & 1.000 & .804 & .731 & .694 & .672 \\ .797 & .816 & .804 & 1.000 & .906 & .875 & .852 \\ .776 & .806 & .731 & .906 & 1.000 & .972 & .824 \\ .729 & .741 & .694 & .875 & .972 & 1.000 & .854 \\ .660 & .675 & .672 & .852 & .824 & .854 & 1.000 \end{bmatrix}$$

Exercise 1.19

1.19. Create the scatter plot and boxplot displays of Figure 1.5 for (a) the mineral-content data in Table 1.8 and (b) the national-track-records data in Table 1.9.

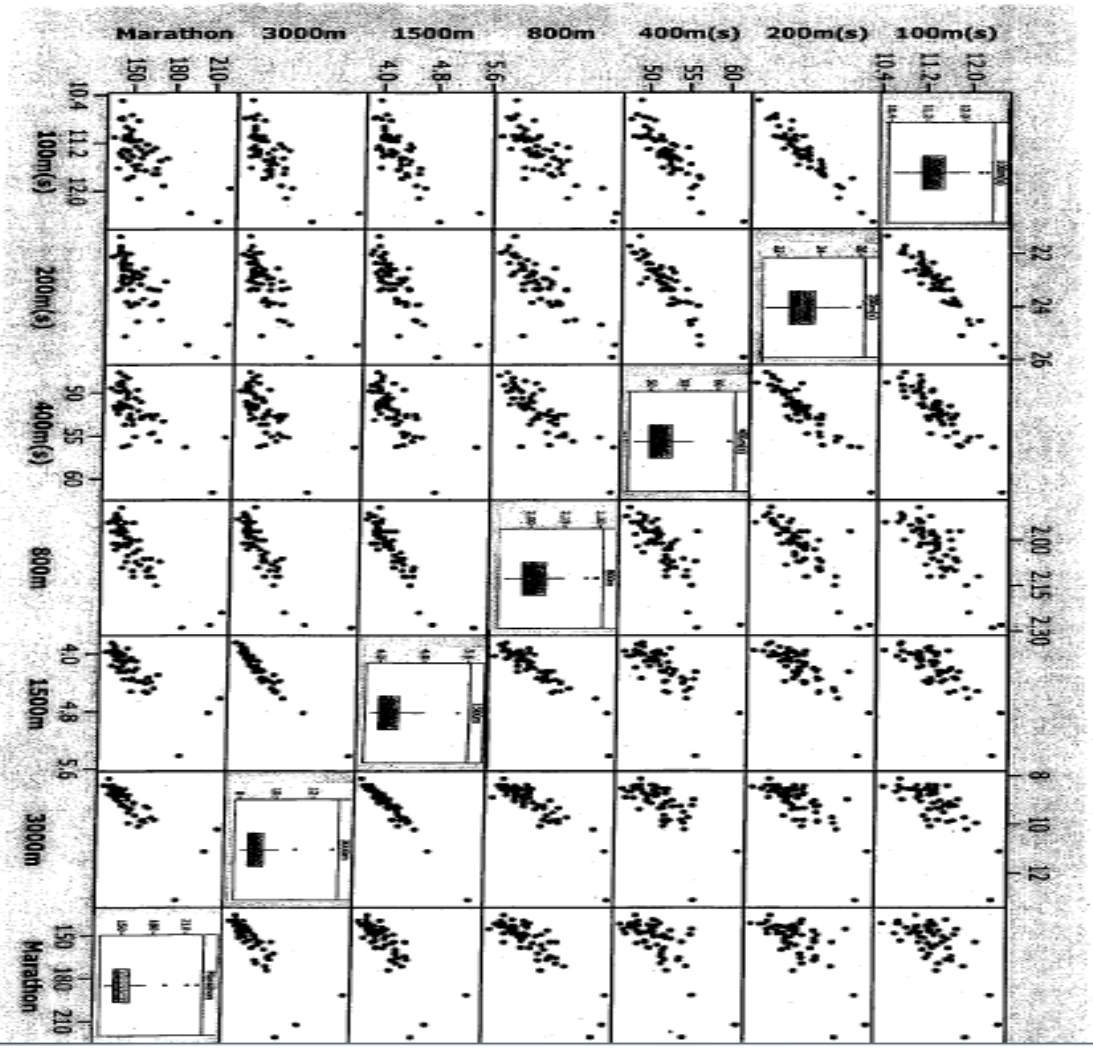
Solution 1.19



Solution 1.19

1.19

(b)



Exercise 1.20

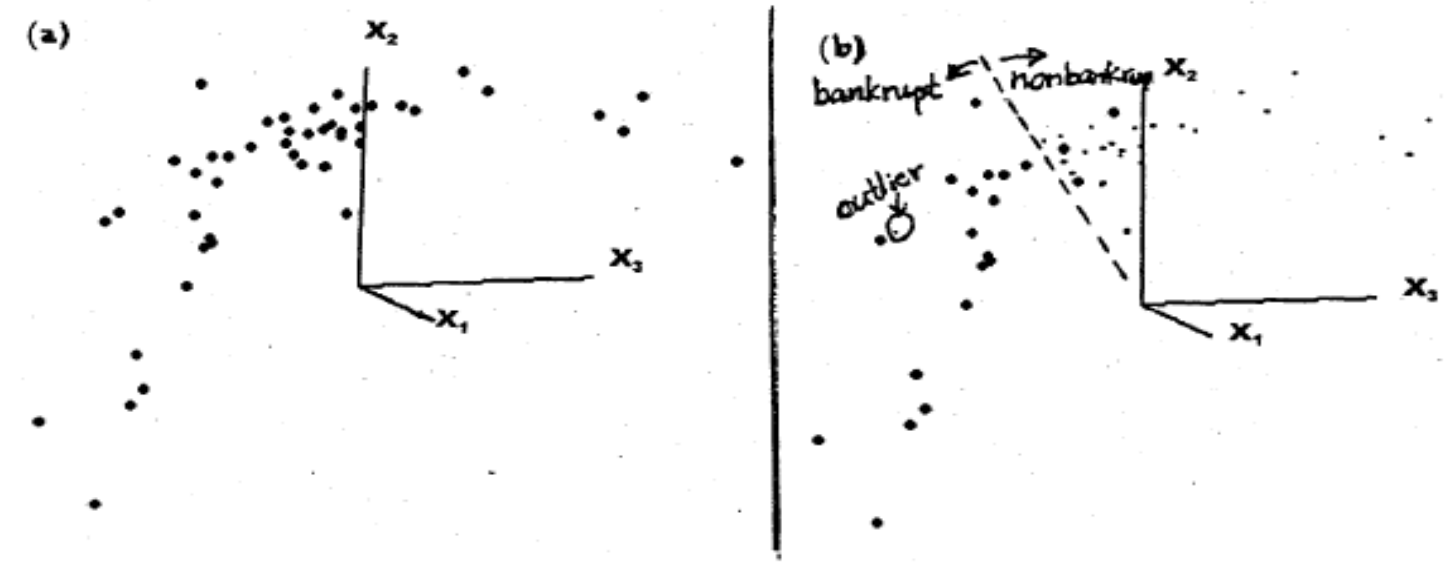
1.20. Refer to the bankruptcy data in Table 11.4, page 657, and on the following website www.prenhall.com/statistics. Using appropriate computer software,

(a) View the entire data set in x_1, x_2, x_3 space. Rotate the coordinate axes in various directions. Check for unusual observations.

(b) Highlight the set of points corresponding to the bankrupt firms. Examine various three-dimensional perspectives. Are there some orientations of three-dimensional space for which the bankrupt firms can be distinguished from the nonbankrupt firms? Are there observations in each of the two groups that are likely to have a significant impact on any rule developed to classify firms based on the sample means, variances, and covariances calculated from these data? (See Exercise 11.24.)

Solution 1.20

1.20



(a) The plot looks like a cigar shape, but bent. Some observations in the lower left hand part could be outliers. From the highlighted plot in (b) (actually non-bankrupt group not highlighted), there is one outlier in the nonbankrupt group, which is apparently located in the bankrupt group, besides the strung out pattern to the right.

(b) The dotted line in the plot would be an orientation for the classification.

Exercise 1.21

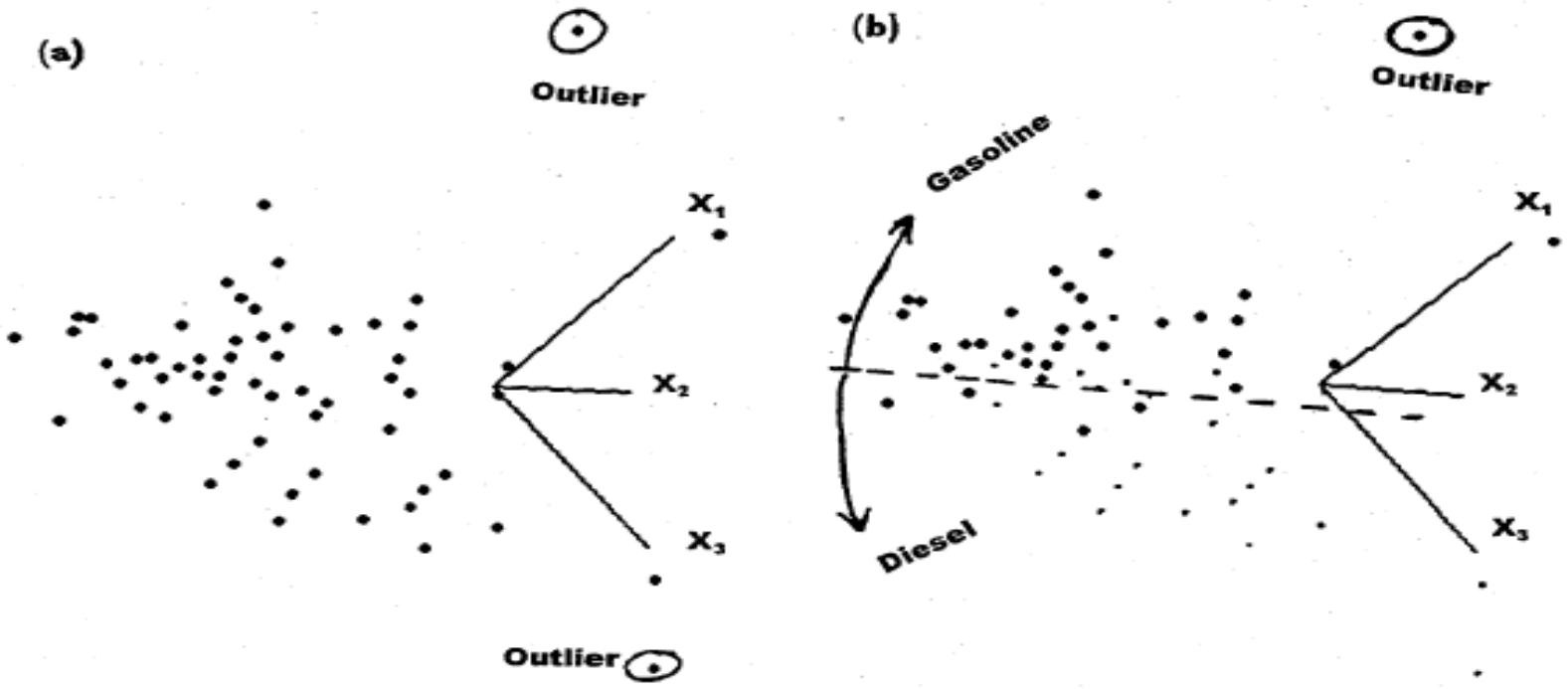
1.21. Refer to the milk transportation-cost data in Table 6.10, page 345, and on the web at www.prenhall.com/statistics. Using appropriate computer software,

(a) View the entire data set in three dimensions. Rotate the coordinate axes in various directions. Check for unusual observations.

(b) Highlight the set of points corresponding to gasoline trucks. Do any of the gasoline-truck points appear to be multivariate outliers? (See Exercise 6.17.) Are there some orientations of x_1, x_2, x_3 space for which the set of points representing gasoline trucks can be readily distinguished from the set of points representing diesel trucks?

Solution 1.21

1.21



(a) There are two outliers in the upper right and lower right corners of the plot.

(b) Only the points in the gasoline group are highlighted. The observation in the upper right is the outlier. As indicated in the plot, there is an orientation to classify into two groups.

Exercise 1.22

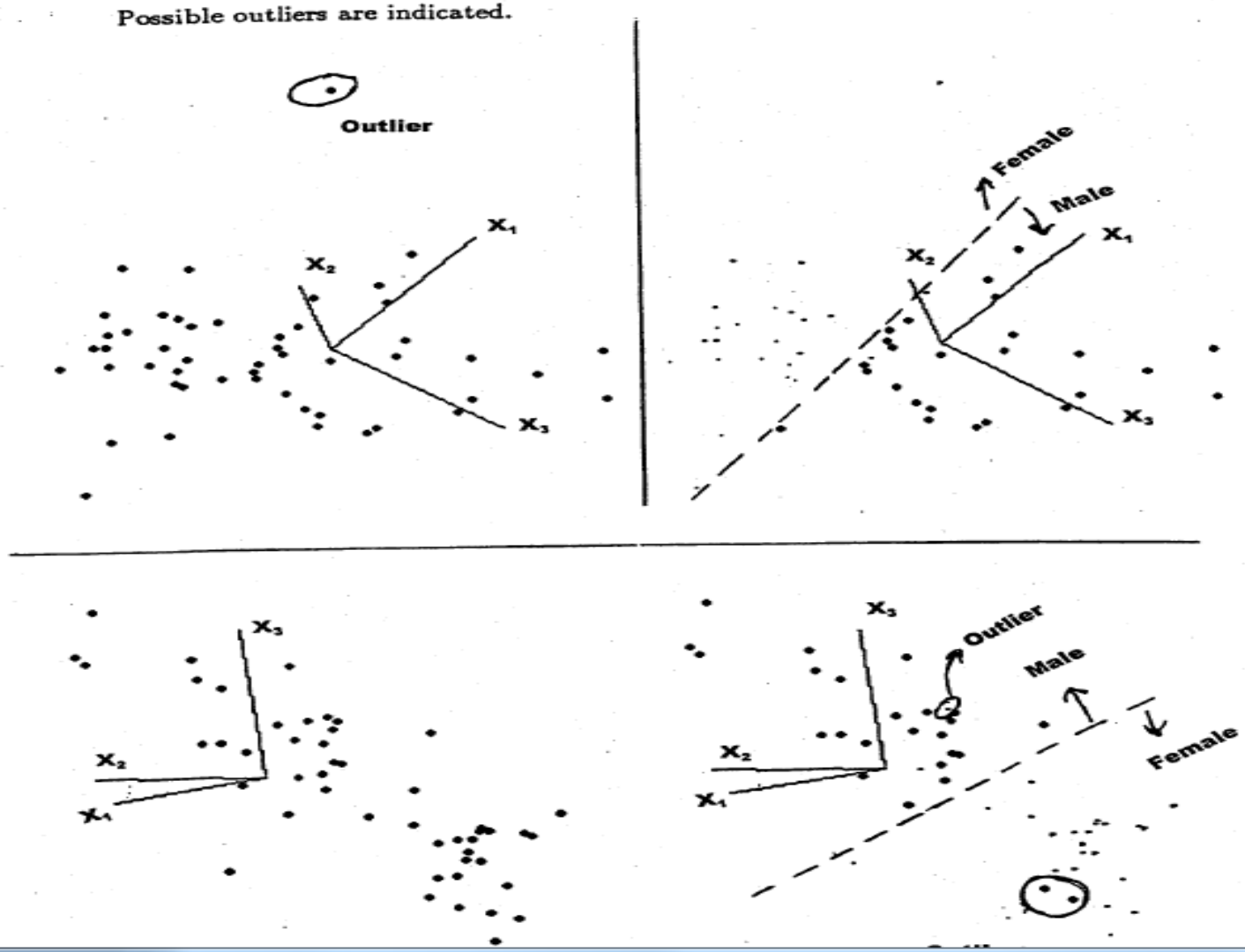
1.22. Refer to the oxygen-consumption data in Table 6.12, page 348, and on the web at www.prenhall.com/statistics. Using appropriate computer software,

(a) View the entire data set in three dimensions employing various combinations of three variables to represent the coordinate axes. Begin with the x_1, x_2, x_3 space.

(b) Check this data set for outliers.

Solution 1.22

1.22 Possible outliers are indicated.



Exercise 1.26

1.26. The data in Table 1.10 (see the bull data on the web at www.prenhall.com/statistics) are the measured characteristics of 76 young (less than two years old) bulls sold at auction. Also included in the table are the selling prices (SalePr) of these bulls. The column headings (variables) are defined as follows:

$$\text{Breed} = \begin{cases} 1 & \text{Angus} \\ 5 & \text{Hereford} \\ 8 & \text{Simental} \end{cases}$$

YrHgt = Yearling height at shoulder (inches)

FtFrBody = Fat free body (pounds)

PrctFFB = Percent fat-free body

Frame = Scale from 1 (small) to 8 (large)

BkFat = Back fat (inches)

SaleHt = Sale height at shoulder (inches)

SaleWt = Sale weight (pounds)

(a) Compute the $\bar{\mathbf{x}}$, $\mathbf{S}_n$, and $\mathbf{R}$ arrays. Interpret the pairwise correlations. Do some of these variables appear to distinguish one breed from another?

(b) View the data in three dimensions using the variables Breed, Frame, and BkFat. Rotate the coordinate axes in various directions. Check for outliers. Are the breeds well separated in this coordinate system?

(c) Repeat part b using Breed, FtFrBody, and SaleHt. Which three-dimensional display appears to result in the best separation of the three breeds of bulls?

Exercise 1.26

Table 1.10 Data on Bulls

Breed	SalePr	YrHgt	FtFrBody	PrctFFB	Frame	BkFat	SaleHt	SaleWt
1	2200	51.0	1128	70.9	7	.25	54.8	1720
1	2250	51.9	1108	72.1	7	.25	55.3	1575
1	1625	49.9	1011	71.6	6	.15	53.1	1410
1	4600	53.1	993	68.9	8	.35	56.4	1595
1	2150	51.2	996	68.6	7	.25	55.0	1488
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
8	1450	51.4	997	73.4	7	.10	55.2	1454
8	1200	49.8	991	70.8	6	.15	54.6	1475
8	1425	50.0	928	70.8	6	.10	53.9	1375
8	1250	50.1	990	71.0	6	.10	54.9	1564
8	1500	51.7	992	70.6	7	.15	55.1	1458

Source: Data courtesy of Mark Ellersieck.

Solution 1.26

1.26 Bull data

(a) XBAR

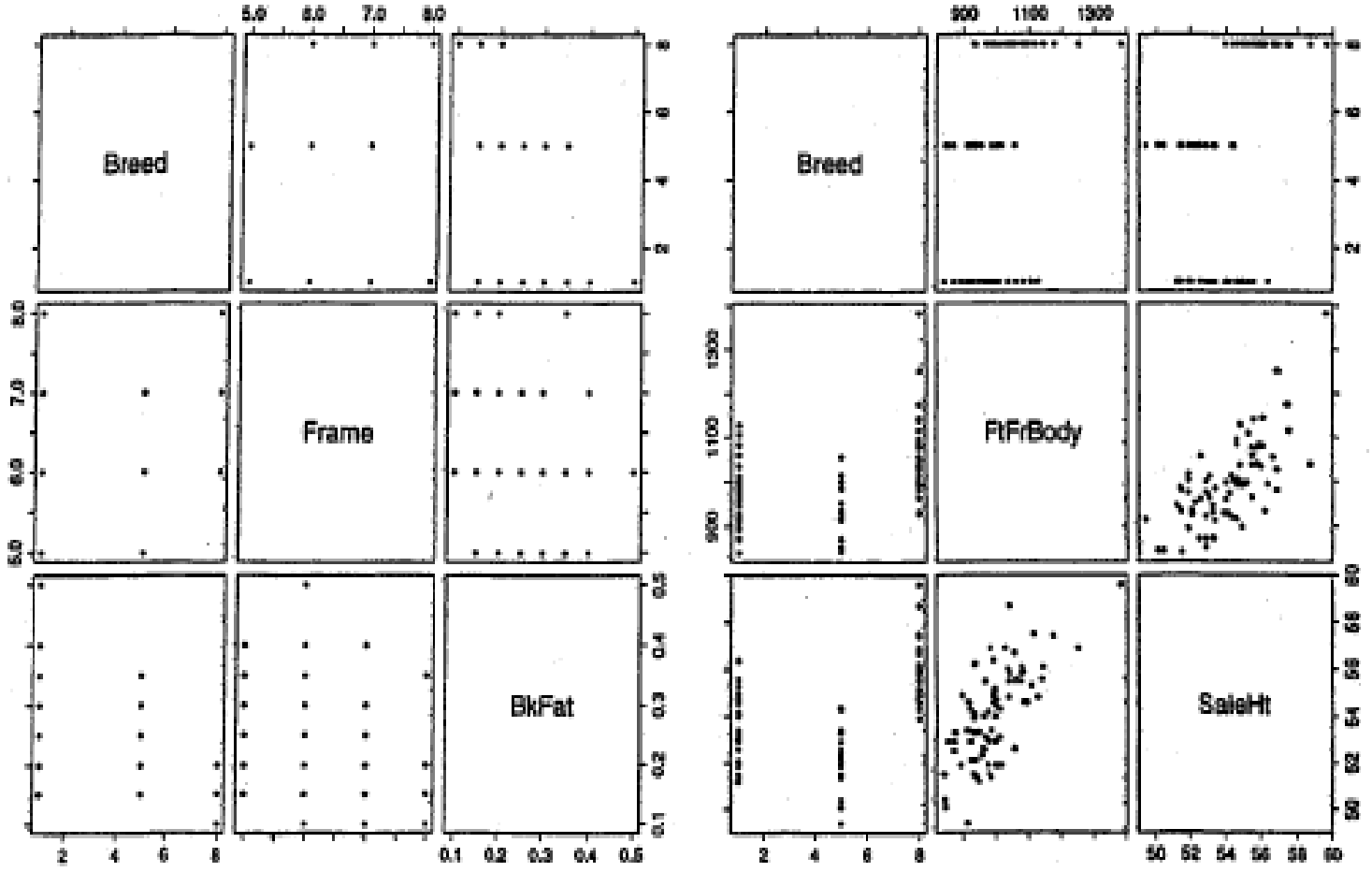
R

	Breed	SalePr	YrHgt	FtFrBody	PrctFFB	Frame	BkFat	SaleHt	SaleWt
4.3816	1.000	-0.224	0.525	0.409	0.472	0.434	-0.615	0.487	0.116
1742.4342	-0.224	1.000	0.423	0.102	-0.113	0.479	0.277	0.390	0.317
50.5224	0.525	0.423	1.000	0.624	0.523	0.940	-0.344	0.860	0.368
995.9474	0.409	0.102	0.624	1.000	0.691	0.605	-0.168	0.699	0.555
70.8816	0.472	-0.113	0.523	0.691	1.000	0.482	-0.488	0.521	0.198
6.3158	0.434	0.479	0.940	0.605	0.482	1.000	-0.260	0.801	0.368
0.1967	-0.615	0.277	-0.344	-0.168	-0.488	-0.260	1.000	-0.282	0.208
54.1263	0.487	0.390	0.860	0.699	0.521	0.801	-0.282	1.000	0.566
1555.2895	0.116	0.317	0.368	0.555	0.198	0.368	0.208	0.566	1.000

Sn

Breed	SalePr	YrHgt	FtFrBody	PrctFFB	Frame	BkFat	SaleHt	SaleWt
9.55	-429.02	2.79	116.28	4.73	1.23	-0.17	3.00	46.32
-429.02	383026.64	450.47	5813.09	-226.46	272.78	15.24	480.56	25308.44
2.79	450.47	2.96	98.81	2.92	1.49	-0.05	2.94	81.72
116.28	5813.09	98.81	8481.26	206.75	51.27	-1.38	128.23	6592.41
4.73	-226.46	2.92	206.75	10.55	1.44	-0.14	3.37	82.82
1.23	272.78	1.49	51.27	1.44	0.85	-0.02	1.47	43.74
-0.17	15.24	-0.05	-1.38	-0.14	-0.02	0.01	-0.05	2.38
3.00	480.56	2.94	128.23	3.37	1.47	-0.05	3.97	145.35
46.32	25308.44	81.72	6592.41	82.82	43.74	2.38	145.35	16628.94

Solution 1.26



Exercise 1.27

1.27. Table 1.11 presents the 2005 attendance (millions) at the fifteen most visited national parks and their size (acres).

(a) Create a scatter plot and calculate the correlation coefficient.

Exercise 1.27

(b) Identify the park that is unusual. Drop this point and recalculate the correlation coefficient. Comment on the effect of this one point on correlation.

(c) Would the correlation in Part b change if you measure size in square miles instead of acres? Explain.

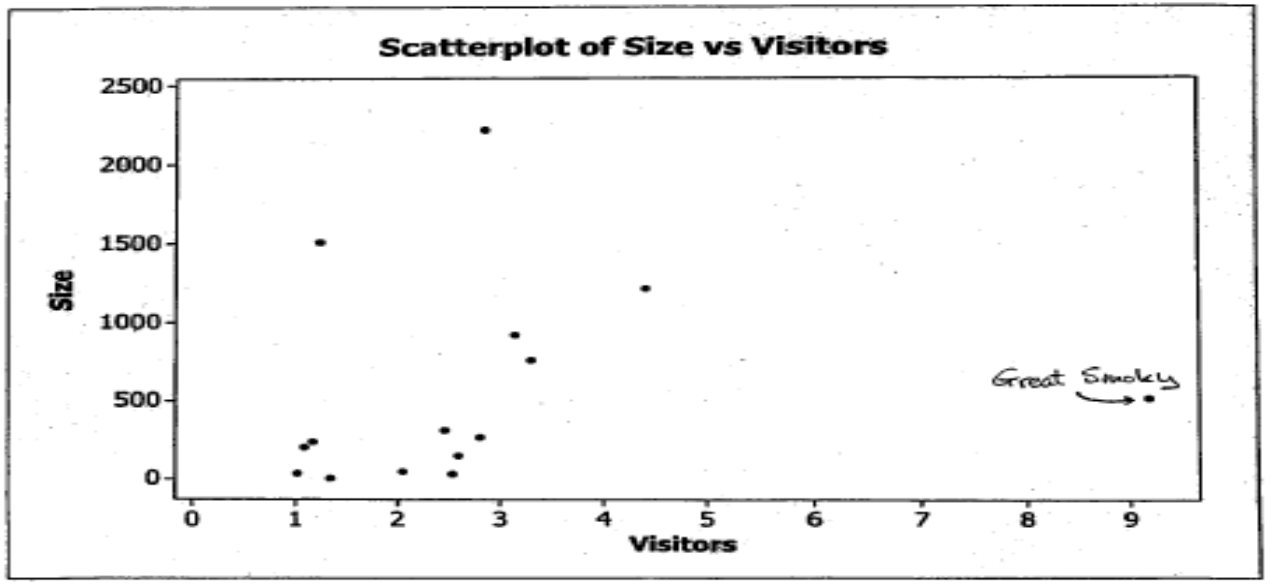
Table 1.11 Attendance and Size of National Parks

National Park	Size (acres)	Visitors (millions)
Arcadia	47.4	2.05
Bruce Canyon	35.8	1.02
Cuyahoga Valley	32.9	2.53
Everglades	1508.5	1.23
Grand Canyon	1217.4	4.40
Grand Teton	310.0	2.46
Great Smoky	521.8	9.19
Hot Springs	5.6	1.34
Olympic	922.7	3.14
Mount Rainier	235.6	1.17
Rocky Mountain	265.8	2.80
Shenandoah	199.0	1.09
Yellowstone	2219.8	2.84
Yosemite	761.3	3.30
Zion	146.6	2.59

Solution 1.27

1.27

(a) Correlation $r = .173$



(b) Great Smoky is unusual park. Correlation with this park removed is $r = .391$. This single point has reasonably large effect on correlation reducing the positive correlation by more than half when added to the national park data set.

(c) The correlation coefficient is a dimensionless measure of association. The correlation in (b) would not change if size were measured in square miles instead of acres.