Chapter 1
Analyzing data

1.1 INTRODUCTION

Imagine being asked to analyze the results of a market survey in which 500 individuals respond to about 40 questions. The questions deal with personal characteristics of the respondents (age, sex, marital status, level of education, income, etc.), the respondents' expenditures on selected products and services, and the number, type, and frequency of purchase of various newspapers and magazines. The survey results can be visualized as arranged in the form of a table, the rows of which correspond to respondents and the columns to the questions. Though such a survey is by no means large, as market surveys go, clearly it would be very difficult to draw any meaningful conclusions simply by glancing at the long list of figures.

However, by suitable reduction—very much as common sense and the techniques that follow describe—it is possible to make comparisons, measure relationships, display graphics, identify trends, and so on. It may be possible, for example, to form a "profile" of the "typical" reader of a sports magazine, to find out if there is any relationship between age and readership of newspapers, to measure the relationship between income and travel expenditures, or to determine trends in magazine readership. This analysis could be of use not only to publishers, but also to advertisers and firms having to decide in which publications a certain product will be advertised.

Techniques and methods by which large sets of data can be summarized and analyzed form the subject of this chapter.

1.2 DISTRIBUTIONS

Almost every type of analysis begins with a set of observations on one or more variables or attributes. This set of observations forms the raw material for statistical analysis. When the number of observations is large, it may be difficult to consider them in the raw form in which they were obtained—hence the need for various measures that reduce the data in a meaningful way.

A first step in the reduction process is usually the classification of the observations into a number of classes (categories, intervals), together with a count of the number ("frequency") of observations falling into each class. It will be convenient always to construct these classes so that they form a mutually exclusive and collectively exhaustive set. A set of classes is called *mutually exclusive* if no observation can be classified into more than one class, and *collectively exhaustive* if every observation can be classified into one of the classes.

A list of mutually exclusive and collectively exhaustive classes and of the corresponding frequencies of observations is called a *frequency distribution.* If the frequencies are divided by the total number of observations, a *relative frequency distribution* is obtained. The following examples illustrate the construction of such distributions.

Example 1.1 The supervisor of an assembly department records the day's output (measured by the number of product units assembled) of the twenty workers in the department:

Assembly department, raw data

Worker No.:	Output	Worker No.:	Output	Worker No.:	Output	Worker No.:	Output
1	10	7	9	13	9	19	10
2	8	8	10	14	8	20	10
3	9	9	11	15	10		
4	10	10	10	16	11		
5	10	11	10	17	9		
6	11	12	9	18	10		

Certain observations occur more than once. Little information is lost if the department's output is summarized in the form of a frequency or relative frequency distribution:

Assembly department, distribution of output

Output (1)	Frequency (2)	Relative frequency (3)
8	2	0.10
9	5	0.25
10	10	0.50
11	3	0.15
Total	20	1.00

Columns (1) and (2) show the frequency distribution of output; columns (1) and (3) show the relative frequency distribution of output. For example, 2 of the 20 workers (10% of the number of workers employed in the department) had an output of 8 units each; 5 workers (25%) had an output of 9 units; etc.

Example 1.2 Table 1.1 shows the age distribution of drivers insured with an automobile company. The raw material in this case can be visualized as a long list showing the age of each of about 18,700 drivers. Needless to say, such a list, even if it could be reproduced here, would be useless—who, after all, can grasp the pattern of nearly 19,000 numbers?

Table 1.1
Distribution of insured drivers by age

Age	Frequency	Relative frequency
Less than 20	2,238	0.1197
20 to 25	2,634	0.1409
25 to 30	2,362	0.1264
30 to 35	2,158	0.1155
35 to 40	1,716	0.0918
40 to 45	1,455	0.0779
45 to 50	1,448	0.0775
50 to 55	1,396	0.0747
55 to 60	1,317	0.0705
60 to 65	1,051	0.0562
65 or more	913	0.0489
Total	18,688	1.0000

A few points concerning the construction of Table 1.1 should be noted. A person's age at a given point in time can conceivably be determined to any degree of accuracy—to the nearest year, month, day, hour, minute, or second. Age is a *continuous variable*. In principle, at least, it can take any value within a specified interval. In Table 1.1, each observation is classified into one of a number of age *intervals*. The intervals are mutually exclusive (an observation cannot be classified into more than one interval) and collectively exhaustive (the intervals cover all possible age values). The width of all but the first and last intervals is constant and equal to five years. Obviously, the intervals could have been made unequal (for example, 20 to 35, 35 to 45), or, if equal, their width could have been any number of years. Needless to say, it is for the investigator to choose the format appropriate for a given situation or problem.

Example 1.3 The sex distribution of the same group of insured drivers is shown in Table 1.2.

Table 1.2
Distribution of insured drivers by sex

Sex	Frequency	Relative frequency
Male	10,331	0.5528
Female	8,357	0.4472
Total	18,688	1.0000

In the first two examples, the observations assumed numerical values and the categories or intervals into which they were classified were also in numerical form. In such cases, we speak of distributions of a *variable*; for example, we say that age or output is a variable with a certain frequency or relative frequency distribution. In the third example, however, there is no natural numerical description of the observations. A person is either male or female; sex is an attribute. *Variables*, then, assume numerical values; *attributes* have no natural numerical description. For example, age, temperature, distance, weight are variables. Sex (male, female), marital status (single, married, divorced, other), nationality (German, French, other) are attributes.

One type of *cumulative frequency distribution* shows the number of observations having values less than or equal to the indicated ones. Similarly, one type of *cumulative relative frequency distribution* shows the proportion of observations with values less than or equal to the indicated ones.

Example 1.1 (Continued) Using the distribution of output in the assembly department, we can construct the following cumulative distributions:

Output (1)	Cumulative frequency (2)	Cumulative relative frequency (3)
8	2	0.10
9	7	0.35
10	17	0.85
11	20	1.00

Columns (1) and (2) form the cumulative frequency distribution, and columns (1) and (3) the cumulative relative frequency distribution of output. For example, 7 workers (35% of the total number of workers) have output less than or equal to 9 units; 17 workers (85%) have output less than or equal to 10 units.

Example 1.2 (Continued) The cumulative age distributions of insured drivers are shown in Table 1.3. For example, 9,392 drivers (50.25% of the total) are under 35.

Table 1.3
Cumulative age distribution of insured drivers

Age interval	Cumulative frequency	Cumulative rel. frequency
To 20	2,238	0.1197
20 to 25	4,872	0.2606
25 to 30	7,234	0.3870
30 to 35	9,392	0.5025
35 to 40	11,108	0.5943
40 to 45	12,563	0.6722
45 to 50	14,011	0.7497
50 to 55	15,407	0.8244
55 to 60	16,724	0.8949
60 to 65	17,775	0.9511
65 or more	18,688	1.0000

In a similar manner, we can construct cumulative distributions showing the number (or proportion) of observations with values greater than, greater than or equal to, or less than the indicated ones.

1.3 GRAPHS

The form of a distribution is often better understood with the help of a graph. Many types of charts are used in books, magazines, and newspapers to present distributions graphically.

Perhaps the most frequently used type is the *bar chart*, in which the frequency or relative frequency of a given value or category equals the height of the corresponding bar. Figure 1.1 shows the relative frequency distribution of output of Example 1.1 in the form of a bar chart. The bars in this

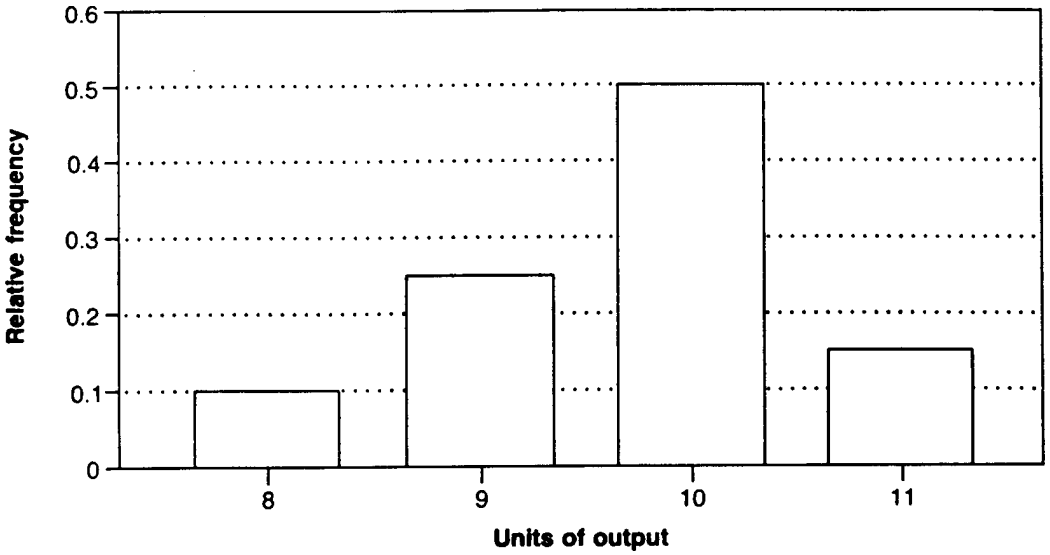

Figure 1.1
Bar chart, Example 1.1

illustration are rather thick, but this is not mandatory; the width of the bar has no special meaning.

Another popular type of graph is the *pie chart*, in which the relative size of the "slice" representing a value or category is made equal to its relative frequency. Figure 1.2 is a pie chart of the distribution of output in Example 1.1.

A lesser-known variant of the bar chart is the *histogram*, most commonly used to display the distribution of a continuous variable. In a histogram, the frequencies or relative frequencies of the variable are shown as rectangles; the width of each rectangle equals the width of the interval, and its *area* the corresponding frequency or relative frequency.

When all intervals have the same width, a histogram looks identical to a bar chart with contiguous bars, except for the scale of the vertical axis. A histogram, however, provides a more sensible display when the intervals have unequal width.

Example 1.4 Columns (1) and (2) of Table 1.4 show the distribution of income for families having more than \$5,000 and less than \$75,000 annual income. (For example, 9.76% of these families have income between \$25,000 and \$30,000.)

All income intervals have the same width. Figure 1.3.a is a bar chart

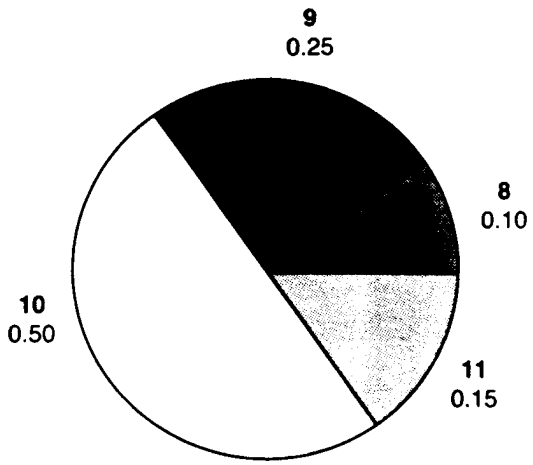

Figure 1.2
Pie chart, Example 1.1

Table 1.4
Distributions of family income

Income ($000) (1)	Rel. frequ. (%) (2)	Income ($000) (3)	Rel. frequ. (%) (4)	Width ($000) (5)	Height (4)÷(5) (6)
5 to 10	4.17	5 to 15	12.73	10	1.27
10 to 15	8.56	15 to 30	28.87	15	1.93
15 to 20	9.76	30 to 75	58.40	45	1.30
20 to 25	9.33		100.00		
25 to 30	9.76				
30 to 35	10.09				
35 to 40	9.76				
40 to 45	9.00				
45 to 50	7.68				
50 to 55	6.69				
55 to 60	5.26				
60 to 65	4.28				
65 to 70	3.07				
70 to 75	2.52				
	100.00				

of this distribution. The bars are contiguous, but this is not necessary—it is the height of the bar that matters.

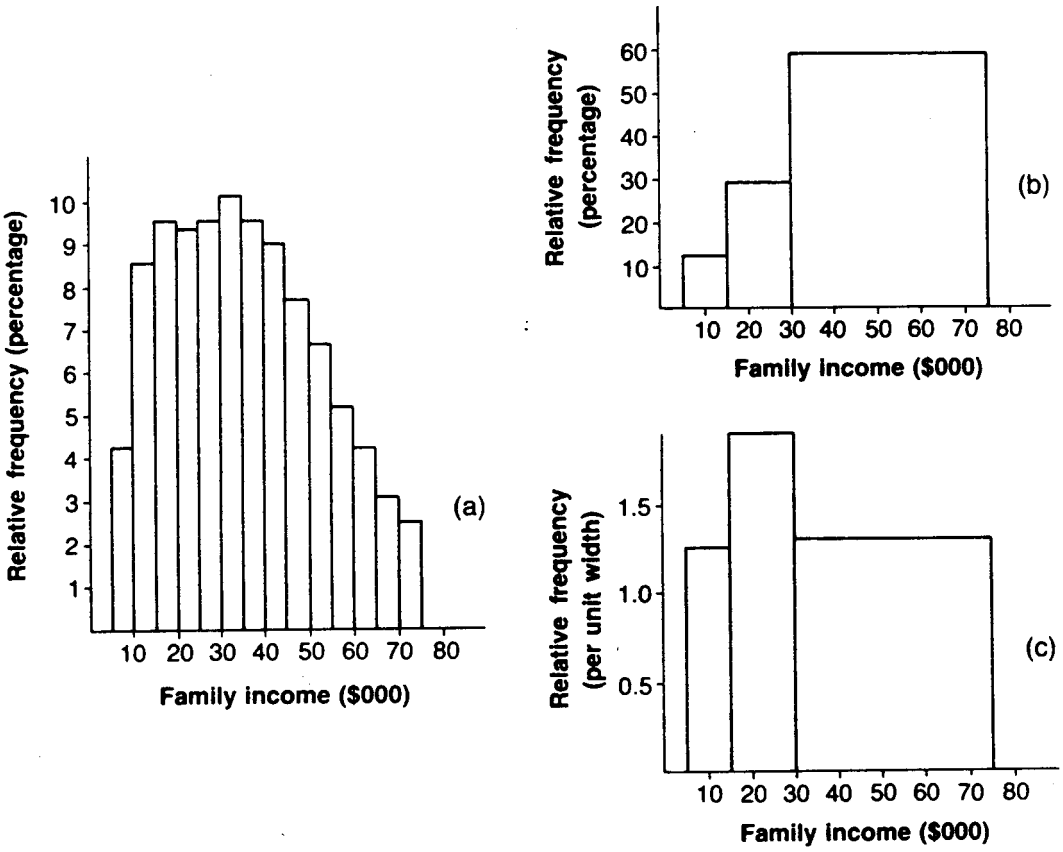

Figure 1.3
Distributions of family income

Suppose, however, that only the abridged distribution of income shown in columns (3) and (4) of Table 1.4 was available. It is consistent with the detailed distribution, but the intervals are fewer and of unequal width. A bar chart of the abridged distribution is shown in Figure 1.3.b. Note that it gives quite a different impression of the form of the income distribution from that given by Figure 1.3.a.

Consider now the histogram shown in Figure 1.3.c. The relative frequency of each income interval is given by the area (width $\times$ height) of the rectangle. The calculation of the heights is shown in column (6) of Table

1.4. The vertical axis of Figure 1.3.c is labelled "relative frequency per unit width," as a reminder that we are dealing with a histogram. In this case, of course, the original distribution has been drastically abridged, but, despite this, it can be seen that the histogram preserves better than the bar chart the form of the detailed distribution.

A histogram has two properties, which we shall use later on: the area of the bar equals the frequency or relative frequency; and the total area of the bars equals 1 (if the histogram depicts relative frequencies) or the total number of observations (if it shows frequencies).

1.4 MEASURES OF TENDENCY

In addition to—or in place of—a frequency distribution, a cumulative distribution, or a graph, it is often desirable to further reduce the information contained in the set of observations to a single number that is, in a certain sense, a representative measure of the entire distribution of the variable. Two types of such measures are usually encountered: (a) measures of location or tendency; and (b) measures of dispersion, indicating the degree of variation or dispersion of the observations around a measure of tendency or location.

Perhaps the most familiar and most widely used measure of location is the arithmetic average (more simply, the *average* or *mean*) of the variable. If there are n observations with values $x_1, x_2, \ldots, x_n$, their average or mean ($\bar{x}$) is defined as

$$\begin{aligned} \bar{x} &= \frac{1}{n}(x_1 + x_2 + \cdots + x_n) \\ &= \frac{1}{n}\sum x. \end{aligned} \tag{1.1}$$

The last expression is a shorthand version of the first; this *summation* notation is explained in Appendix 1.

When the observations have been grouped in the form of a frequency or relative frequency distribution, the calculation of the mean becomes easier. If a discrete variable takes values $x_1, x_2, \ldots, x_m$ with respective frequencies $f(x_1), f(x_2), \ldots, f(x_m)$, the mean, $\bar{x}$, is

$$\begin{aligned} \bar{x} &= \frac{1}{n}[x_1 f(x_1) + x_2 f(x_2) + \cdots + x_m f(x_m)] \\ &= \frac{1}{n}\sum x f(x), \end{aligned} \tag{1.2}$$

where $n = f(x_1) + f(x_2) + \cdots + f(x_m) = \sum f(x)$.

It should be clear that Equation (1.2) follows from (1.1), since there are $f(x_1)$ observations having the value x_1, $f(x_2)$ observations having the

value x_2, and so on. Equation (1.2) can also be written in terms of relative frequencies:

$$\begin{aligned}\bar{x} &= x_1\frac{f(x_1)}{n} + x_2\frac{f(x_2)}{n} + \cdots + x_m\frac{f(x_m)}{n} \\ &= x_1 r(x_1) + x_2 r(x_2) + \cdots + x_m r(x_m) \\ &= \sum xr(x),\end{aligned} \tag{1.3}$$

where $r(x) = f(x)/n$ are the relative frequencies of the variable.

Example 1.1 (Continued) We illustrate the two methods for calculating the mean using the distribution of output of the assembly department.

Calculation of mean output

x	$f(x)$	$r(x)$	$xf(x)$	$xr(x)$
8	2	0.10	16	0.80
9	5	0.25	45	2.25
10	10	0.50	100	5.00
11	3	0.15	33	1.65
	20	1.00	194	9.70

The average output per worker can be calculated using frequencies,

$$\bar{x} = \frac{1}{n}\sum xf(x) = \frac{194}{20} = 9.70,$$

or relative frequencies,

$$\bar{x} = \sum xr(x) = 9.70.$$

The result is the same.

When the observations have been classified into intervals and the raw data are not available, the exact value of the mean cannot be calculated. An approximate value can be obtained by treating all observations in the interval as being equal to the midpoint of the interval; the mean is then calculated using either Equation (1.2) or Equation (1.3) with x_1, x_2, ... , x_m now denoting the midpoints of the class intervals. It can be shown that the accuracy of this approximation depends on how close the midpoints are to the averages of the observations in each interval.

Table 1.5
Calculation of mean of age distribution of insured drivers

Age interval	Midpoint, x	Frequency, $f(x)$	Rel. frequ., $r(x)$	$xf(x)$	$xr(x)$
To 20	18.0	2,238	0.1197	40,284.0	2.15460
20 to 25	22.5	2,634	0.1409	59,265.0	3.17025
25 to 30	27.5	2,362	0.1264	64,955.0	3.47600
30 to 35	32.5	2,158	0.1155	70,135.0	3.75375
35 to 40	37.5	1,716	0.0918	64,350.0	3.44250
40 to 45	42.5	1,455	0.0779	61,837.5	3.31075
45 to 50	47.5	1,448	0.0775	68,780.0	3.68125
50 to 55	52.5	1,396	0.0747	73,290.0	3.92175
55 to 60	57.5	1,317	0.0705	75,727.5	4.05375
60 to 65	62.5	1,051	0.0562	65,687.5	3.51250
65+	68.0	913	0.0489	62,084.0	3.32520
	Total	18,688	1.0000	706,395.5	37.80230

Example 1.2 (Continued) The approach is illustrated in calculating the average age of insured drivers, as shown in Table 1.5.

The first and last intervals are open: 18 and 68 were arbitrarily chosen as the midpoints of the intervals in the belief that they would be close to the average age of drivers under 20 and over 65.

The approximate average age can be calculated using either the frequency distribution,

$$\bar{x} = \frac{1}{n}\sum xf(x) = (706,395.5)/(18,688) = 37.799,$$

or the relative frequency distribution,

$$\bar{x} = \sum xr(x) = 37.802.$$

The small discrepancy is due to rounding. We conclude that the average age of insured drivers is approximately 37.8 years.

Measures of tendency other than the mean are less frequently used. The *mode* of a distribution is the most frequently occurring observation or interval. For example, the mode of the distribution of output in the assembly department is 10; similarly, the modal five-year age interval of the

drivers' age distribution is 20 to 24. Obviously, a distribution will not have a unique mode if more than one value or interval are tied as most frequently occurring.

The *median* of a distribution is the value of the variable which divides the distribution into two equal halves. Think of the observations as arranged in increasing or decreasing order: the median is the value of the observation in the middle of this list. For example, the median of the distribution of output in the assembly department is 10; the median age of insured drivers is in the interval 30 to 35. The values of the variable that divide the distribution (arranged in order of increasing values of the observations) into four equal parts are called, respectively, the *first quartile*, the *second quartile* (which is the median), and the *third quartile* of the distribution. The values of the variable which divide the distribution into five equal parts are called *quintiles*; those dividing the distribution into ten equal parts are called *deciles*; and so on.

1.5 MEASURES OF DISPERSION

Measures of dispersion variously attempt to describe the "scatter," "variation," or "spread" of the observations around a central value. Almost invariably, the mean serves as that central value, but any other measure of tendency, such as the median or the mode of the distribution, can in principle be used.

Suppose that n (ungrouped) observations $x_1, x_2, \ldots, x_n$ are available. An obvious candidate for a measure of dispersion is the *average deviation*:

$$\frac{1}{n}\sum(x - \bar{x}). \tag{1.4}$$

$(x - \bar{x})$ measures the difference between an observation and the mean, and the average deviation is the average of these differences. It is easy to show, however, that the average deviation is *always* equal to zero (see next example), and is therefore useless as a measure of dispersion. An alternative measure of dispersion is the *average absolute deviation*, defined as

$$\frac{1}{n}\sum|x - \bar{x}|. \tag{1.5}$$

Despite its intuitive appeal, this measure of dispersion is seldom used, primarily because it is not tractable mathematically. The most widely used measure of dispersion, one very similar to the average absolute deviation, is the *variance*, s^2:

$$s^2 = \frac{1}{n}\sum(x - \bar{x})^2. \tag{1.6}$$

The (positive) square root of the variance is the *standard deviation* or *standard error*, s:

$$s = +\sqrt{s^2}. \tag{1.7}$$

For example, suppose that the starting annual salaries of $n = 3$ accountants are 33, 34, and 38 ($000). The three measures of dispersion defined above can be calculated as follows:

x	$(x - \bar{x})$	$\lvert x - \bar{x} \rvert$	$(x - \bar{x})^2$
33	-2	2	4
34	-1	1	1
38	3	3	9
105	0	6	14
$(\bar{x} = 35)$	$\sum(x - \bar{x})$	$\sum \lvert x - \bar{x} \rvert$	$\sum(x - \bar{x})^2$

The average starting salary is 35. The sum of deviations from the mean, $\sum(x - \bar{x})$, equals zero, as is always the case. The average absolute deviation is 6/3 or 2. The variance equals 14/3 or 4.667. The standard deviation is $\sqrt{4.667}$ or 2.16.

Measures of dispersion are used primarily to compare the variation of two or more sets of observations. From now on, we shall use either the variance or the standard deviation as such a measure.

To illustrate, suppose we are attempting to compare starting salaries of accountants and economists. The starting salaries of four economists were 25, 29, 32, 40 ($000) and are shown in Figure 1.4 together with the salaries of the three accountants.

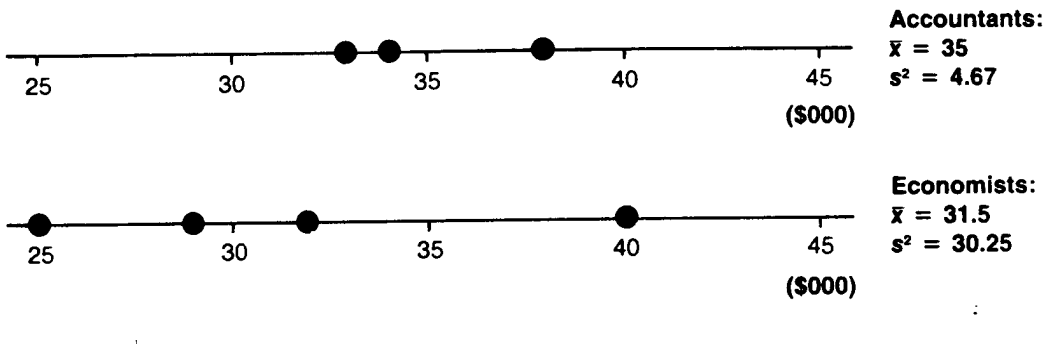

Figure 1.4
Comparison of starting salaries

It is clear from a visual inspection of Figure 1.4 that accountants, on the average, earn more than economists; also, that their salaries tend to vary less around their average salary than do the salaries of economists. We

would have reached exactly the same conclusions had we compared instead the means (35 *vs.* 31.5) and variances (4.667 *vs.* 30.250) of the two sets of observations.

Of course, in this small example, visual inspection is convenient and reliable; however, the value of the summary measure will be appreciated in cases involving many observations, where the pattern of the observations cannot be ascertained easily even with the help of a graph.

When the observations have been grouped in the form of a frequency or relative frequency distribution, the calculation of the variance can be simplified. If a discrete variable takes values $x_1, x_2, \ldots, x_m$, with frequencies $f(x_1), f(x_2), \ldots, f(x_m)$, then the variance of the distribution can be calculated from the following expression:

$$\begin{aligned} s^2 &= \frac{1}{n}[(x_1-\bar{x})^2 f(x_1) + (x_2-\bar{x})^2 f(x_2) + \cdots + (x_m-\bar{x})^2 f(x_m)] \\ &= \frac{1}{n}\sum (x-\bar{x})^2 f(x). \end{aligned} \tag{1.8}$$

This formula is derived from Equation (1.6). It follows because $f(x_1)$ observations have the value x_1 and deviation $(x_1 - \bar{x})$, $f(x_2)$ observations have deviation $(x_2 - \bar{x})$, and so on. The variance can also be calculated using relative frequencies. Since $r(x) = f(x)/n$,

$$s^2 = \sum (x-\bar{x})^2 \frac{f(x)}{n} = \sum (x-\bar{x})^2 r(x). \tag{1.9}$$

Example 1.1 (Continued) The mean of the distribution of output in the assembly department was found earlier to be 9.70 units. The variance of the distribution can be calculated using the frequencies in column (2) as shown in columns (4) to (6).

x	$f(x)$	$r(x)$	$(x-\bar{x})$	$(x-\bar{x})^2$	$(x-\bar{x})^2 f(x)$	$(x-\bar{x})^2 r(x)$
(1)	(2)	(3)	(4)	(5)	(6)	(7)
8	2	0.10	−1.70	2.89	5.78	0.2890
9	5	0.25	−0.70	0.49	2.45	0.1225
10	10	0.50	0.30	0.09	0.90	0.0450
11	3	0.15	1.30	1.69	5.07	0.2535
	$n = 20$	1.00			14.20	0.7100

Thus, the variance is

$$s^2 = \frac{1}{n}\sum (x-\bar{x})^2 f(x) = \frac{1}{20}14.20 = 0.71.$$

Alternatively, the variance may be calculated using the relative frequencies of column (3), as shown in column (7):

$$s^2 = \sum (x - \bar{x})^2 r(x) = 0.71.$$

The result is the same.

Just as in raw data, the variance measures the dispersion of the observations about their mean. The greater the dispersion, the greater the variance. To illustrate, suppose that the distributions of starting salary of accountants and economists were as depicted in the histograms of Figure 1.5 (the data are fictitious). The same scale is used in both panels.

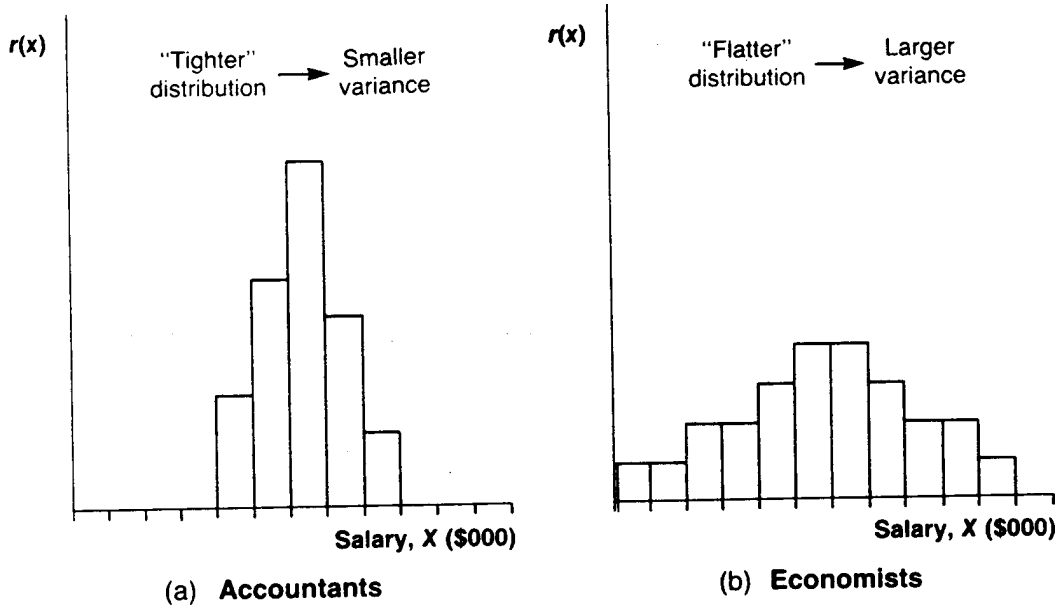

Figure 1.5
Distributions of starting salaries

Accountants' salaries tend to vary less around their average salary than do economists' salaries; the variance of accountants' salaries will therefore be smaller than that of economists' salaries.

The variance of a distribution can also be written in a form more convenient for hand calculations:

$$s^2 = \frac{1}{n} \sum x^2 f(x) - \bar{x}^2. \tag{1.10}$$

In terms of relative frequencies, this is

$$s^2 = \sum x^2 r(x) - \bar{x}^2. \tag{1.11}$$

When the observations are grouped into class intervals, the variance of a distribution can be approximated by using the midpoint as the value of all the observations in the interval and then applying the above expressions in a straightforward way.

Example 1.2 (Continued) Table 1.6 illustrates the use of midpoints and of Equation (1.11) in calculating the variance of the age distribution of insured drivers.

Table 1.6
Calculation of variance of drivers' age distribution

Age interval	Midpoint, x	Rel. frequ., $r(x)$	x^2	$x^2 r(x)$
To 20	18.0	0.1197	324.00	38.783
20 to 25	22.5	0.1409	506.25	71.331
25 to 30	27.5	0.1264	756.25	95.590
30 to 35	32.5	0.1155	1,056.25	121.997
35 to 40	37.5	0.0918	1,406.25	129.094
40 to 45	42.5	0.0779	1,806.25	140.707
45 to 50	47.5	0.0775	2,256.25	174.859
50 to 55	52.5	0.0747	2,756.25	205.892
55 to 60	57.5	0.0705	3,306.25	233.091
60 to 65	62.5	0.0562	3,906.25	219.531
65+	68.0	0.0489	4,624.00	226.114
		1.0000		1,656.989

Applying Equation (1.11), and recalling that $\bar{x} = 37.80$ from Table 1.5, we get:

$$s^2 = \sum x^2 r(x) - \bar{x}^2 = (1,656.989) - (37.80)^2 = 228.15.$$

The standard deviation of the distribution is $s = 15.10$.

It should always be kept in mind that neither the means nor the variances of dissimilar distributions can be compared. Obviously, it does not make sense to compare a distribution of age with a distribution of income, the mean age with mean income, or the variance of the age distribution with

the variance of the income distribution. We may, however, compare similar distributions and their associated summary measures, provided that the comparison makes sense. For example, it would be appropriate to compare the mean and variance of the distribution of men's income with those of women's income; or the current distribution of age with that of ten years ago.

1.6 INDEX NUMBERS

Index numbers are measures designed to indicate the relative changes in the overall level of such variables as prices, production, wages, or employment.

The simplest type of *index* expresses a series of measurements on one particular variable in terms of one member of the series. For example, Table 1.7 shows the price of a "standard" refrigerator during the period 19X7 to 19X9.

Table 1.7
Price and price index of refrigerators, 19X7 to 19X9

Year	Price ($)	Price index, 19X7 = 100
19X7	600	100
19X8	700	117
19X9	750	125

This price series can be expressed in terms of the 19X7 price by dividing the price each year by the 19X7 price, and multiplying the result by 100 to form an index of the price of refrigerators with the year 19X7 serving as the *base* of the index. In relation to the 19X7 price, then, the price of refrigerators in 19X8 was $(700/600) \times 100$ or about 117%, while that in 19X9 was about 125%. Put a different way, the price in 19X8 showed a 17% increase relative to the 19X7 price; the 19X9 price was 25% greater than the 19X7 price.

The most usual type of index is the *composite index*, which summarizes a number of series into one index series. Let us consider, as an example, how one may construct an index of *appliance* prices based on the information shown in Table 1.8.

Perhaps the simplest way of combining these prices to form a composite index is to calculate each year the ordinary average price of the appliances, and to express the terms of this series in relation to the average price in the base year, as shown in Table 1.9.

The obvious shortcoming of this method is that all the price series are given the same weight (namely, 1/3) in the calculation of the average price

Table 1.8
Appliance prices, 19X7 to 19X9

Year	Price of: Stoves	Refrigerators	Freezers
19X7	$475	$600	$350
19X8	$490	$700	$400
19X9	$510	$750	$440

Table 1.9
Index of appliance prices, 19X7 to 19X9, based on average of prices

Year	Average price of appliances ($)	Index of appliance prices, 19X7=100
19X7	475	100
19X8	530	112
19X9	567	119

and hence in the index. Since the appliances sell in different numbers, it would seem that a better index could be constructed by assigning different weights to the individual price series—weights reflecting the relative contributions of the various appliance types in the market.

Suppose that the numbers of appliances sold in 19X7 were as follows:

Appliance	Number sold (000)	Share
Stoves	600	0.4
Refrigerators	750	0.5
Freezers	150	0.1
Total	1,500	1.0

A weighted average appliance price can be constructed by multiplying the price of each type by the market share of that appliance type and summing the products. For example, the weighted average appliance price in 19X7 would be

$$(0.4)(475) + (0.5)(600) + (0.1)(350) = 525.$$

An index can then be constructed by dividing the weighted average price each year by that in the base year and multiplying the result by 100. These calculations are shown in Table 1.10.

According to this composite index, 19X8 appliance prices increased by 12% and 19X9 prices by 19% over appliance prices in 19X7.

Table 1.10
Index of appliance prices, 19X7 to 19X9,
based on weighted average of prices

Year	Weighted ave. price ($)	Index, 19X7=100
19X7	525	100
19X8	586	112
19X9	623	119

This simple example illustrates a method that can be used to form *any* composite index. Depending on the case and the purpose for which the index is constructed, a decision must be made regarding the appropriate number and type of the time series to be incorporated in the index, the period to serve as the base of the index, and the weights to be used in combining the individual time series. The weights in particular are usually chosen so as to reflect the importance of the individual series in forming the composite index. The base period need not be the same as that of the weights. The weights, however, must remain constant if the purpose of the index is to show changes in the component series.

For example, the *Consumer Price Index (CPI)* in the United States and Canada is a composite index of prices of commodities and services, with weights reflecting the shares of these items in consumer expenditures. The *Dow-Jones Industrial Average (DJIA)* is a composite index of the prices of 30 selected companies ("representative of Corporate America") with equal weights. The *New York Stock Exchange (NYSE) Index* is a composite index of the prices of *all* stocks listed on the Exchange, weighted according to the total market value of outstanding shares. Other stock exchange indexes—e.g., the *Toronto Stock Exchange (TSE) Index*—are calculated in a similar fashion on the basis of the prices of a *sample* of stocks traded.

1.7 BIVARIATE DISTRIBUTIONS

Until now, we have assumed that the observations are classified into categories or classes according to a single variable or attribute. The frequency and relative frequency distributions we have examined may thus be called *univariate* frequency or relative frequency distributions. There are cases, however, where the observations may be classified according to two, three, or more variables or attributes jointly, yielding *bivariate*, *trivariate*, or in general *multivariate frequency* and *relative frequency distributions*—alternatively, yielding *joint distributions* of two, three, or more specified variables or attributes. (The term *cross-tabulation* is also used to indicate a bivariate distribution.)

Example 1.5 The population of insured drivers, classified according to a single variable—age—yielded the distribution shown in Table 1.1. The same population, classified according to another single attribute—sex—yielded the distribution shown in Table 1.2. Still the same population, classified jointly according to sex and age, yields the joint distribution shown in Table 1.11.

Table 1.11
Age and sex distribution of insured drivers

	Relative frequencies (%)		
Age interval	Male	Female	Totals
Less than 20	7.68	4.29	11.97
20 to 25	8.29	5.80	14.09
25 to 30	7.29	5.35	12.64
30 to 35	6.61	4.94	11.55
35 to 40	5.06	4.12	9.18
40 to 45	4.14	3.65	7.79
45 to 50	4.12	3.63	7.75
50 to 55	3.87	3.60	4.74
55 to 60	3.53	3.52	7.05
60 to 65	2.59	3.03	5.62
65 or more	2.10	2.79	4.89
Totals	55.28	44.72	100.00

Table 1.11 shows the relative frequencies with which the indicated combinations of age and sex occur. For example, of the entire population, 7.68% were male under 20 years old; 4.29% were female under 20; 8.29% were male 20 to 25 years old; and so on. Note that the sum of the relative frequencies for each age interval or sex category shown on the margin of the table is identical to the relative frequency for that age interval or sex category given by the corresponding univariate relative frequency distributions of age and sex. For example, since 7.68% of drivers are male under 20 years old and 4.29% are female under 20 years old, the percentage of drivers under 20 years old (regardless of their sex) ought to be equal to $7.68 + 4.29 = 11.97$, a figure which can be obtained directly from the univariate distribution of age shown in Table 1.1. Similarly, since 7.68% of drivers were male less than 20 years old, 8.29% were male 20 to 25 years old, ... , and 2.10% were male 65 or more years old, the percentage of male drivers (regardless of age) is the sum of these relative frequencies (55.28%), which is the figure given in the univariate distribution of sex shown in Table 1.2. This identity always holds

when the classes or class intervals are mutually exclusive and collectively exhaustive.

The mean, variance, and other summary measures of a variable may be calculated in a straightforward way using the values of the variable and the corresponding frequencies or relative frequencies shown on the margin of a joint distribution.

A bivariate frequency or relative frequency distribution of a variable can be shown graphically in the form of a *stereogram*, which is an extension of the histogram to the bivariate case (see Figure 1.6). The frequency or relative frequency of a given pair of intervals is indicated by a pillar, the volume of which is equal to the frequency or relative frequency, while its width and depth are equal to the length of the corresponding intervals. Such a diagram is, of course, difficult to construct, but it can be borne in mind as an aid in visualizing bivariate distributions.

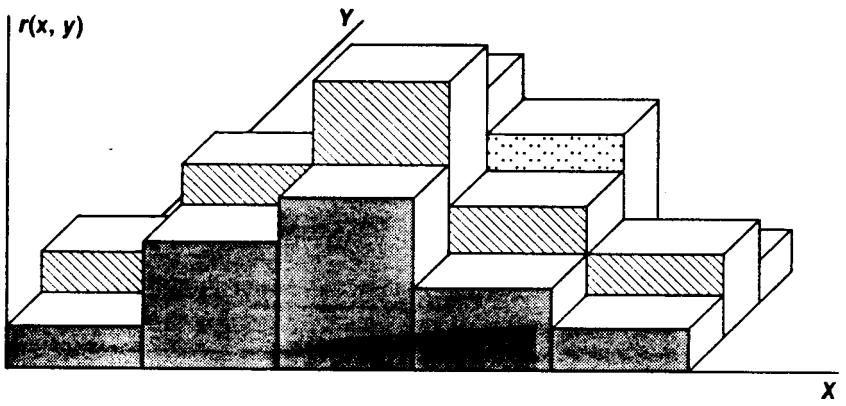

Figure 1.6
Bivariate distribution

1.8 CONDITIONAL DISTRIBUTIONS

The frequency or relative frequency distribution of one variable or attribute given that the other variable or attribute has a certain value or falls into a certain category is called a *conditional distribution.* There are as many conditional distributions of one variable or attribute as there are values or classes of the other.

Example 1.6 The joint frequency distribution of drivers according to age and accident involvement during a recent one-year period was as follows (all figures in thousands):

	Involved in accident?		
Age	Yes	No	Total
Under 25	127	805	932
25 and over	239	3,392	3,631
Total	366	4,197	4,563

Thus, of the 4,563 registered drivers, 127 were 24 years old or younger and were involved in one or more traffic accidents; 3,392 were 25 years old or older and were not involved in any traffic accidents; and so on.

The distribution of accident involvement for drivers under 25 can be determined as follows. Of the 4,563 drivers, 932 were under 25; of these, 127 or about 14% had an accident, while 805 or 86% had no accidents. Therefore, the *conditional relative frequency distribution* of accident involvement for drivers under 25 is:

Drivers under 25	
Involved in accident?	Rel. frequency
Yes	0.14
No	0.86
	1.00

Similarly, the conditional relative frequency distribution of accident involvement for drivers 25 and over is:

Drivers 25 and over		
Involved in accident?	Rel. frequency	
Yes	(239/3,631=)	0.07
No	(3,392/3,631=)	0.93
		1.00

These distributions confirm a rather well-known fact, namely, that the accident rate among young drivers is greater than among older drivers, although the difference in these rates (14% *vs.* 7%) is not as great as many people believe.

The same joint frequency distribution can be used to construct two other conditional distributions. The conditional distribution of age for drivers involved in an accident is:

Drivers with accidents		
Age	Rel. frequency	
Under 25	(127/366=)	0.35
25 and over	(239/366=)	0.65
		1.00

Thus, of the 366 drivers involved in an accident, 127 or 35% were under 25, while 239 or 65% were 25 and over.

Similarly, the conditional distribution of age for drivers who had no accidents is:

Drivers without accidents	
Age	Rel. frequency
Under 25	0.19
25 and over	0.81
	1.00

It may be noted that 932 or 20.4% of the 4,563 drivers were under 25; but among the drivers who had an accident, younger drivers accounted for a much higher proportion (34.7%).

Conditional relative frequency distributions, therefore, are calculated by dividing the joint frequencies by the appropriate row or column totals. Exactly the same method is used to calculate conditional distributions on the basis of a joint relative frequency distribution, as the following illustrates.

Example 1.6 (Continued) Suppose that only the joint relative frequency distribution of drivers by age and accident involvement was available:

	Involved in accident?		
Age	Yes	No	Total
Under 25	0.028	0.176	0.204
25 and over	0.052	0.744	0.796
Total	0.080	0.920	1.000

This distribution, of course, is based on the joint frequency distribution listed earlier; for example, 127 or 2.8% of the 4,563 drivers were under 25 and were involved in an accident.

To construct the conditional distribution of, say, age, given that a driver was involved in an accident, we could argue as follows. Out of every 1,000 drivers, 80 were involved in an accident; out of these 80, 28 were under

25 and 52 were 25 years old or older. Therefore, the proportion of "under 25s" among drivers who had an accident is 28/80 or 35%. Similarly, the proportion of drivers 25 and over among those drivers who had an accident is 65%.

Drivers with accidents

Age	Rel. frequency	
Under 25	(0.028/0.080=)	0.35
25 and over	(0.052/0.080=)	0.65
		1.00

This is exactly the same conditional distribution as the one calculated using the joint frequency distribution.

The reader can verify that all the other conditional distributions can also be obtained by dividing the joint relative frequencies in the appropriate row and column by that row or column's marginal relative frequency. The "appropriate" row or column is the one that corresponds to the "given" category or value.

Let us write this definition with symbols (doing so may appear pedantic at this point, but will help us better understand similar definitions for probability distributions in the next chapter). Imagine the joint relative frequency distribution of two variables or attributes, X and Y, arranged in the form of a table:

		Y		
X	$\cdots$	y	$\cdots$	Total
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
x	$\cdots$	$r(x,y)$	$\cdots$	$r(x)$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
Total	$\cdots$	$r(y)$	$\cdots$	1.0

The relative frequency of the pair of classes (x, y) is denoted by $r(x, y)$; $r(x)$ and $r(y)$ are the marginal relative frequencies of classes x and y respectively. The conditional relative frequency of class y given that the class of attribute X is x, which we write as $r(Y = y|X = x)$ and abbreviate as $r(y|x)$, is given by

$$r(y|x) = \frac{r(x,y)}{r(x)}. \tag{1.12}$$

Similarly,

$$r(x|y) = \frac{r(x,y)}{r(y)}. \tag{1.13}$$

The above imply that

$$r(x, y) = r(x)r(y|x) = r(y)r(x|y), \tag{1.14}$$

an expression useful for calculating joint from marginal and conditional relative frequencies.

1.9 INDEPENDENCE

In business, as in other fields, we often want to determine whether or not two variables or attributes are related to one another, and, if so, to measure the strength of that relationship. This is particularly true when we wish to use one variable or attribute to predict the other. Think, for example, of the relationship between a test for admission to a graduate program and performance in that program. The test is taken before admission. If the test is well designed, we expect that it will be related to performance in the program. In particular, we expect people who do well in the test to do well in the program also. Conversely, people who do not do well in the test may be denied admission to the program on the grounds that they are expected to do poorly if admitted to the program.

Let us begin with a very simple example, and first consider how to establish whether or not there is *any* relationship between two attributes. Once we determine that a relationship exists, we can then consider how to measure the strength of that relationship.

Suppose that it is possible to classify unambiguously drivers as Good or Bad. (This is hardly a simple task, but we need not elaborate further.) Suppose further that 200 drivers were classified according to sex and skill as shown in Table 1.12.

Table 1.12
Distribution of drivers by skill and sex

	Sex		
Skill	Male	Female	Total
Good	77 (0.385)	33 (0.165)	110 (0.550)
Bad	63 (0.315)	27 (0.135)	90 (0.450)
Total	140 (0.700)	60 (0.300)	200 (1.000)

The joint relative frequencies are shown in parentheses. The question is: Is there any relationship between sex and driving skill?

It would be inappropriate, of course, to claim that male drivers are worse than female drivers because "there are more male bad drivers (63) than female bad drivers (27)"; there are, after all, more male drivers in

Table 1.13
Conditional relative frequency distributions

Skill	Male drivers		Female drivers		Marginal distribution, all drivers	
Good	(77/140=)	0.55	(33/60=)	0.55	(110/200=)	0.55
Bad	(63/140=)	0.45	(27/60=)	0.45	(90/200=)	0.45
Total		1.00		1.00		1.00

total than female drivers. A more appropriate comparison should be based on the conditional distributions of skill given sex shown in Table 1.13.

We see that 77 of the 140, or 55% of the male drivers are Good, and 45% are Bad; 55% of the female drivers are Good and 45% are Bad. In this case, the conditional distribution of driving skill is the same regardless of the sex of the driver. To put it roughly, as far as skill is concerned, it does not make any difference whether the driver is a man or a woman. We are justified therefore in claiming that the two attributes, sex and driving skill, are unrelated to one another. (We would have reached the same conclusion had we considered instead the conditional distributions of sex given skill, as the reader can easily verify.)

Two attributes, then, are said to be *unrelated* to or *independent* of one another if all the conditional distributions of one attribute given the other are identical.

Unrelated attributes have two features which are also illustrated in this example. *First*, when two attributes are unrelated, all the conditional relative frequency distributions of one attribute are the same and are equal to the marginal relative frequency distribution of the attribute. Note that the proportions of Good and Bad drivers among *all* drivers are 55% and 45% respectively—the same as among male or female drivers. *Second*, when two attributes are unrelated, all joint relative frequencies are equal to the product of the corresponding marginal relative frequencies. To see this, refer to Table 1.12 and note that the relative frequency of Male Good drivers (0.385) is equal to the product of the marginal relative frequencies of Male (0.700) and Good (0.550) drivers; similarly, (0.165) = (0.300)(0.550), (0.315) = (0.700)(0.450), and (0.135) = (0.300)(0.450). (Incidentally, the opposite is also true: if either of these features is present in a joint distribution, we may conclude that the attributes are independent.)

This definition of independence applies to variables as well. Suppose that the files of a number of insured drivers are examined, and the number

Table 1.15
Joint distribution B

	Sex		
Skill	Male	Female	Total
Good	0	110	110
Bad	90	0	90
Total	90	110	200

of claims they made in two successive years is recorded.

Conditional distributions of claims

Number of	Number of claims, 19X5			
claims, 19X4	0	1	2	Total
0	0.90	0.08	0.02	1.0
1	0.90	0.08	0.02	1.0
2	0.90	0.08	0.02	1.0

You will note that the conditional distribution of the number of claims in 19X5 is the same regardless of the number of claims in 19X4. In this case, we can say that the two variables are independent of one another. (This example is not an accurate reflection of reality. In fact, drivers who make more claims in a particular year also tend to make more in the next year.)

Independence is an extreme condition. Most relationships arising in practice vary in strength from very weak to very strong. The next two sections describe measures of the strength of a relationship between attributes and between variables.

1.10 MEASURING THE RELATIONSHIP BETWEEN ATTRIBUTES

The conclusion that sex and driving skill are unrelated derives, of course, from the fictitious data presented in Table 1.12. Suppose instead that the joint distribution of sex and driving skill was as shown in Table 1.14 or 1.15.

Table 1.14
Joint distribution A

	Sex		
Skill	Male	Female	Total
Good	110	0	110
Bad	0	90	90
Total	110	90	200

According to Table 1.14, all male drivers are Good and all female drivers are Bad. According to Table 1.15, all male drivers are Bad and all female drivers Good. In either case, it is reasonable to claim that the two attributes are perfectly related, in the sense that knowing one attribute is tantamount to knowing the other. For example, given Table 1.15, knowing that a driver is female is knowing that she is a Good driver; knowing that a driver is male is knowing that he is a Bad driver.

An attribute, X, is said to be *perfectly related* to another attribute, Y, if knowledge of Y implies knowledge of X. A perfect relationship, as defined here, may not be symmetric; it is not difficult to construct examples where X is perfectly related to Y, but not Y to X.

Real-world relationships usually fall somewhere between the two extremes of no relationship and perfect relationship. Therefore, we wish to construct a measure of the "extent" ("degree," "strength") of a relationship between two attributes.

One such measure is the *coefficient of association* (P), defined as

$$P = \frac{1}{q-1} \sum\sum \frac{[r(x,y) - r(x)r(y)]^2}{r(x)r(y)}. \tag{1.15}$$

We have in mind a joint distribution of two attributes, X and Y, arranged in the form of a table. $r(x,y)$ denotes the relative frequency of a pair of classes (x,y). $r(x)$ and $r(y)$ are the marginal relative frequencies of classes x and y respectively. q is the smaller of the number of rows (m) and of columns (k) of the table, that is, $q = \min(m,k)$. The double summation symbol indicates that P is a sum of terms, one for each *pair* of classes (x,y).

Let us first illustrate the calculation of this coefficient, and then discuss its properties. Suppose that the joint relative frequency distribution of sex and skill is as shown in Table 1.16.

Table 1.16
Joint distribution C

	Sex		
Skill	Male	Female	Total
Good	0.25	0.20	0.45
Bad	0.45	0.10	0.55
Total	0.70	0.30	1.00

There are $m = 2$ rows and $k = 2$ columns; hence, $q = \min(2,2) = 2$,

and

$$P = \frac{1}{(2-1)}\Big[\frac{(0.25 - 0.45 \times 0.70)^2}{0.45 \times 0.70} + \frac{(0.20 - 0.45 \times 0.30)^2}{0.45 \times 0.30} + \\ + \frac{(0.45 - 0.55 \times 0.70)^2}{0.55 \times 0.70} + \frac{(0.10 - 0.55 \times 0.30)^2}{0.55 \times 0.30}\Big] \\ = 0.0813.$$

What are the properties of this coefficient? It can be shown of the coefficient of association that:

(a) it has a value always in the range 0 to 1;

(b) it is equal to 0 when the two attributes are independent, in which case $r(x, y) = r(x)r(y)$;

(c) it is equal to 1 when one attribute is perfectly related to the other; this will be the case when each row (if the number of rows is greater than or equal to the number of columns), or each column (if the number of rows is less than or equal to the number of columns) of the joint frequency or relative frequency distribution contains a single non-zero entry; and finally,

(d) the greater the deviations of the actual joint relative frequencies, $r(x, y)$, from those that would be expected had the two attributes been independent—in which case, $r(x, y) = r(x)r(y)$—the greater is the value of the coefficient.

The P-coefficient, therefore, may be interpreted as a standardized measure of the degree of association between two attributes.

Example 1.7 In order to investigate the relationship between the price of ten-speed bicycles purchased and various buyer characteristics, a manufacturer mailed the following short questionnaire to 200 bicycle buyers.

1. What price did you pay for your bicycle?
   ___ $100 to $199
   ___ $200 to $299
   ___ $300 and over
2. How old are you?
   ___ under 18
   ___ 18 to 30
   ___ over 30
3. Are you ___ male? ___ female?
4. What is your family income?
   ___ under $20,000
   ___ $20,000 to $30,000
   ___ over $30,000
5. In what area do you live?
   ___ urban
   ___ suburban
   ___ rural
6. Do you live in a
   ___ house?
   ___ apartment?
   ___ other?
7. Do any of your friends own bicycles?
   ___ yes
   ___ no
8. How long did you contemplate buying before making the actual purchase?
   ___ less than 1 month
   ___ 1 to 3 months
   ___ over 3 months

It was felt that by knowing if age, sex, family income, etc. were related to the price of the bicycle purchased, an advertising strategy could be drawn up that would take advantage of this information.

Of the 200 questionnaries sent out, 84 were returned. Table 1.17 summarizes the responses to the first two questions. The numbers in parentheses are relative frequencies. For example, 21 buyers (25% of the 84 who responded) stated that their age was between 18 and 30, and that the price they paid for their bicycles was between $200 and $299.

Table 1.17
Survey of bicycle buyers

	Age (years)			
Price ($)	Under 18	18 to 30	Over 30	Total
100 to 199	20 (0.238)	7 (0.083)	1 (0.012)	28 (0.333)
200 to 299	6 (0.071)	21 (0.250)	6 (0.071)	33 (0.393)
300 and more	3 (0.036)	7 (0.083)	13 (0.155)	23 (0.274)
Total	29 (0.345)	35 (0.417)	20 (0.238)	84 (1.000)

The coefficient of association between price and age can be calculated using Equation (1.15):

$$P = \frac{1}{3-1}\Big[\frac{(0.238 - 0.345 \times 0.333)^2}{0.345 \times 0.333} + \cdots + \frac{(0.155 - 0.238 \times 0.274)^2}{0.238 \times 0.274}\Big] = 0.230.$$

Six other, similar tabulations can be made. In each case, price is one of the two attributes examined, and the question is whether there is any relationship between price and the other attribute. The results are summarized in the following table:

Price *vs.:*	*P*
Age	0.230
Sex	0.014
Family income	0.037
Location of home	0.005
Type of residence	0.047
Friends own bicycles	0.097
Time before purchase	0.054

The attribute most strongly related to price, therefore, is age. To see what type of relationship exists between price and age, we examine the three conditional distributions of price given age:

Conditional distributions of price given age

Price ($)	Under 18	18 to 30	Over 30
100 to 199	0.69	0.20	0.05
200 to 299	0.21	0.60	0.30
300 and more	0.10	0.20	0.65
	1.00	1.00	1.00

For example, of the 35 buyers in the 18 to 30 age group, 7 (20%) paid $100 to $199; 21 (60%) paid $200 to $299; and so on. It is clear that those under 18 tend to buy in the low price range; those in the 18 to 30 age group tend to buy in the medium price range; and those over 30 tend to buy in the high price range.

1.11 CORRELATION

We turn now to the measurement of the strength of the relationship between two variables.

Example 1.8 Figure 1.7 shows the behavior of two indexes of stock prices over a five-year period. The first is an index of the prices of shares of food companies; the second is an index of the prices of shares of chemical companies.

It appears that the two time series are closely related, in the sense that they tend to vary in unison. The relationship is more clearly illustrated

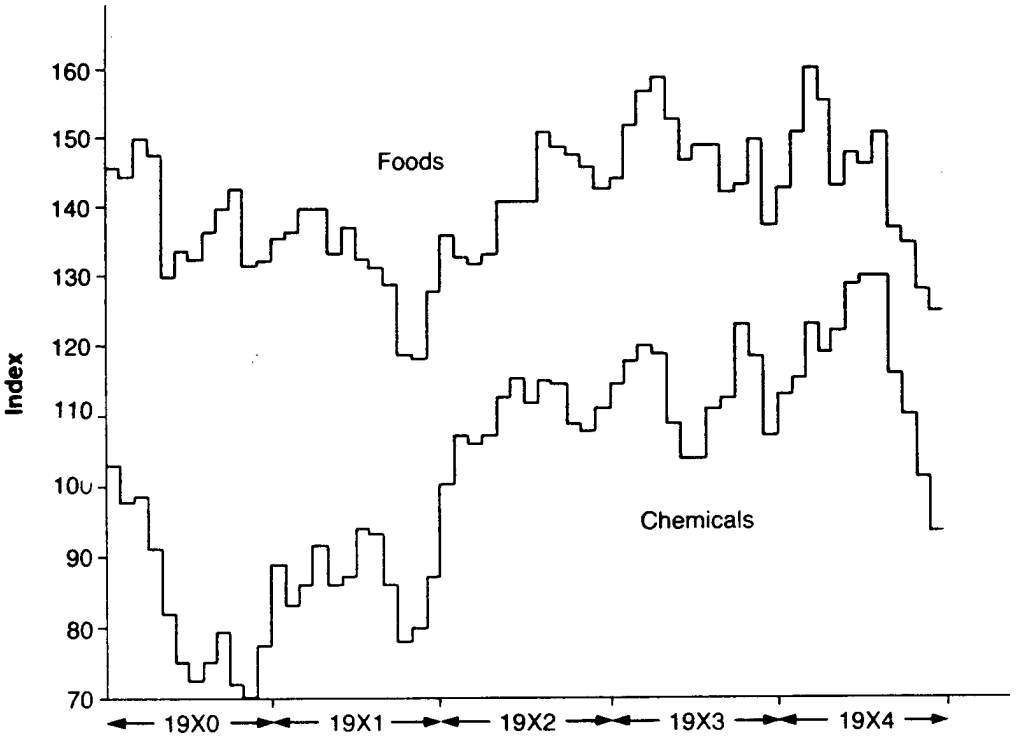

Figure 1.7
Stock price indexes

in the *scatter diagram* shown in Figure 1.8. Each point in this diagram represents one pair of monthly values of the two indexes. For example, the values of the two indexes in the first three months of 19X4 are shown in the following table.

Year	Month	Foods index	Chemicals index
19X4	Jan.	143.0	113.3
	Feb.	151.1	114.6
	Mar.	160.5	122.8
	...	...	...

These three pairs of values are identified in Figure 1.8.

Figure 1.8 confirms the impression that the two indexes tend to move together. It seems that high values of one index tend to be associated with high values of the other index, and that low values of one index are associated with low values of the other. The relationship between the two indexes can

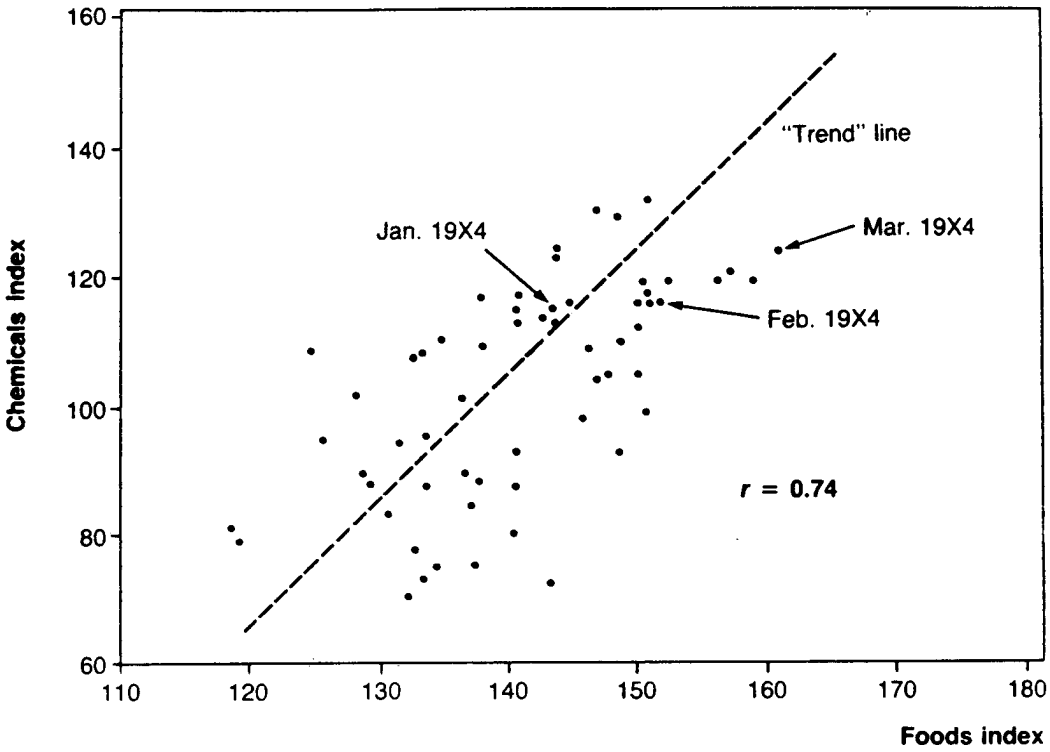

Figure 1.8
Scatter diagram: foods and chemicals

be described roughly by an upward-sloping line, the "trend" line in Figure 1.8.

Figure 1.9 is a scatter diagram of two other monthly indexes of stock prices over the same period. The first is an index of the price of shares of textile and clothing firms, and the second is an index of the price of shares of companies involved in the extraction of primary metals. As in the case of foods and chemicals, the relationship between the primary metals and the textile and clothing indexes is such that the greater the value of one index, the greater tends to be the value of the other index, and vice versa. However, when we compare the two scatter diagrams (Figures 1.8 and 1.9), it appears that there is a difference in the "strength" ("extent," "degree") of the relationship between the two pairs of variables: it seems that the foods and chemicals indexes are more closely related than the primary metals and textile and clothing indexes; that is, the values of the first two indexes tend to cluster more closely around the trend line than the values of the other

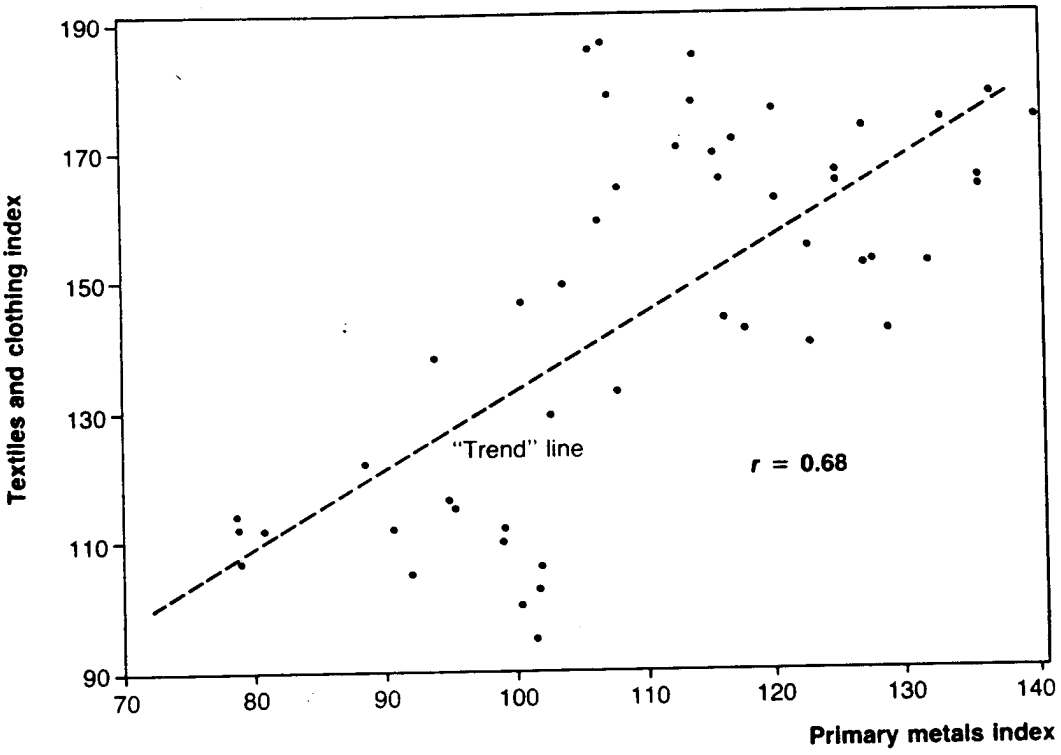

Figure 1.9
Scatter diagram: textiles and primary metals

two indexes.

The scatter diagram is a useful tool for obtaining an impression of the nature and extent of the relationship between two variables. Graphs, however, also have certain limitations: their construction is time-consuming, and sometimes appearances may be deceptive. The question then may be posed: Can the degree of the relationship between two variables be summarized into a single number? With certain qualifications, the answer is yes. A widely used measure for this purpose is the *correlation coefficient*, which we shall now describe.

Suppose there are n pairs (x, y) of values of two variables, X and Y. The *correlation coefficient* of X and Y is defined as:

$$r = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{\sum(x - \bar{x})^2}\sqrt{\sum(y - \bar{y})^2}}. \tag{1.16}$$

We illustrate first the calculation of r and then describe its properties. Suppose we have three pairs of observations for variables X and Y: (1, 1), (−1, 0), and (0, 2). Then the following are calculated:

x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$
1	1	1	0	0	1	0
−1	0	−1	−1	1	1	1
0	2	0	1	0	0	1
0	3	0	0	1	2	2
$(\bar{x}=0)$	$(\bar{y}=1)$			$\sum(x-\bar{x})(y-\bar{y})$	$\sum(x-\bar{x})^2$	$\sum(y-\bar{y})^2$

Therefore, the correlation coefficient of X and Y is

$$r = \frac{1}{\sqrt{2}\sqrt{2}} = 0.5.$$

The properties of the correlation coefficient can be best understood with the help of Figure 1.10, which shows six types of scatter diagrams and the associated approximate value of the correlation coefficient. These properties are:

(a) The value of the correlation coefficient is always between −1 and +1.

(b) When all the pairs of values of X and Y lie on a straight line, r is equal to +1 if the line is upward-sloping (Figure 1.10.a, or to −1 if the line is downward-sloping (Figure 1.10.d).

(c) When the pairs of (x, y) values tend to cluster along an upward-sloping line, the value of r will be a positive number between 0 and 1; the closer the points cluster around the line, the closer r will be to +1 (Figure 1.10.b). Similarly, the closer the points cluster around a downward-sloping line, the closer will r be to −1 (Figure 1.10.c). Depending on whether the line is upward- or downward-sloping, we say that the variables are *positively* or *negatively correlated.*

(d) When, as in Figure 1.10.e, there is no apparent linear "trend" in the relationship between X and Y, r will tend to be near 0. In fact, $r = 0$ if X and Y are independent. Note, however, that r will be near 0 also for certain types of curvilinear relationships, as, for example, in Figure 1.10.f.

The correlation coefficient may therefore be described as a standardized measure of the degree to which two variables are *linearly* related.

For calculations by hand, it is convenient to write Equation (1.16) in an alternative form:

$$r = \frac{\sum xy - n\bar{x}\bar{y}}{\sqrt{\sum x^2 - n\bar{x}^2}\sqrt{\sum y^2 - n\bar{y}^2}}. \tag{1.17}$$

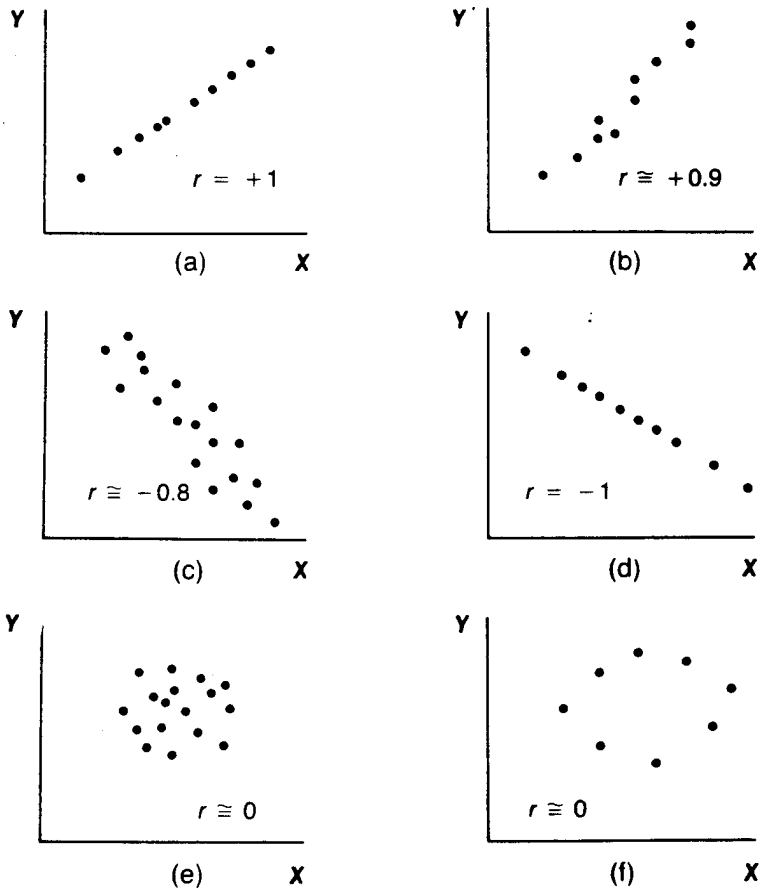

Figure 1.10
Scatter diagrams and associated correlation coefficients

Example 1.8 (Continued) The calculation of the correlation coefficient of the foods and chemical stock price indexes is illustrated in Table 1.18. Available are 60 pairs of observations.

We find, $\bar{x}$= (8,486.0)/60 = 141.433, and $\bar{y}$ = (6,402.1)/60 = 106.702; also,

$$\sum x^2 - n\bar{x}^2 = (1,205,825) - (60)(141.433)^2 = 5,627,$$

$$\sum y^2 - n\bar{y}^2 = (694,064) - (60)(106.702)^2 = 10,945,$$

$$\sum xy - n\bar{x}\bar{y} = (911,340) - (60)(141.433)(106.702) = 5,869,$$

Table 1.18
Calculations, Example 1.8

Year	Month	Index of: Foods, x	Chemicals, y	x^2	y^2	xy
19X0	Jan.	146.6	102.8	21,491	10,568	15,070
	Feb.	144.7	97.8	20,938	9,584	14,166
	Mar.	150.4	98.7	22,620	7,742	14,844
	...	...	...	...	...	...
19X4	Nov.	127.9	101.3	16,358	10,262	12,956
	Dec.	125.4	93.3	15,725	8,705	11,700
		8,486.0	6,402.1	1,205,825	694,064	911,340

and

$$r = \frac{5,869}{\sqrt{5,627}\sqrt{10,945}} = 0.740.$$

The correlation coefficient of the primary metals and the textile and clothing indexes can be shown to be 0.680. The calculated coefficients confirm the visual impression obtained from Figure 1.9, that the relationship between the metals and clothing indexes is not as strong as that between foods and chemicals.

As with all other summary measures, the correlation coefficient* does not convey as much as a diagram does. This disadvantage, however, is often

* Equation (1.16) can also be written as

$$r = \frac{\frac{1}{n}\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\frac{1}{n}\sum(x-\bar{x})^2}\sqrt{\frac{1}{n}\sum(y-\bar{y})^2}}.$$

The numerator is called the *covariance* of X and Y, while the denominator is the product of the standard deviations of X and Y. Thus the correlation coefficient can also be described as

$$Cor(X,Y) = \frac{Cov(X,Y)}{Sd(X)Sd(Y)}.$$

We shall not make use of this relationship in this chapter, and will not elaborate on the interpretation of the covariance. The reader should keep in mind that there are several ways of expressing the correlation coefficient.

compensated by the convenience of being able to deal with only a single number instead of large sets of observations.

Example 1.8 (Continued) Table 1.19 shows the correlation coefficients of pairs of the following sub-indexes of industrial stocks: foods (F), textiles and clothing (T&C), primary metals (PM), metal fabricating (MF), and chemicals (C). The correlation coefficients are calculated on the basis of 60 monthly observations on all these indexes over the same five-year period.

Table 1.19
Correlation coefficients of stock price indexes

	F	T&C	PM	MF	C
F	1.000	0.686	0.910	0.749	0.740
T&C		1.000	0.680	0.779	0.858
PM			1.000	0.693	0.801
MF				1.000	0.626
C					1.000

We have already examined two of these correlation coefficients, that of foods and chemicals (0.740), and that of primary metals and textiles and clothing (0.680). The correlation coefficient of chemicals and foods is, of course, the same as that of foods and chemicals: the correlation coefficient refers to a pair of variables, and the order in which the variables are considered is irrelevant. Table 1.19, therefore, does not show redundant entries. Note also that the correlation coefficient of a variable with itself is 1.0.

The correlation coefficients of Table 1.19 allow us to tell at a glance that all pairs of indexes are positively and fairly strongly correlated. The highest correlation is that between foods and primary metals (0.910), while the least correlated are the metal fabricating and chemicals indexes (0.626).

When the observations on the variables X and Y have been grouped in the form of a joint frequency or relative frequency distribution, the calculation of the correlation coefficient can be simplified. Think of a joint frequency distribution, arranged as usual in the form of a table:

	Values of Y			
Values of X	$\cdots$	y	$\cdots$	Total
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
x	$\cdots$	$f(x,y)$	$\cdots$	$f(x)$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
Total	$\cdots$	$f(y)$	$\cdots$	n

As shown in the table, the frequency of the pair of values (x, y) is denoted by $f(x, y)$. Denote the marginal frequency of x by $f(x)$, and the marginal frequency of y by $f(y)$. The grand total of the frequencies is, of course, n. The correlation coefficient can now be defined as:

$$r = \frac{\sum\sum(x-\bar{x})(y-\bar{y})f(x,y)}{\sqrt{\sum(x-\bar{x})^2 f(x)}\sqrt{\sum(y-\bar{y})^2 f(y)}}. \tag{1.18}$$

The double summation symbol ($\sum\sum$) indicates that the numerator is the sum of terms $(x-\bar{x})(y-\bar{y})$ for all *pairs* of values of x and y.

Equation (1.18) is simply a different version of (1.16). It takes advantage of the fact that there are altogether $f(x, y)$ terms of the form $(x-\bar{x})(y-\bar{y})$, $f(x)$ terms of the form $(x-\bar{x})^2$, and $f(y)$ terms of the form $(y-\bar{y})^2$.

If we use a similar notation, $r(x, y), r(x), r(y)$, for the joint and marginal relative frequencies, (1.18) can also be written in the following form:

$$r = \frac{\sum\sum(x-\bar{x})(y-\bar{y})r(x,y)}{\sqrt{\sum(x-\bar{x})^2 r(x)}\sqrt{\sum(y-\bar{y})^2 r(y)}}. \tag{1.19}$$

It should be realized that Equations (1.18) and (1.19) are identical expressions: (1.19) is obtained from (1.18) by dividing both numerator and denominator by n. Further algebraic manipulations produce yet two more versions, which are more efficient for calculations by hand or by computer:

$$\begin{aligned} r &= \frac{\sum\sum xyf(x,y) - n\bar{x}\bar{y}}{\sqrt{\sum x^2 f(x) - n\bar{x}^2}\sqrt{\sum y^2 f(y) - n\bar{y}^2}} \\ &= \frac{\sum\sum xyr(x,y) - \bar{x}\bar{y}}{\sqrt{\sum x^2 r(x) - \bar{x}^2}\sqrt{\sum y^2 r(y) - \bar{y}^2}}. \end{aligned} \tag{1.20}$$

The above expressions are frequently used also to approximate the correlation coefficient of a joint frequency or relative frequency distribution in which the variables are classified jointly into class intervals. In this case, the midpoints of the intervals are used as representative values of all the observations in an interval.

Example 1.9 Among the criteria used in judging applications to graduate business schools at the time of this writing, the candidate's performance in the Graduate Management Admission Test (GMAT) usually carries substantial weight. The GMAT is not, of course, the only criterion used in admissions decisions. Other sources of information regarding the applicant's potential are the transcript of undergraduate courses and grades, letters of

recommendation, interviews, and the application form itself; different graduate schools attach different weights to these inputs in deciding whether an applicant should be admitted or not. According to its sponsors, the GMAT is currently used by about 700 institutions in the U.S., Canada, and other countries, and is required of every applicant by more than 500 schools.

The GMAT yields three scores: verbal, quantitative, and total. A candidate's "raw" score is based on the number of correct answers to a large number of multiple-choice questions, minus a certain fraction of the wrong answers—an adjustment intended to discourage haphazard guessing. The total raw scores are converted into scaled scores, so that the mean of the scaled total scores is equal to 500 and their standard deviation equal to 100. The same process is applied to verbal and quantitative scores; the mean and standard deviation of the scaled verbal and quantitative scores are 30 and 8 respectively.

The purpose of the test is to predict a student's performance in a graduate business program. The questions are so designed and the scores so calculated that a higher score in the GMAT is taken as an indication of potentially better performance in the graduate school. Thus, it is quite appropriate to ask how well the test predicts performance and how it compares with other criteria used in admissions decisions.

It should be borne in mind that an examination of the relationship between test scores (or any other quantitative measure used in admissions) and performance in the graduate school is possible only for students who have already been admitted by the school. It would have been useful to have an indication of how well a rejected applicant might have performed had he or she been admitted, but, for obvious reasons, such information is not available.

Figure 1.11 is a scatter diagram showing the total GMAT score and the average grade in the first year of the MBA program of 80 full-time students at a northeastern university.

Using the 80 pairs of observations and a computer program, it is an easy matter to calculate the exact value of the correlation coefficient, which can be shown to be equal to 0.517. However, in order to illustrate the accuracy of the approximate method of calculating the correlation coefficient, let us pretend that neither the original observations nor the scatter diagram are available. Instead, suppose that we have been furnished with the joint relative frequency distribution shown in Table 1.20. (This distribution can be obtained from Figure 1.11 by drawing a "grid" of lines to form the corresponding class intervals, counting the number of observations falling into each grid square, and dividing this frequency by 80 to obtain the relative frequency of each cell. Observations lying on the boundary of two intervals were classified into the lower interval.)

Treating the midpoints of the intervals as representative of all observa-

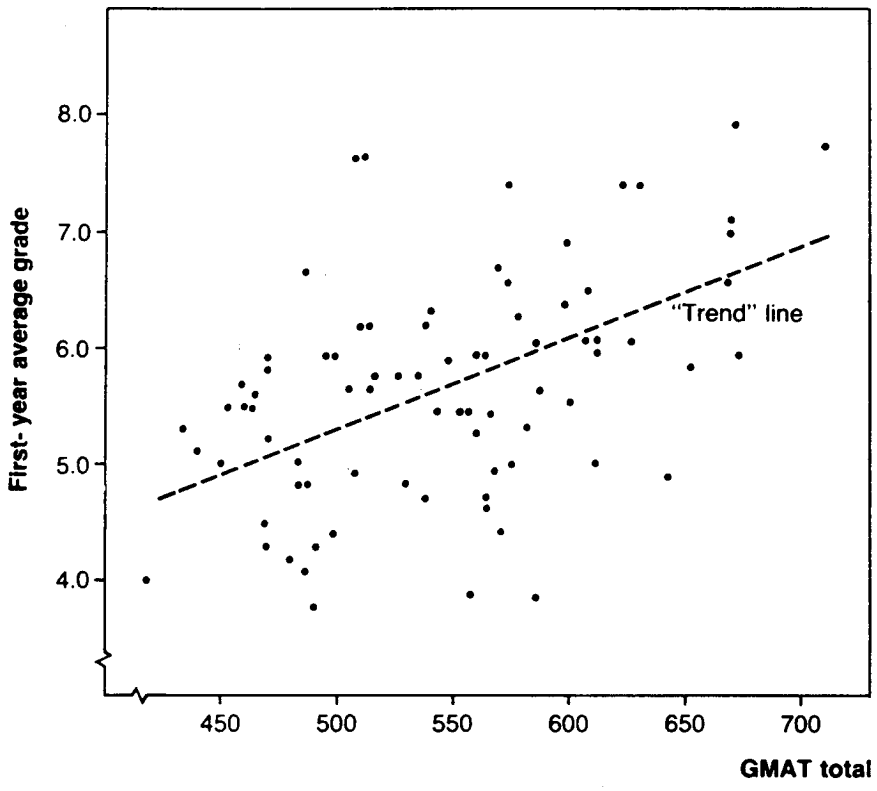

Figure 1.11
GMAT total scores and first-year average grades

Table 1.20
Joint relative frequency distribution of GMAT total score and first-year average grade

GMAT total (X)		First-year average grade (Y)					
Interval:		3 to 4	4 to 5	5 to 6	6 to 7	7 to 8	
	Midpoint:	3.5	4.5	5.5	6.5	7.5	Total
400 to 500	450	0.025	0.125	0.150	0.012		0.312
500 to 600	550	0.025	0.100	0.212	0.112	0.038	0.487
600 to 700	650		0.025	0.075	0.050	0.038	0.188
700 to 800	750					0.012	0.012
Total		0.050	0.250	0.437	0.174	0.088	1.000

tions in the interval, we calculate first

$$\bar{x} = \sum xr(x) = (450)(0.312) + \cdots + (750)(0.012) = 539.450,$$

$$\bar{y} = \sum yr(y) = (3.5)(0.050) + \cdots + (7.5)(0.088) = 5.494,$$

$$\sum x^2 r(x) = (450)^2(0.312) + \cdots + (750)^2(0.012) = 296,678,$$

$$\sum y^2 r(y) = (3.5)^2(0.050) + \cdots + (7.5)^2(0.088) = 31.196,$$

and

$$\sum\sum xyr(x,y) = (450)(3.5)(0.025) + \cdots + (750)(7.5)(0.012) = 2,988.2.$$

Therefore, the correlation coefficient is approximately

$$\begin{aligned} r &= \frac{\sum\sum xyr(x,y) - \bar{x}\bar{y}}{\sqrt{\sum x^2 r(x) - \bar{x}^2}\sqrt{\sum y^2 r(y) - \bar{y}^2}} \\ &= \frac{2,998.2 - (539.45)(5.494)}{\sqrt{296,678 - (539.45)^2}\sqrt{31.196 - (5.494)^2}} \\ &= \frac{34.462}{(75.311)(1.006)} \\ &= 0.45. \end{aligned}$$

The difference between this value and the correct one (0.517) is rather large in this case. In general, the success of the approximation depends on the degree to which the midpoints are representative of the observations in the cells. A glance at Figure 1.11 will show that the midpoints of the outer cells do not represent well the observations in these cells. The reader should not forget, however, that an exact value can be calculated only when all the observations are available; a certain loss in accuracy must be expected when they are not.

Let us now return to the question: Is the GMAT a good predictor of student performance in the MBA program? From Figure 1.11 or the calculated correlation coefficient, it is obvious that the two variables (GMAT total and first-year average grade) are positively correlated, that is, high values of one variable tend to be associated with high values of the other variable, and vice versa; this tendency is roughly described by the trend line in Figure 1.11 (regression, a method to be discussed later in this text, was used to determine the plotted line). On the other hand, it is also clear that the relationship is far from being exact, and thus the reliability of the GMAT total as a predictor of student performance is not very high.

1.12 MULTIVARIATE DISTRIBUTIONS

It is, of course, possible to form joint distributions of any number of variables or attributes. Table 1.21, for example, shows the joint relative frequency distribution of subscribers to a magazine according to sex, duration of subscription, and place of residence.

Table 1.21
A trivariate distribution

Sex	Duration of subscription	Place of residence	%
Male	Less than 1 yr.	Urban	15
		Rural	9
	1 yr. or more	Urban	16
		Rural	26
Female	Less than 1 yr.	Urban	3
		Rural	6
	1 yr. or more	Urban	7
		Rural	8
	Total		100

From such a distribution we may obtain all the bivariate or univariate distributions. Joint distributions of more than three variables or attributes can be formed in like manner. Measures of correlation or association can be defined as extensions of the bivariate measures. We shall not pursue these topics further, however, as they appear infrequently in business analysis.

1.13 COMPUTER PROGRAMS FOR DATA ANALYSIS

It goes without saying that the processing by hand of raw data of the size commonly found in business would be very tedious, boring, time-consuming, and expensive. Fortunately, many special computer programs are available to perform the calculations and to display the results in a clear and elegant manner.

Such computer programs range from relatively simple undertakings in the public domain, to massive, elaborate, all-but-omniscient, and expensive commercial software. Some are designed for large mainframe computers, but others fit the capacity of even a very modest personal computer.

Especially helpful for serious work in business are computing systems such as BMDP, MINITAB, SAS, and SPSS. Some of these come with separate mainframe and personal computer versions. Each system has its own requirements, specifications, commands, and style of output. In general, however, a system accepts a computer file containing the raw data, and

allows the analyst to create new variables or attributes from the existing ones, to delete or add observations, to summarize the data not only in the manner described in this chapter but in other, more ingenious ways, to display handsome charts and graphs, and to perform further statistical operations. Included in these operations are calculations on special probability distributions, the solution of complicated decision problems, the design and selection of samples, the drawing of inferences from samples, and the execution of simulation, regression, and time series analyses. These statistical operations are described in the chapters that follow.

1.14 IN SUMMARY

We examined various methods by which large sets of data can be reduced or summarized in a manner suitable for the problem at hand and the purpose for which the data were collected. A *frequency* or *relative frequency distribution* shows the number or proportion of observations falling into each class, interval, or category. The *average* or *mean* is the most familiar summary measure of tendency; among others are the *median*, the *mode*, the *quartiles*, and the *deciles* of a distribution. The variation (dispersion) of a distribution about the mean is usually measured by either the *variance* or the *standard deviation* of the distribution.

Index numbers show the relative change in the level of one or more time series.

Bivariate frequency or relative frequency distributions show the number or proportion of the observations falling into classes formed according to two variables or attributes. The univariate measures of tendency and dispersion have their counterparts in bivariate distributions as well. The degree to which two variables are linearly related can be measured by the *correlation coefficient*. A similar measure of association between two attributes is the *P-coefficient*.

PROBLEMS

1.1 The distribution of the number of cars per household is estimated to be as follows:

Number of cars	Number of households (000)
0	1,665
1	4,138
2	1,443
3 or more	313
Total	7,559

Calculate and interpret the mean, the variance, and the standard deviation of this distribution. Assume that all households with 3 or more cars have exactly 3 cars. Construct a graph of the relative frequency distribution of the number of cars per household.

1.2 The distributions of families by number of persons in the years 19X1, 19X6, and 19X8 are given below:

Number of persons	Number of families (000) 19X1	19X6	19X8
2	1,589	2,010	2,088
3	1,042	1,220	1,268
4	1,052	1,289	1,422
5	663	687	696
6 or more	713	521	419
Total	5,059	5,727	5,893

(a) Determine the relative frequency distributions of the number of persons per family in 19X1, 19X6, and 19X8.

(b) Calculate and interpret the mean and the variance of these distributions.

(c) Compare graphically the three relative frequency distributions.

(d) In the light of the above, comment on the changes in the distribution of the number of persons per family in the period 19X1 to 19X8.

1.3 Suppose that variable Y is linearly related to variable X as follows:

$$Y = a + bX,$$

where a and b are some constants. The following three problems illustrate such a case. Suppose that there are n observations on X and Y.

(a) Use Equations (1.1) and (1.6) to show that the mean ($\bar{y}$), variance (s_y^2) and standard deviation (s_y) of the observations on Y are related to the corresponding characteristics of the observations on X as follows:

$$\bar{y} = a + b\bar{x},$$

$$s_y^2 = b^2 s_x^2,$$

$$s_y = |b| s_x.$$

(b) Same as (a), but using Equations (1.3) and (1.9).

1.4 Use the results of Problem 1.3 to find the mean and standard deviation of the age distribution of insured drivers of Example 1.2 with age measured in *months*. Recall that the mean age in *years* is 37.8 and the standard deviation of the age distribution, again expressed in years, is 15.10.

1.5 According to the instructions of a popular almanac, "to convert degrees Fahrenheit into degrees centigrade, subtract 32, multiply by 5, and divide by 9; to convert degrees centigrade to degrees Fahrenheit, multiply by 9, divide by 5, and add 32."

(a) Express degrees centigrade (C) as a linear function of degrees Fahrenheit (F).

(b) Express degrees Fahrenheit (F) as a linear function of degrees centigrade (C).

(c) The distribution of the maximum daily temperature in July, based on past records and expressed in degrees Fahrenheit, has mean 88 and variance 5. Calculate the mean, the variance, and the standard deviation of the same temperature expressed in degrees centigrade. (Use the results of Problem 1.3.)

1.6 As was stated in Example 1.9, the GMAT "raw" total scores are converted into "scaled" total scores in such a way that the mean of the scaled scores is equal to 500 and their standard deviation equals 100. Let X be the raw and Y the scaled score of a candidate. Let $\bar{x}$ and s_x^2 be the mean and variance of all n raw scores.

(a) Show that the scaled scores will have mean 500 and standard deviation 100 if they are calculated as a linear transformation of the raw scores, $y = a + bx$, where $b = 100/s_x$ and $a = 500 - b\bar{x}$.

(b) Suppose that the mean and variance of the raw scores are $\bar{x} = 450$ and $s_x^2 = 14{,}884$ respectively. Calculate the scaled score of a candidate whose raw score is 585.

1.7 Investigate and report on the methods used to construct the following indexes: (a) the national consumer price index; (b) the national wholesale price index; (c) the Dow-Jones averages; (d) the Standard and Poor's stock price indexes; (e) the indexes of the principal Stock Exchanges in the country.

1.8 (a) Explain how you would combine the prices (or price indexes) of refrigerators, ranges, washing machines, dryers, vacuum cleaners, and small appliances to form an index of the price of household appliances. What weights would you use?

(b) Explain how you would combine the prices (or price indexes) of gasoline, motor oil, tires, batteries, automobile insurance, automobile repairs, and automobile registration fees to form an index of the cost of automobile operation and maintenance. What weights would you use?

(c) Explain how you would combine the prices (or price indexes) of men's wear, ladies' wear, and children's wear to form an index of the price of clothing. What weights would you use in the construction of the index?

(d) Explain how you would combine the prices (or price indexes) of television sets, audio equipment, movie projectors, and cameras to form an index of the cost of home entertainment. What weights would you use?

1.9 Consider the data shown in Table 1.22.

Table 1.22
Labor statistics, Problem 1.9

Year/ month	Population 15 years of age and over	Labor force[1]	Em- ployed	Unem- ployed	Partici- pation rate[2]	Unem- ployment rate[3]
	(Thousands of persons)				(%)	(%)
19X0 Dec.	18,139	11,445	10,635	810	63.1	7.1
19X1 Jan.	18,165	11,407	10,566			
19X1 Feb.	18,192	11,511	10,433			
19X1 Mar.	18,213	11,609	10,528			
19X1 Apr.	18,235	11,585	10,319			
19X1 May	18,258	11,880	11,026	854	65.1	7.2

Notes: [1] Number of persons 15 years of age and over who are actively seeking work. [2] Labor force as a percentage of the population 15 years of age and over. [3] Number unemployed as a percentage of the labor force.

(a) Fill in the missing entries. For each of the above six series, construct an index with base January 19X1 = 100.

(b) Construct an index for the number employed with base March 19X1 = 100. Use the original series for the conversion.

(c) Convert the index series for the number unemployed with base January 19X1 = 100 to one with base March 19X1 = 100. In general, how do you change the base of an index series?

1.10 Construct a consumer price index for the years 19X7, 19X8, and 19X9, with base 19X7 = 100, using as weights the consumer expenditures in 19X8 shown in Table 1.23.

Table 1.23
Data for Problem 1.10

Category	Consumer expenditure 19X8 ($ million)	Consumer price indexes, 19X1 = 100		
		19X7	19X8	19X9
Food	3,364	180.1	208.0	235.4
Shelter	3,449	159.3	170.8	180.5
Household operation	819	177.4	194.3	211.6
Furnishings and equipment	847	144.3	150.0	162.0
Clothing	1,353	141.0	146.4	159.9
Personal care	337	157.5	169.4	183.7
Medical and health care	397	151.0	160.9	176.7
Smoking and alcoholic beverages	640	143.8	155.5	166.7
Transportation	2,291	153.3	162.2	178.0
Recreation	955	141.5	146.4	156.3
Reading	122	153.5	159.7	176.4
Education	149	117.2	124.0	129.4
Miscellaneous expenses	435	n.a.	n.a.	n.a.

1.11 The following table shows the annual income (X) and savings (Y) of five families:

Family No.:	Income, X ($000)	Savings, Y ($000)
1	30	12
2	24	6
3	20	8
4	40	10
5	16	4

(a) Plot the observations in a scatter diagram. Does there appear to be a relationship between income and savings?

(b) Calculate and interpret the correlation coefficient of income and savings.

1.12 Suppose that the joint relative frequency distribution of family size (Y) and

the number of cars owned (X) is as follows:

Number of cars	Family size 2	3	4	Total
0	0.3	0.1		0.4
1	0.1	0.2	0.1	0.4
2	—	0.1	0.1	0.2
Total	0.4	0.4	0.2	1.0

That is, 30% of the families consist of 2 members and have no car, and so on. (a) Calculate and interpret the mean and variance of X and Y. (b) Calculate and interpret the correlation coefficient of X and Y.

1.13 Two variables, X and Y, are transformed linearly to:

$$X' = a + bX,$$
$$Y' = c + dY,$$

where a, b, c, and d are some constants. Use Equation (1.16) to determine the correlation coefficient of X' and Y' as a function of the correlation coefficient of X and Y.

1.14 Suppose that X and Y are linearly related to the *same* third variable, Z. That is, suppose that

$$X = a + bZ, \;\; Y = c + dZ,$$

where a, b, c, and d are some constants. Suppose that n observations are available on X, Y, and Z.

(a) Using Equation (1.16), show that the correlation coefficient of X and Z is equal to $+1$ if $b > 0$, or to -1 if $b < 0$.

(b) Again using Equation (1.16), show that the correlation coefficient of X and Y is equal to $+1$ if the constants b and d have the same sign, or to -1 if b and d have opposite signs.

1.15 As a member of a consumer research firm, you are asked to formulate a plan for monitoring supermarket prices on a regular basis. Once a week, investigators will visit a number of stores in the city and record the prices of items appearing on a carefully prepared list. Each item will be specified precisely by brand name, size, quantity, etc. You realize, of course, that it would be impossible to monitor the prices of *all* items carried by a supermarket—a typical supermarket carries between 20,000 and 30,000 different items. Also, the items are not of the same importance in the consumers' budgets, so their prices must be weighted in some fashion. You have to consider the possibility that some items on your list are not carried by all supermarkets, and that there may or may not be a close substitute offered. Finally, the procedure for monitoring the prices must be fair in the sense that it should not be unduly influenced by a store's "specials" or "loss leaders." As briefly and precisely as possible, explain how you would go about setting up the monitoring procedure.

1.16 Five hundred male adults were interviewed concerning their consumption of beer, other alcoholic beverages, and tobacco. They were classified according to their level of consumption as Light and Heavy consumers, with the following results:

Beer consumption	Alcoholic beverage consumption		
	Light	Heavy	Total
Light	140	160	300
Heavy	110	90	200
Total	250	250	500

Beer consumption	Tobacco consumption		
	Light	Heavy	Total
Light	200	100	300
Heavy	40	160	200
Total	240	260	500

Of the two attributes, consumption of other alcoholic beverages and consumption of tobacco, which is more closely related to beer consumption? Explain and justify your conclusions.

1.17 Last year's distribution of drivers by age and accident involvement was as follows:

Age of driver (years)	Licensed drivers (000)	Involved in accident(s) (000)
16 to 19	327	56
20 to 24	605	71
25 to 34	1,188	97
35 to 44	846	57
45 to 54	739	44
55 to 64	518	27
65+	339	14
Not known or reported		17
Total	4,562	383

(a) "Age and accident involvement are obviously not independent of one another." Formally justify this statement.

(b) Calculate the accident rate for 16- to 19-year-old drivers. Calculate the accident rates for all listed age intervals.

(c) Without actually doing any calculations, but with the help of a diagram, describe how the above data can be used to estimate the accident rate for any one year of age (e.g., 19, 21, or 57).

1.18 As we know, the correlation coefficient is a measure of the degree of a linear relationship between two variables. It is not intended to measure the strength of a relationship between attributes, since the latter do not have a natural numerical representation. It is, however, tempting to use numerical codes for the categories

of the attributes (for example, 0 for "not satisfied," 1 for "moderately satisfied," and 2 for "very satisfied"), and then calculate the correlation coefficient as if the attributes were variables.

Construct a simple numerical example to show that the value of the correlation coefficient will vary depending on the codes used. Comment on this approach.

1.19 The National Automobile Association conducts annually a survey of its members. Included in the questionnaire are such questions as: What type of car do you own? How is it equipped? What distance do you drive annually? How satisfied are you with your car? 15,300 car owners responded to the 19X5 survey.

(a) The 19X5 responses to the question "How satisfied are you with your car?" tabulated according to the origin of the car, were as follows:

Origin	Very satisfied	Moderately satisfied	Unsatisfied	Total
North American	6,400	2,600	1,000	10,000
Japanese	3,200	500	100	3,800
European	1,000	300	200	1,500
Total	10,600	3,400	1,300	15,300

Would you say that an owner's satisfaction depends on where the car came from? Carefully justify your answer.

(b) The same 19X5 responses, tabulated according to the size of the respondents' cars, were as follows:

Size of car	Very satisfied	Moderately satisfied	Unsatisfied	Total
Large	5,500	1,400	800	7,700
Medium	3,150	1,100	350	4,600
Small	1,950	900	150	3,000
Total	10,600	3,400	1,300	15,300

Does an owner's satisfaction depend on the size of the car? Carefully justify your answer.

(c) Would you say that an owner's satisfaction depends *more* on the size of the car *than* on the car's origin? Carefully justify your answer.

1.20 The annual surveys of consumer finances conducted by the Survey Research Center provide useful information on such aspects of household financing as distribution of income, ownership of durables, and possession of financial assets. The survey reported here paid special attention to the increasing usage of credit cards. 2,576 heads of households throughout the country were interviewed during the months of January, February, April, and May. The questions were identical in all these interviews and related to the respondents' use of and attitude toward credit cards, and to characteristics of their households (income, age of the head, etc.). Of the 2,576 respondents, 1,290 used some type of credit card; 872 used a gasoline card (issued by oil companies), 414 used a bank card (issued under the names: Mastercard, Unicard, BankAmericard, etc.), 239 used a travel and entertainment card (Diners Club, Carte Blanche, American Express), and 913 used a store credit

Table 1.24
Use of credit cards by type and income

Annual family income	Percentage of families who use credit cards				
	Gas	Bank	T&E	Store	Any type
Less than $3,000	8	2	2	11	17
$3,000 to 4,999	14	3	4	12	24
$5,000 to 7,499	24	11	5	23	39
$7,500 to 9,999	32	14	8	36	54
$10,000 to 14,999	45	22	10	50	67
$15,000 to 19,999	65	30	14	56	74
$20,000 to 24,999	68	40	30	66	84
$25,000 and over	67	37	40	61	81
All families	872	414	239	913	1,290

card (issued by major retail firms, such as Sears, Macy's, etc.). Table 1.24 shows the percentage of households using credit cards in each income interval. (The information in this and the other table for this problem is presented very much as it appeared in the original source. Read carefully.)

In one of the questions, "respondents who reported using a credit card were asked how much they spent and charged on that particular type of credit card in the previous month. The question was formed in this manner because of suspected decline of recall of transactions further in the past." Table 1.25 shows the conditional distributions of the amount charged by income and by age of the head of the household.

(a) In a diagram, show the relationship between family income, on the one hand, and, on the other, each of the following variables: the percentage of families using gasoline, bank, travel and entertainment, and retail store cards, and cards of all types. Does income affect all these use rates in the same way?

(b) Using the midpoints of the listed income intervals, approximate the mean and the variance of the distribution of family income for users of credit cards. (Use $2,000 and $30,000 as the "midpoints" of the first and last income intervals.)

(c) Using the midpoints of the intervals of the amount charged, approximate the average amount charged in one month on all credit cards. (Use $250 as the "midpoint" of the last interval; do not include in your calculations families who did not report the amount charged.)

(d) Indicate how to calculate the correlation coefficients of the amount charged, on the one hand, and the family income and age of the head of the household on the other.

(e) Calculate the correlation coefficients of (i) amount charged and family income, and (ii) amount charged and age of head of household. (You may want to use a computer program for these calculations.) Which of the two variables (family income, age of head) is more strongly related to the amount charged?

(f) Calculate and plot the relationship between the average amount charged and income. Calculate and plot the relationship between average amount charged and age of head of household.

(g) The respondents were asked to state the amount charged on their cards *in the month previous to the survey.* How does the manner in which the survey was conducted affect the validity of the survey results?

Table 1.25
Amounts charged on all credit cards
(percentage distribution of families[1])

	Amount charged on all credit cards during one month										
	No charges	$1 -14	$15 -29	$30 -49	$50 -74	$75 -99	$100 -149	$150 -199	$200 or more	D.K./ N.A.[3]	No. of families
All families[2]	21	9	14	13	13	6	9	3	8	4	1,290
Total family income:											
Less than $3,000	33	5	23	16	13	3	2	*	2	3	59
$3,000 to 4,999	29	19	22	12	8	2	4	*	3	1	72
$5,000 to 7,499	30	13	18	11	12	5	6	2	2	1	156
$7,500 to 9,999	23	9	19	14	11	5	7	3	4	5	220
$10,000 to 14,999	23	8	15	13	11	6	10	3	7	4	421
$15,000 to 19,999	17	7	9	16	13	9	10	7	6	6	203
$20,000 to 24,999	9	5	5	10	17	5	14	6	21	8	78
$25,000 and over	1	5	2	9	18	8	15	5	31	6	81
Age of head:											
Under 25	19	11	23	16	15	1	4	3	5	3	105
25 to 34	22	6	17	14	12	8	8	4	5	4	284
35 to 44	21	8	12	16	13	6	8	4	10	2	275
45 to 54	23	8	12	10	10	5	11	5	11	5	305
55 to 64	18	8	13	15	15	6	11	2	6	6	192
65 and older	27	14	16	8	12	6	8	2	4	3	129

Notes: [1]All row totals equal 100%. [2]Includes only those families who use credit cards. [3]Declined/no answer. ∗Less than 0.5%.