2 Descriptive Statistics: Tabular and Graphical Presentations

CHAPTER CONTENTS

Statistics in Practice Marks and Spencer: not just any statistical graphics

LEARNING OBJECTIVES

After studying this chapter and doing the exercises, you should be able to construct and interpret several different types of tabular and graphical data summaries.

1 For single qualitative variables: frequency, relative frequency and percentage frequency distributions; bar charts and pie charts.

2 For single quantitative variables: frequency, relative frequency and percentage frequency distributions; cumulative frequency, relative cumulative frequency and percentage cumulative frequency distributions; dot plots, stem-and-leaf plots, histograms and cumulative distribution plots (ogives).

3 For pairs of qualitative and quantitative data: cross-tabulations, with row and column percentages.

4 For pairs of quantitative variables: scatter diagrams.

5 You should be able to give an example of Simpson's paradox and explain the relevance of this paradox to the cross-tabulation of variables.

As explained in Chapter 1, data can be classified as either qualitative or quantitative. **Qualitative data** use labels or names to identify categories of like items. **Quantitative data** are numerical values that indicate how much or how many.

This chapter introduces tabular and graphical methods commonly used to summarize both qualitative and quantitative data. Everyone is exposed to these types of presentation in annual reports (see Statistics in Practice), newspaper articles and research studies. It is important to understand how they are prepared and how they should be interpreted. We begin with methods for summarizing single variables. Section 2.3 introduces methods for summarizing the relationship between two variables.

Modern spreadsheet and statistical software packages provide extensive capabilities for summarizing data and preparing graphical presentations. EXCEL, IBM SPSS and MINITAB are three widely available packages. There are guides to some of their capabilities on the associated online platform.

STATISTICS IN PRACTICE

Marks & Spencer: not just any statistical graphics

Marks & Spencer has a company history going back to 1884. The group is based in London, but has offices across the UK as well as overseas. Most people are likely to have come across its promotional activities and its advertising slogan 'Your M&S'. Marks & Spencer advertisements have featured a long list of well-known faces, including Twiggy, Erin O'Connor, David Beckham, Claudia Schiffer, Rosie Huntington-Whiteley and Antonio Banderas.

Marks & Spencer's shares are traded on the London Stock Exchange and it is a constituent of the FTSE 100 Index. Like all public companies, Marks & Spencer publishes an annual report. In the annual report, alongside many photographs of its ambassadors and models, there are pictures of a different nature: statistical charts illustrating in particular the financial performance of the company. The examples here are from Marks and Spencer's 2013 Annual Report. First is a chart showing Marks & Spencer's governance framework, then a bar chart showing the breakdown of Marks & Spencer's international revenue, and finally a line graph showing mystery shopper feedback.

We are exposed to statistical charts of this type almost daily: in newspapers and magazines, on TV, online and in business reports such as the Marks & Spencer Annual Report. In this chapter, you will learn about tabular and graphical methods of descriptive statistics such as frequency distributions, bar charts, histograms, stem-and-leaf displays, cross-tabulations and others. The goal of these methods is to summarize data so that they can be easily understood and interpreted.

A window display showing an array of personalities who have modelled for Marks & Spencer

Our Committees and Committee Chairmen

For more on our Governance framework go to **marksandspencer.com/the company**

International revenue

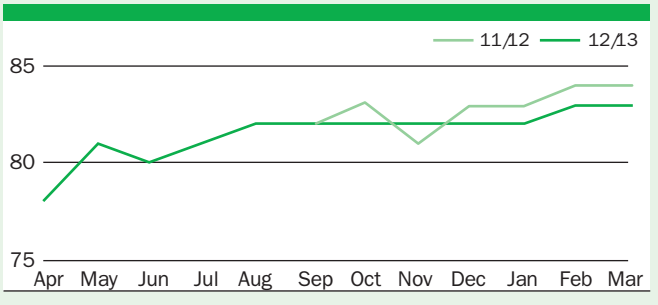
£1,075.4 m
↑4.5%

11/12	£1,066.1 m
10/11	£1,007.3 m
09/10	£968.7 m

Analysis We are continuing to transform M&S into a more internationally focused business and are making progress against our target of increasing international sales by £300 m to £500 m by 2013/14.

12/13	£1,075.4 m
11/12	£1,066.1 m
10/11	£1,077.3 m
09/10	£968.7 m

UK Mystery Shopping scores



Average score
81%

Analysis Mystery Shop scores remained high this year at 81%. However, to help us be more in touch with customers we plan to replace our monthly Mystery Shop programme with a more regular, in-depth customer satisfaction survey.

Annual space growth

2.8%

Analysis As consumer's shopping habits change, we continue to evolve our space selectively. We expect the planned opening of new space will add c.2% to the UK in 2013/14.

2.1 SUMMARIZING QUALITATIVE DATA

Frequency distribution

We begin with a definition.

Frequency distribution

A frequency distribution is a tabular summary of data showing the number (frequency) of items in each of several non-overlapping classes.

The following example demonstrates the construction and interpretation of a **frequency distribution** for qualitative data. Audi, BMW, Mercedes, Opel and VW are five popular brands of car in Germany. The data in Table 2.1 are for a sample of 50 new car purchases of these five brands.

To construct a frequency distribution, we count the number of times each brand appears in Table 2.1. VW appears 19 times, Mercedes appears 13 times and so on. These counts are summarized in the frequency distribution in Table 2.2. The summary offers more insight than the original data. We see that VW is the leader, Mercedes is second, Audi is third. Opel and BMW are tied for fourth.

TABLE 2.1 Data from a sample of 50 new car purchases

VW	BMW	Mercedes	Audi	VW
VW	Mercedes	Audi	VW	Audi
VW	VW	VW	Audi	Mercedes
VW	VW	Opel	Opel	BMW
VW	Audi	Mercedes	Audi	Mercedes
VW	Mercedes	Mercedes	VW	Mercedes
VW	VW	Mercedes	Opel	Mercedes
Mercedes	BMW	VW	VW	VW
BMW	Opel	Audi	Opel	Mercedes
VW	Mercedes	BMW	VW	Audi

TABLE 2.2 Frequency distribution of new car purchases

CAR BRANDS

Brand	Frequency
Audi	8
BMW	5
Mercedes	13
Opel	5
VW	19
Total	50

Relative frequency and percentage frequency distributions

A frequency distribution shows the number (frequency) of items in each of several non-overlapping classes. We are often interested in the proportion, or percentage, of items in each class. The *relative frequency* of a class is the fraction or proportion of items belonging to a class. For a data set with n observations, the relative frequency of each class is:

Relative frequency

$$\text{Relative frequency of a class} = \frac{\text{Frequency of the class}}{n} \qquad \textbf{(2.1)}$$

The *percentage frequency* of a class is the relative frequency multiplied by 100.

A **relative frequency distribution** is a tabular summary showing the relative frequency for each class. A **percentage frequency distribution** summarizes the percentage frequency for each class. Table 2.3 shows these distributions for the car purchase data. The relative frequency for VW is 19/50 = 0.38, the relative frequency for Mercedes is 13/50 = 0.26 and so on. From the percentage frequency distribution, we see that 38 per cent of the purchases were VW, 26 per cent were Mercedes and so on. We can also note, for example, that 38 + 26 = 64 per cent of the purchases were of the top two car brands.

TABLE 2.3 Relative and percentage frequency distributions of new car purchases

Brand	Relative frequency	Percentage frequency
Audi	0.16	16
BMW	0.10	10
Mercedes	0.26	26
Opel	0.10	10
VW	0.38	38
Total	1.00	100

Bar charts and pie charts

A **bar chart**, or **bar graph**, is a pictorial representation of a frequency, relative frequency, or percentage frequency distribution. On one axis of the chart (usually the horizontal), we specify the labels for the classes (categories) of data. A frequency, relative frequency or percentage frequency scale can be used for the other axis of the chart (usually the vertical). Then, using a bar of fixed width drawn above each class label, we make the length of the bar equal the frequency, relative frequency or percentage frequency of the class. For qualitative data, the bars should be separated to emphasize the fact that each class is separate. Figure 2.1 shows a bar chart of the frequency distribution for the 50 new car purchases.

FIGURE 2.1
Bar chart of new car purchases

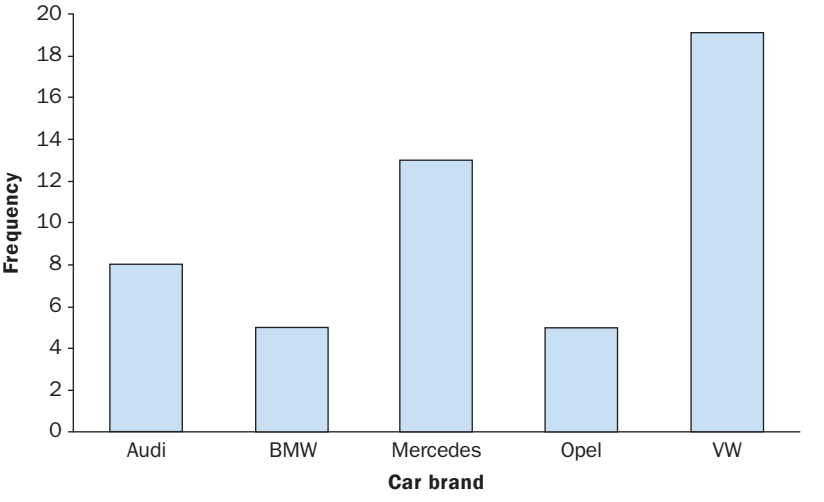
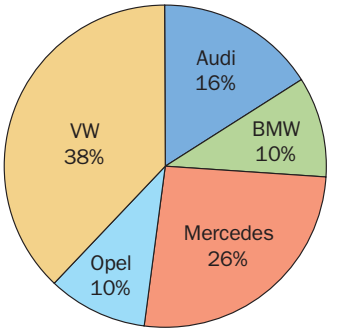


FIGURE 2.2
Pie chart of new car purchases



A **pie chart** is another way of presenting relative frequency and percentage frequency distributions. We first draw a circle to represent all the data. Then we subdivide the circle into sectors that correspond to the relative frequency for each class. For example, because a circle contains 360 degrees and VW shows a relative frequency of 0.38, the sector of the pie chart labelled VW consists of 0.38(360) = 136.8 degrees. The sector of the pie chart labelled Mercedes consists of 0.26(360) = 93.6 degrees. Similar calculations for the other classes give the pie chart in Figure 2.2. The numerical values shown for each sector can be frequencies, relative frequencies or percentage frequencies.

Often the number of classes in a frequency distribution is the same as the number of categories in the data, as for the car purchase data in this section. Data that included all car brands would require many categories, most of which would have a small number of purchases. Classes with smaller frequencies can be grouped into an aggregate class labelled 'other'. Classes with frequencies of 5 per cent or less would most often be treated in this way.

In quality control applications, bar charts are used to summarize the most important causes of problems. When the bars are arranged in descending order of height from left to right with the most frequently occurring cause appearing first, the bar chart is called a *Pareto diagram*, named after its founder, Vilfredo Pareto, an Italian economist.

EXERCISES

Methods

1. The response to a question has three alternatives: A, B and C. A sample of 120 responses provides 60 A, 24 B and 36 C. Construct the frequency and relative frequency distributions.

COMPLETE SOLUTIONS

2. A partial relative frequency distribution is given below.

Class	*Relative frequency*
A	0.22
B	0.18
C	0.40
D	

a. What is the relative frequency of class D?
b. The total sample size is 200. What is the frequency of class D?
c. Construct the frequency distribution.
d. Construct the percentage frequency distribution.

3. A questionnaire provides 58 Yes, 42 No and 20 No-opinion answers.

a. In the construction of a pie chart, how many degrees would be in the sector of the pie showing the Yes answers?

b. How many degrees would be in the sector of the pie showing the No answers?
c. Construct a pie chart.
d. Construct a bar chart.

Applications

4. CEM4Mobile is a customer experience management company based in Finland. The company does extensive market research in the mobile telecommunications field. Its research shows that the four most popular mobile operating systems in Nordic countries are Apple iOS, Symbian OS, Android and Nokia OS. A sample of 50 page loads from mobile browsing services follows.

Android	Android	Android	Symbian	Apple	Apple	Symbian	Apple	Apple	Android
Android	Symbian	Android	Apple	Nokia	Android	Apple	Apple	Apple	Nokia
Nokia	Apple	Symbian	Apple	Nokia	Symbian	Android	Nokia	Android	Apple
Android	Symbian	Symbian	Apple	Android	Android	Apple	Android	Android	Apple
Apple	Nokia	Symbian	Symbian	Android	Android	Apple	Symbian	Symbian	Android

a. Are these data qualitative or quantitative?
b. Construct frequency and percentage frequency distributions.
c. Construct a bar chart and a pie chart.
d. On the basis of the sample, which mobile operating system was the most popular? Which one was second?


NORDIC OS

5. A Wikipedia article listed the six most common last names in Belgium as (in alphabetical order): Jacobs, Janssens, Maes, Mertens, Peeters and Willems. A sample of 50 individuals with one of these last names provided the following data.

Peeters	Peeters	Willems	Janssens	Janssens	Peeters	Jacobs	Maes	Janssens	Mertens
Jacobs	Maes	Peeters	Willems	Jacobs	Maes	Peeters	Janssens	Maes	Maes
Peeters	Maes	Peeters	Maes	Janssens	Janssens	Mertens	Jacobs	Jacobs	Peeters
Mertens	Maes	Peeters	Janssens	Willems	Willems	Peeters	Janssens	Willems	Mertens
Jacobs	Willems	Peeters	Janssens	Mertens	Janssens	Peeters	Mertens	Mertens	Janssens


BELGIUM NAMES

Summarize the data by constructing the following:

a. Relative and percentage frequency distributions.
b. A bar chart.
c. A pie chart.
d. Based on these data, what are the three most common last names?

6. The flextime system at Electronics Associates allows employees to begin their working day at 7:00, 7:30, 8:00, 8:30 or 9:00 a.m. The following data represent a sample of the starting times selected by the employees.

7:00	8:30	9:00	8:00	7:30	7:30	8:30	8:30	7:30	7:00
8:30	8:30	8:00	8:00	7:30	8:30	7:00	9:00	8:30	8:00


COMPLETE SOLUTIONS

Summarize the data by constructing the following:

a. A frequency distribution.
b. A percentage frequency distribution.
c. A bar chart.
d. A pie chart.
e. What do the summaries tell you about employee preferences in the flextime system?

7. A Merrill Lynch Client Satisfaction Survey asked clients to indicate how satisfied they were with their financial consultant. Client responses were coded 1 to 7, with 1 indicating 'not at all satisfied' and

7 indicating 'extremely satisfied'. The following data are from a sample of 60 responses for a particular financial consultant.

5	7	6	6	7	5	5	7	3	6
7	7	6	6	6	5	5	6	7	7
6	6	4	4	7	6	7	6	7	6
5	7	5	7	6	4	7	5	7	6
6	5	3	7	7	6	6	6	6	5
5	6	6	7	7	5	6	4	6	6

a. Construct a frequency distribution and a relative frequency distribution for the data.
b. Construct a bar chart.
c. On the basis of your summaries, comment on the clients' overall evaluation of the financial consultant.

2.2 SUMMARIZING QUANTITATIVE DATA

Frequency distribution

As defined in Section 2.1, a frequency distribution is a tabular summary of data showing the number (frequency) of items in each of several non-overlapping classes. This definition holds for quantitative as well as qualitative data. However, with quantitative data there is usually more work involved in defining the non-overlapping classes.

Consider the quantitative data in Table 2.4. These data show the time in days required to complete year-end audits for a sample of 20 clients of Sanderson and Clifford, a small accounting firm. The data are rounded to the nearest day. The three steps necessary to define the classes for a frequency distribution with quantitative data are:

1 Determine the number of non-overlapping classes.
2 Determine the width of each class.
3 Determine the class limits.

We demonstrate these steps using the audit time data in Table 2.4.

Number of classes

Classes are formed by specifying ranges that will be used to group the data. As a general guideline, we recommend using between 5 and 20 classes. For a small sample of data, as few as five or six classes may be used to summarize the data. For larger samples, more classes are usually required. The aim is to use enough classes to show the pattern of variation in the data, but not so many classes that some contain very few data points. Because the sample in Table 2.4 is relatively small ($n = 20$), we chose to construct a frequency distribution with five classes.


AUDIT

TABLE 2.4 Year-end audit times (in days)

12	14	19	18	15	15	18	17	20	27
22	23	22	21	33	28	14	18	16	13

Width of the classes

The second step is to choose a width for the classes. As a general guideline, we recommend using the same width for each class. This reduces the chance of inappropriate interpretations. The choices for the number and the width of classes are not independent decisions. More classes means a smaller class width and vice versa. To determine an approximate class width, we identify the largest and smallest data values. Then we can use the following expression to determine the approximate class width.

Approximate class width

$$\frac{\text{Largest data value} - \text{Smallest data value}}{\text{Number of classes}} \qquad \textbf{(2.2)}$$

The approximate width given by equation (2.2) can be rounded to a more convenient value. For example, an approximate class width of 9.28 might be rounded to 10.

For the year-end audit times, the largest value is 33 and the smallest value is 12. We decided to summarize the data with five classes, so equation (2.2) provides an approximate class width of $(33 - 12)/5 = 4.2$. We decided to round up and use a class width of five days.

In practice, the number of classes and the appropriate class width are determined by trial and error. Once a possible number of classes is chosen, equation (2.2) is used to find the approximate class width. The process can be repeated for a different number of classes. Ultimately, the analyst uses judgement to determine the number of classes and class width that provide a good summary of the data. Different people may construct different, but equally acceptable, frequency distributions. The goal is to reveal the natural grouping and variation in the data.

For the audit time data, after deciding to use five classes, each with a width of five days, the next task is to specify the class limits for each of the classes.

Class limits

Class limits must be chosen so that each data item belongs to one and only one class. The *lower class limit* identifies the smallest possible data value assigned to the class. The *upper class limit* identifies the largest possible data value assigned to the class. In constructing frequency distributions for qualitative data, we did not need to specify class limits because each data item naturally fell into a separate class (category). But with quantitative data, class limits are necessary to determine where each data value belongs.

Using the audit time data, we selected ten days as the lower class limit and 14 days as the upper class limit for the first class. This class is denoted 10–14 in Table 2.5. The smallest data value, 12, is included in the 10–14 class. We then selected 15 days as the lower class limit and 19 days as the upper class limit of the next class. We continued defining the lower and upper class limits to obtain a total of five classes: 10–14, 15–19, 20–24, 25–29 and 30–34. The largest data value, 33, is included in the 30–34 class. The difference between the lower class limits of adjacent classes is the class width. Using the first two lower class limits of 10 and 15, we see that the class width is $15 - 10 = 5$.

A frequency distribution can now be constructed by counting the number of data values belonging to each class. For example, the data in Table 2.5 show that four values (12, 14, 14 and 13) belong to the 10–14 class. The frequency for the 10–14 class is 4. Continuing this counting process for the 15–19, 20–24, 25–29 and 30–34 classes provides the frequency distribution in Table 2.5. Using this frequency distribution, we can observe that:

1. The most frequently occurring audit times are in the class 15–19 days. Eight of the 20 audit times belong to this class.
2. Only one audit required 30 or more days.

Other comments are possible, depending on the interests of the person viewing the frequency distribution. The value of a frequency distribution is that it provides insights about the data not easily obtained from the data in their original unorganized form.

TABLE 2.5 Frequency distribution for the audit time data

Audit time (days)	Frequency
10–14	4
15–19	8
20–24	5
25–29	2
30–34	1
Total	20

The appropriate values for the class limits with quantitative data depend on the level of accuracy of the data. For instance, with the audit time data, the limits used were integer values because the data were rounded to the nearest day. If the data were rounded to the nearest one-tenth of a day (e.g. 12.3, 14.4), the limits would be stated in tenths of days. For example, the first class would be 10.0–14.9. If the data were rounded to the nearest one-hundredth of a day (e.g. 12.34, 14.45), the limits would be stated in hundredths of days, e.g. the first class would be 10.00–14.99.

An *open-ended* class requires only a lower class limit or an upper class limit. For example, in the audit time data, suppose two of the audits had taken 58 and 65 days. Rather than continuing with classes 35–39, 40–44, 45–49 and so on, we could simplify the frequency distribution to show an open-ended class of '35 or more'. This class would have a frequency count of 2. Most often the open-ended class appears at the upper end of the distribution. Sometimes an open-ended class appears at the lower end of the distribution and occasionally such classes appear at both ends.

Class midpoint

In some applications, we want to know the midpoints of the classes in a frequency distribution for quantitative data. The **class midpoint** is the value halfway between the lower and upper class limits. For the audit time data, the five class midpoints are 12, 17, 22, 27 and 32.

Relative frequency and percentage frequency distributions

We define the relative frequency and percentage frequency distributions for quantitative data in the same way as for qualitative data. The relative frequency is simply the proportion of the observations belonging to a class. With n observations,

$$\text{Relative frequency of a class} = \frac{\text{Frequency of the class}}{n}$$

The percentage frequency of a class is the relative frequency multiplied by 100.

Based on the class frequencies in Table 2.5 and with $n = 20$, Table 2.6 shows the relative frequency and percentage frequency distributions for the audit time data. Note that 0.40 of the audits, or 40 per cent, required from 15 to 19 days. Only 0.05 of the audits, or 5 per cent, required 30 or more days. Again, additional interpretations and insights can be obtained by using Table 2.6.

TABLE 2.6 Relative and percentage frequency distributions for the audit time data

Audit time (days)	Relative frequency	Percentage frequency
10–14	0.20	20
15–19	0.40	40
20–24	0.25	25
25–29	0.10	10
30–34	0.05	5
Total	1.00	100

FIGURE 2.3
Dot plot for the audit time data

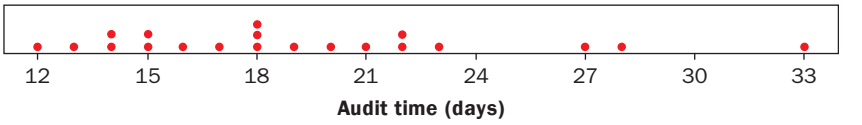
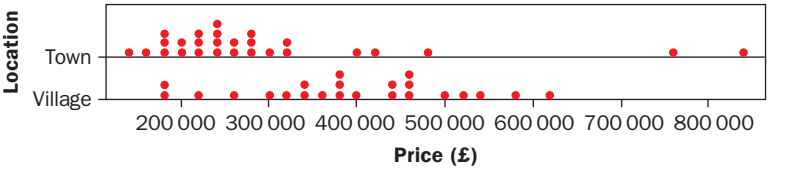


FIGURE 2.4
Dot plot comparing selling prices for houses in town and village locations



Dot plot

One of the simplest graphical summaries of data is a **dot plot**. A horizontal axis shows the range of values for the observations. Each data value is represented by a dot placed above the axis. Figure 2.3 is a dot plot produced in MINITAB for the audit time data in Table 2.4. The three dots located above 18 on the horizontal axis indicate that three audit times of 18 days occurred.

Dot plots show the details of the data and are useful for comparing data distributions for two or more samples. For example, Figure 2.4 shows a MINITAB dot plot comparing the selling prices of houses for two samples of houses: one in town locations and the other in village locations.

Histogram

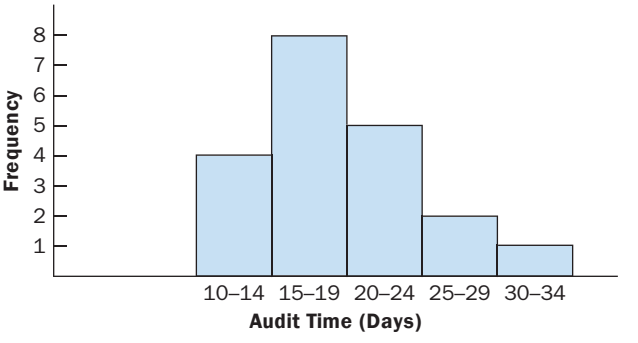
A **histogram** is a chart showing quantitative data previously summarized in a frequency, relative frequency or percentage frequency distribution. The variable of interest is placed on the horizontal axis and the frequency, relative frequency or percentage frequency on the vertical axis. The frequency, relative frequency or percentage frequency of each class is shown by drawing a rectangle whose base is determined by the class limits on the horizontal axis and whose height is the corresponding frequency, relative frequency or percentage frequency.

Figure 2.5 is a histogram for the audit time data. The class with the greatest frequency is shown by the rectangle above the class 15–19 days. The height of the rectangle shows that the frequency of this class is 8. A histogram for the relative or percentage frequency distribution of this data would look the same as the histogram in Figure 2.5 except that the vertical axis would be labelled with relative or percentage frequency values.

As Figure 2.5 shows, the adjacent rectangles of a histogram touch one another. This is the usual convention for a histogram, unlike a bar chart. Because the classes for the audit time data are stated as 10–14, 15–19, 20–24 and so on, one-unit spaces of 14 to 15, 19 to 20, etc. would seem to be needed between the classes. Eliminating the spaces in the histogram for the audit-time data helps show that, even though the data are rounded to the nearest full day, all values between the lower limit of the first class and the upper limit of the last class are possible.

One of the most important uses of a histogram is to provide information about the shape, or form, of a distribution. Figure 2.6 contains four histograms constructed from relative frequency distributions. Panel A shows the histogram for a set of data moderately skewed to the left. A histogram is skewed to the left, or

FIGURE 2.5
Histogram for the audit time data



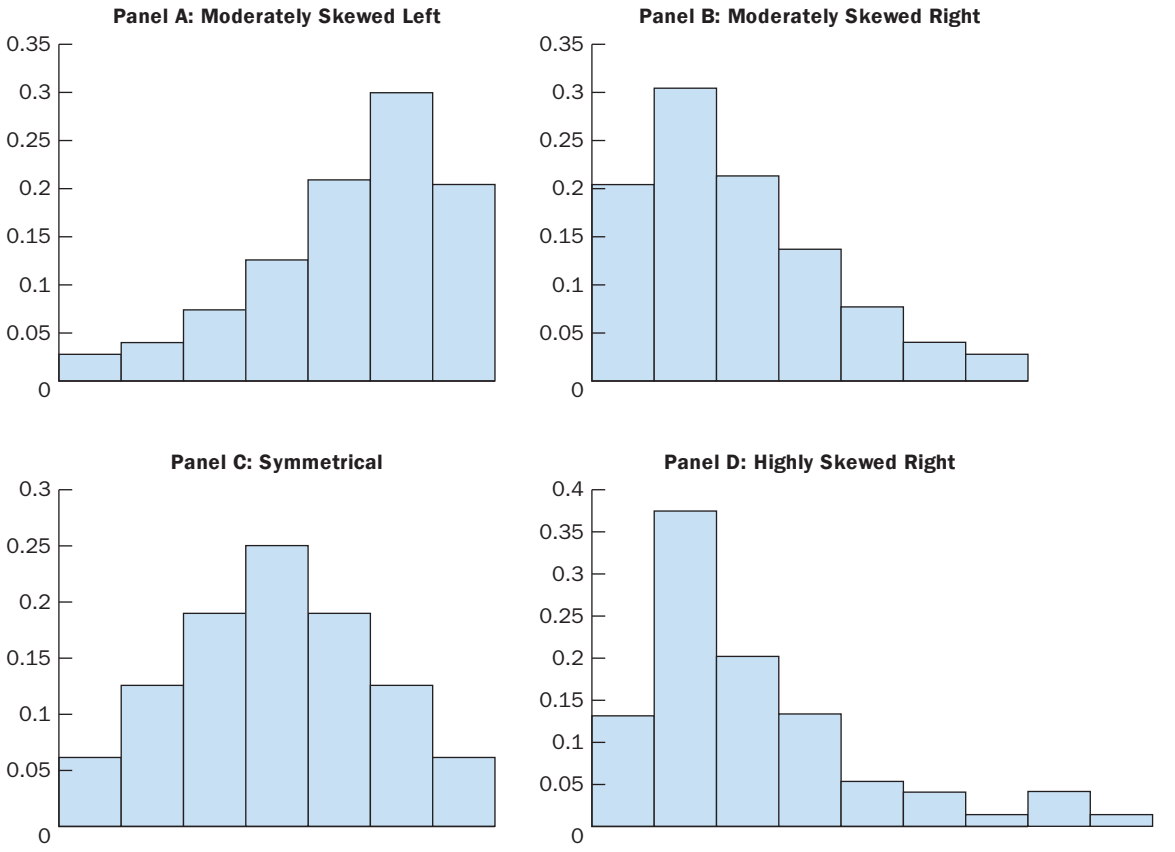


FIGURE 2.6
Histograms showing differing levels of skewness

negatively skewed, if its tail extends further to the left. A histogram like this might be seen for scores from a relatively simple test. There are no scores above 100 per cent, most of the scores are above 70 per cent and only a few really low scores occur. Panel B shows the histogram for a set of data moderately skewed to the right. A histogram is skewed to the right, or positively skewed, if its tail extends further to the right. An example of this type of histogram would be for data such as house values. A relatively small number of expensive homes create the skewness in the right tail.

Panel C shows a symmetrical histogram. In a symmetrical histogram, the left tail mirrors the shape of the right tail. Histograms for real data are never perfectly symmetrical, but for many applications may be roughly symmetrical. Data for IQ scores, heights and weights of people and so on, lead to histograms that are roughly symmetrical. Panel D shows a histogram highly skewed to the right (positively skewed). This histogram was constructed from data on the amount of customer purchases over one day at a women's clothing store. Data from applications in business and economics often lead to histograms that are skewed to the right: for instance, data on wealth, salaries, purchase amounts and so on.

Cumulative distributions

A variation of the frequency distribution that provides another tabular summary of quantitative data is the **cumulative frequency distribution**. The cumulative frequency distribution uses the number of classes, class widths and class limits adopted for the frequency distribution. However, rather than showing the frequency of each class, the cumulative frequency distribution shows the number of data items with values *less than or equal to the upper class limit* of each class. The first two columns of Table 2.7 show the cumulative frequency distribution for the audit time data.

TABLE 2.7 Cumulative frequency, cumulative relative frequency and cumulative percentage frequency distributions for the audit time data

Audit time (days)	Cumulative frequency	Cumulative relative frequency	Cumulative percentage frequency
Less than or equal to 14	4	0.20	20
Less than or equal to 19	12	0.60	60
Less than or equal to 24	17	0.85	85
Less than or equal to 29	19	0.95	95
Less than or equal to 34	20	1.00	100

Consider the class with the description 'less than or equal to 24'. The cumulative frequency for this class is simply the sum of the frequencies for all classes with data values less than or equal to 24. For the frequency distribution in Table 2.5, the sum of the frequencies for classes 10–14, 15–19 and 20–24 indicates that 4 + 8 + 5 = 17 data values are less than or equal to 24. The cumulative frequency distribution in Table 2.7 also shows that four audits were completed in 14 days or less and 19 audits were completed in 29 days or less.

A **cumulative relative frequency distribution** shows the proportion of data items and a **cumulative percentage frequency distribution** shows the percentage of data items with values less than or equal to the upper limit of each class. The cumulative relative frequency distribution can be computed either by summing the relative frequencies in the relative frequency distribution, or by dividing the cumulative frequencies by the total number of items. Using the latter approach, we found the cumulative relative frequencies in column 3 of Table 2.7 by dividing the cumulative frequencies in column 2 by the total number of items (n = 20). The cumulative percentage frequencies were computed by multiplying the cumulative relative frequencies by 100.

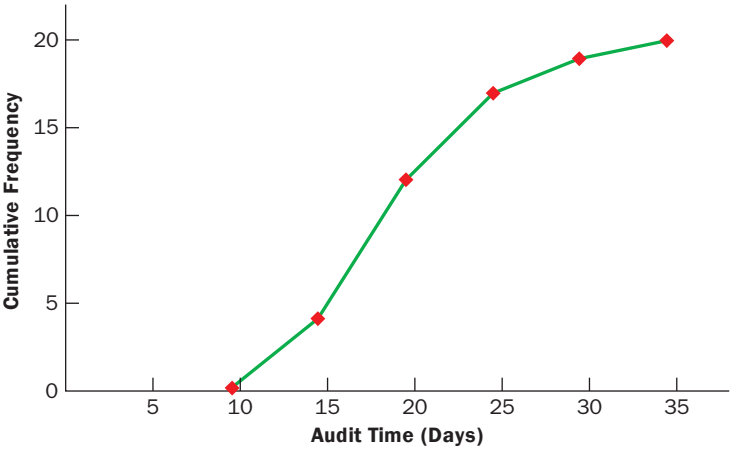
The cumulative relative and percentage frequency distributions show that 0.85 of the audits, or 85 per cent, were completed in 24 days or less; 0.95 of the audits, or 95 per cent, were completed in 29 days or less and so on.

The last entry in a cumulative frequency distribution always equals the total number of observations. The last entry in a cumulative relative frequency distribution always equals 1.00 and the last entry in a cumulative percentage frequency distribution always equals 100.

Cumulative distribution plot (ogive)

A graph of a cumulative distribution, called an **ogive**, shows data values on the horizontal axis and either the cumulative frequencies, the cumulative relative frequencies, or the cumulative percentage frequencies on the vertical axis. Figure 2.7 illustrates a cumulative distribution plot or ogive for the cumulative frequencies of the audit time data.

FIGURE 2.7
Ogive for the audit time data



The ogive is constructed by plotting a point corresponding to the cumulative frequency of each class. Because the classes for the audit time data are 10–14, 15–19, 20–24 and so on, one-unit gaps appear from 14 to 15, 19 to 20 and so on. These gaps are eliminated by plotting points halfway between the class limits. So, 14.5 is used for the 10–14 class, 19.5 is used for the 15–19 class and so on. The 'less than or equal to 14' class with a cumulative frequency of four is shown on the ogive in Figure 2.7 by the point located at 14.5 on the horizontal axis and 4 on the vertical axis. The 'less than or equal to 19' class with a cumulative frequency of 12 is shown by the point located at 19.5 on the horizontal axis and 12 on the vertical axis. Note that one additional point is plotted at the left end of the ogive. This point starts the ogive by showing that no data values fall below the 10–14 class. It is plotted at 9.5 on the horizontal axis and 0 on the vertical axis. The plotted points are connected by straight lines to complete the ogive.

Exploratory data analysis: stem-and-leaf display

Exploratory data analysis techniques consist of simple arithmetic and easy-to-draw graphs that can be used to summarize data quickly. One technique – referred to as a **stem-and-leaf display** – can be used to show both the rank order and shape of a data set simultaneously. To illustrate the stem-and-leaf display, consider the data in Table 2.8. These came from a 150-question aptitude test given to 50 individuals recently interviewed for a position at Hawkins Manufacturing. The data indicate the number of questions answered correctly.

To construct a stem-and-leaf display, we first arrange the leading digits of each data value to the left of a vertical line. To the right of the vertical line, on the line corresponding to the appropriate first digit, we record the last digit for each data value as we pass through the observations in the order they were recorded.

6	9 8
7	2 3 6 3 6 5
8	6 2 3 1 1 0 4 5
9	7 2 2 6 2 1 5 8 8 5 4
10	7 4 8 0 2 6 6 0 6
11	2 8 5 9 3 5 9
12	6 8 7 4
13	2 4
14	1

Sorting the digits on each line into rank order is now relatively simple. This leads to the stem-and-leaf display shown here.

6	8 9
7	2 3 3 5 6 6
8	0 1 1 2 3 4 5 6
9	1 2 2 2 4 5 5 6 7 8 8
10	0 0 2 4 6 6 6 7 8
11	2 3 5 5 8 9 9
12	4 6 7 8
13	2 4
14	1

The numbers to the left of the vertical line (6, 7, …, 14) form the *stem*, and each digit to the right of the vertical line is a *leaf*. For example, the first row has a stem value of 6 and leaves of 8 and 9. It indicates that two data values have a first digit of six. The leaves show that the data values are 68 and 69. Similarly, the second row indicates that six data values have a first digit of 7. The leaves show that the data values are 72, 73, 73, 75, 76 and 76. Rotating the page counter-clockwise onto its side provides a picture of the data that is similar to a histogram with classes of 60–69, 70–79, 80–89 and so on.

TABLE 2.8 Number of questions answered correctly on an aptitude test

112	72	69	97	107	73	92	76	86	73
126	128	118	127	124	82	104	132	134	83
92	108	96	100	92	115	76	91	102	81
95	141	81	80	106	84	119	113	98	75
68	98	115	106	95	100	85	94	106	119

Although the stem-and-leaf display may appear to offer the same information as a histogram, it has two primary advantages.

1. The stem-and-leaf display is easier to construct by hand for small data sets.
2. Within a class interval, the stem-and-leaf display provides more information than the histogram because the stem-and-leaf shows the actual data.

Just as a frequency distribution or histogram has no absolute number of classes, neither does a stem-and-leaf display have an absolute number of rows or stems. If we believe that our original stem-and-leaf display condensed the data too much, we can stretch the display by using two stems for each leading digit (using five stems for each leading digit is also a possibility). Using two stems for each leading digit, we would place all data values ending in 0, 1, 2, 3 and 4 in one row and all values ending in 5, 6, 7, 8 and 9 in a second row. The following display illustrates this approach. This stretched stem-and-leaf display is similar to a frequency distribution with intervals of 65–69, 70–74, 75–79 and so on.

6	8 9
7	2 3 3
7	5 6 6
8	0 1 1 2 3 4
8	5 6
9	1 2 2 2 4
9	5 5 6 7 8 8
10	0 0 2 4
10	6 6 6 7 8
11	2 3
11	5 5 8 9 9
12	4
12	6 7 8
13	2 4
14	1

The preceding example shows a stem-and-leaf display for data with three digits. Stem-and-leaf displays for data with more than three digits are possible. For example, consider the following data on the number of burgers sold by a fast-food restaurant for each of 15 weeks.

1565	1852	1644	1766	1888	1912	2044	1812
1790	1679	2008	1852	1967	1954	1733	

A stem-and-leaf display of these data follows.

Leaf unit = 10

15	6
16	4 7
17	3 6 9
18	1 5 5 8
19	1 5 6
20	0 4

A single digit is used to define each leaf, and only the first three digits of each observation have been used to construct the display. At the top of the display we have specified leaf unit = 10. Consider the first stem (15) and its associated leaf (6). Combining these numbers gives 156. To reconstruct an approximation of the original data value, we must multiply this number by 10, the value of the *leaf unit*: $156 \times 10 = 1560$. Although it is not possible to reconstruct the exact data value from the display, using a single digit for each leaf enables stem-and-leaf displays to be constructed for data having a large number of digits. Leaf units may be 100, 10, 1, 0.1 and so on. Where the leaf unit is not shown on the display, it is assumed to equal 1.

EXERCISES

Methods


FREQUENCY

8. Consider the following data.

14	21	23	21	16	19	22	25	16	16
24	24	25	19	16	19	18	19	21	12
16	17	18	23	25	20	23	16	20	19
24	26	15	22	24	20	22	24	22	20

a. Construct a frequency distribution using classes of 12–14, 15–17, 18–20, 21–23 and 24–26.
b. Construct a relative frequency distribution and a percentage frequency distribution using the classes in (a).


COMPLETE SOLUTIONS

9. Consider the following frequency distribution. Construct a cumulative frequency distribution and a cumulative relative frequency distribution.

Class	*Frequency*
10–19	10
20–29	14
30–39	17
40–49	7
50–59	2

10. Construct a histogram and an ogive for the data in Exercise 9.

11. Consider the following data.

8.9	10.2	11.5	7.8	10.0	12.2	13.5	14.1	10.0	12.2
6.8	9.5	11.5	11.2	14.9	7.5	10.0	6.0	15.8	11.5

a. Construct a dot plot.
b. Construct a frequency distribution.
c. Construct a percentage frequency distribution.


COMPLETE SOLUTIONS

12. Construct a stem-and-leaf display for the following data.

70	72	75	64	58	83	80	82	76	75	68	65	57	78	85	72

13. Construct a stem-and-leaf display for the following data.

11.3	9.6	10.4	7.5	8.3	10.5	10.0	9.3	8.1	7.7	7.5	8.4	6.3	8.8

Applications

14. A doctor's office staff studied the waiting times for patients who arrive at the office with a request for emergency service. The following data with waiting times in minutes were collected over a one-month period.

2 5 10 12 4 4 5 17 11 8 9 8 12 21 6 8 7 13 18 3

Use classes of 0–4, 5–9 and so on in the following:

a. Show the frequency distribution.
b. Show the relative frequency distribution.
c. Show the cumulative frequency distribution.
d. Show the cumulative relative frequency distribution.
e. What proportion of these patients wait nine minutes or less?

15. Data for the numbers of units produced by a production employee during the most recent 20 days are shown here.

160	170	181	156	176	148	198	179	162	150
162	156	179	178	151	157	154	179	148	156

Summarize the data by constructing the following:

a. A frequency distribution.
b. A relative frequency distribution.
c. A cumulative frequency distribution.
d. A cumulative relative frequency distribution.
e. A cumulative distribution plot (ogive).

16. The closing prices of 40 company shares (in Kuwaiti dinar) follow.

44.00	0.80	69.00	226.00	68.00	51.00	265.00	130.00
172.00	202.00	52.00	134.00	81.00	50.00	550.00	28.50
13.00	435.00	218.00	270.00	52.00	108.00	248.00	0.45
188.00	800.00	59.00	65.00	355.00	410.00	102.00	174.00
136.00	34.00	64.00	660.00	122.00	62.00	290.00	90.00


SHARES

a. Construct frequency and relative frequency distributions.
b. Construct cumulative frequency and cumulative relative frequency distributions.
c. Construct a histogram.
d. Using your summaries, make comments and observations about the price of shares.

17. The table below shows the estimated 2013 mid-year population of Kenya, by age group, rounded to the nearest thousand (from the US Census Bureau International Data Base).

Age group	*Population (000s)*
0–9	13 310
10–19	9 601
20–29	7 904
30–39	5 975
40–49	3 273
50–59	2 076
60–69	1 171
70–79	555
80+	171

a. Construct a percentage frequency distribution.
b. Construct a cumulative percentage frequency distribution.

c. Construct a cumulative distribution plot (ogive).
d. Using the ogive, estimate the age that divides the population into halves (you will learn in Chapter 3 that this is called the *median*).

18. The *Nielsen Home Technology Report* provided information about home technology and its usage by individuals aged 12 and older. The following data are the hours of personal computer usage during one week for a sample of 50 individuals.


COMPUTER

4.1	1.5	5.9	3.4	5.7	1.6	6.1	3.0	3.7	3.1	4.8	2.0	3.3
11.1	3.5	4.1	4.1	8.8	5.6	4.3	7.1	10.3	6.2	7.6	10.8	0.7
4.0	9.2	4.4	5.7	7.2	6.1	5.7	5.9	4.7	3.9	3.7	3.1	12.1
14.8	5.4	4.2	3.9	4.1	2.8	9.5	12.9	6.1	3.1	10.4		

Summarize the data by constructing the following:


COMPLETE SOLUTIONS

a. A frequency distribution (use a class width of three hours).
b. A relative frequency distribution.
c. A histogram.
d. A cumulative distribution plot (ogive).
e. Comment on what the data indicate about personal computer usage at home.

19. The daily high and low temperatures (in degrees Celsius) for 20 cities on one particular day follow.

City	*High*	*Low*	*City*	*High*	*Low*
Athens	24	12	Melbourne	19	10
Bangkok	33	23	Montreal	18	11
Cairo	29	14	Paris	25	13
Copenhagen	18	4	Rio de Janeiro	27	16
Dublin	18	8	Rome	27	12
Havana	30	20	Seoul	18	10
Hong Kong	27	22	Singapore	32	24
Johannesburg	16	10	Sydney	20	13
London	23	9	Tokyo	26	15
Manila	34	24	Vancouver	14	6

a. Prepare a stem-and-leaf display for the high temperatures.
b. Prepare a stem-and-leaf display for the low temperatures.
c. Compare the stem-and-leaf displays from parts (a) and (b), and comment on the differences between daily high and low temperatures.
d. Use the stem-and-leaf display from part (a) to determine the number of cities having a high temperature of 25 degrees or above.
e. Provide frequency distributions for both high and low temperature data.

2.3 CROSS-TABULATIONS AND SCATTER DIAGRAMS

So far in this chapter, we have focused on methods for summarizing *one variable at a time.* Often a manager or decision-maker requires tabular and graphical methods that will assist in the understanding of the *relationship between two variables.* Cross-tabulation and scatter diagrams are two such methods.

Cross-tabulation

A **cross-tabulation** is a tabular summary of data for two variables. Consider the following data from a consumer restaurant review, based on a sample of 300 restaurants in a large European city. Table 2.9 shows the data for the first five restaurants: the restaurant's quality rating and typical meal price. Quality

TABLE 2.9 Quality rating and meal price for 300 restaurants

Restaurant	Quality rating	Meal price (€)
1	Disappointing	18
2	Good	22
3	Disappointing	28
4	Excellent	38
5	Good	33
.	.	.
.	.	.
.	.	.

RESTAURANT

TABLE 2.10 Cross-tabulation of quality rating and meal price for 300 restaurants

	Meal price				
Quality rating	**€10–19**	**€20–29**	**€30–39**	**€40–49**	**Total**
Disappointing	42	40	2	0	84
Good	34	64	46	6	150
Excellent	2	14	28	22	66
Total	78	118	76	28	300

rating is a qualitative variable with categories 'disappointing', 'good' and 'excellent'. Meal price is a quantitative variable that ranges from €10 to €49.

A cross-tabulation of the data is shown in Table 2.10. The left and top margin labels define the classes for the two variables. In the left margin, the row labels (disappointing, good and excellent) correspond to the three classes of the quality rating variable. In the top margin, the column labels (€10–19, €20–29, €30–39 and €40–49) correspond to the four classes of the meal price variable. Each restaurant in the sample provides a quality rating and a meal price, and so is associated with a cell in one of the rows and one of the columns of the cross-tabulation. For example, restaurant 5 has a good quality rating and a meal price of €33. This restaurant belongs to the cell in row 2 and column 3 of Table 2.10. In constructing a cross-tabulation, we simply count the number of restaurants that belong to each of the cells in the cross-tabulation.

We see that the greatest number of restaurants in the sample (64) have a good rating and a meal price in the €20–29 range. Only two restaurants have an excellent rating and a meal price in the €10–19 range. In addition, note that the right and bottom margins of the cross-tabulation provide the frequency distributions for quality rating and meal price separately. From the frequency distribution in the right margin, we see the quality rating data showing 84 disappointing restaurants, 150 good restaurants and 66 excellent restaurants.

Dividing the totals in the right margin by the total for that column provides relative and percentage frequency distributions for the quality rating variable.

Quality rating	*Relative frequency*	*Percentage frequency*
Disappointing	0.28	28
Good	0.50	50
Excellent	0.22	22
Total	1.00	100

We see that 28 per cent of the restaurants were rated disappointing, 50 per cent were rated good and 22 per cent were rated excellent.

Dividing the totals in the bottom row of the cross-tabulation by the total for that row provides relative and percentage frequency distributions for the meal price variable. In this case the values do not add exactly to 100, because the values being summed are rounded. From the percentage frequency distribution we quickly see that 26 per cent of the meal prices are in the lowest price class (€10–19), 39 per cent are in the next higher class and so on.

Meal price	*Relative frequency*	*Percentage frequency*
€10–19	0.26	26
€20–29	0.39	39
€30–39	0.25	25
€40–49	0.09	9
Total	1.00	100

The frequency and relative frequency distributions constructed from the margins of a cross-tabulation provide information about each of the variables individually, but they do not shed any light on the relationship between the variables. The primary value of a cross-tabulation lies in the insight it offers about this relationship. Converting the entries in a cross-tabulation into row percentages or column percentages can provide the insight.

For row percentages, the results of dividing each frequency in Table 2.10 by its corresponding row total are shown in Table 2.11. Each row of Table 2.11 is a percentage frequency distribution of meal price for one of the quality rating categories. Of the restaurants with the lowest quality rating (disappointing), we see that the greatest percentages are for the less expensive restaurants (50.0 per cent have €10–19 meal prices and 47.6 per cent have €20–29 meal prices). Of the restaurants with the highest quality rating (excellent), we see that the greatest percentages are for the more expensive restaurants (42.4 per cent have €30–39 meal prices and 33.4 per cent have €40–49 meal prices). Hence, the cross-tabulation reveals that higher meal prices are associated with the higher quality restaurants, and the lower meal prices are associated with the lower quality restaurants.

Cross-tabulation is widely used for examining the relationship between two variables. The final reports for many statistical studies include a large number of cross-tabulations. In the restaurant survey, the cross-tabulation is based on one qualitative variable (quality rating) and one quantitative variable (meal price). Cross-tabulations can also be constructed when both variables are qualitative and when both variables are quantitative. When quantitative variables are used, we must first create classes for the values of the variable. For instance, in the restaurant example we grouped the meal prices into four classes (€10–19, €20–29, €30–39 and €40–49).

Simpson's paradox

In many cases, a summary cross-tabulation showing how two variables are related has in effect been aggregated across a third variable (or across more than one variable). If so, we must be careful in drawing conclusions about the relationship between the two variables in the aggregated cross-tabulation. In some cases the conclusions based upon the aggregated cross-tabulation can be completely reversed if we look at the non-aggregated data, something known as **Simpson's paradox**. To provide an illustration, we consider an example involving the analysis of sales success for two sales executives in a mobile telephone company.

TABLE 2.11 Row percentages for each quality rating category

	Meal Price				
Quality rating	**€10–19**	**€20–29**	**€30–39**	**€40–49**	**Total**
Disappointing	50.0	47.6	2.4	0.0	100
Good	22.7	42.7	30.6	4.0	100
Excellent	3.0	21.2	42.4	33.4	100

The two sales executives are Aaron and Theo. They handle enquiries for renewal of two types of mobile telephone agreement: pre-pay contracts and pay-as-you-go (PAYG) agreements. The cross-tabulation below shows the outcomes for 200 enquiries each for Aaron and Theo, aggregated across the two types of agreement. The cross-tabulation involves two variables: outcome (sale or no sale) and sales executive (Aaron or Theo). It shows the number of sales and the number of no-sales for each executive, along with the column percentages in parentheses next to each value.

	Sales executive		
	Aaron	*Theo*	*Total*
Sales	82 (41%)	102 (51%)	184
No-sales	118 (59%)	98 (49%)	216
Total	200 (100%)	200 (100%)	400

The column percentages indicate that Aaron's overall sales success rate was 41 per cent, compared with Theo's 51 per cent success rate, suggesting that Theo has the better sales performance. A problem arises with this conclusion, however. The following cross-tabulations show the enquiries handled by Aaron and Theo for the two types of agreement separately.

	Pre-pay				*PAYG*		
	Aaron	*Theo*	*Total*		*Aaron*	*Theo*	*Total*
Sales	56 (35%)	18 (30%)	74	*Sales*	26 (65%)	84 (60%)	110
No-sales	104 (65%)	42 (70%)	146	*No-sales*	14 (35%)	56 (40%)	70
Total	160 (100%)	60 (100%)	220	*Total*	40 (100%)	140 (100%)	180

We see that Aaron achieved a 35 per cent success rate for pre-pay contracts and 65 per cent for PAYG agreements. Theo had a 30 per cent success rate for pre-pay and 60 per cent for PAYG. This comparison suggests that Aaron has a better success rate than Theo for both types of agreement, a result that contradicts the conclusion reached when the data were aggregated across the two types of agreement. This example illustrates Simpson's paradox.

Note that for both sales executives the sales success rate was much higher for PAYG than for pre-pay contracts. Because Theo handled a much higher proportion of PAYG enquiries than Aaron, the aggregated data favoured Theo. When we look at the cross-tabulations for the two types of agreement separately, however, Aaron shows the better record. Hence, for the original cross-tabulation, we see that the *type of agreement* is a hidden variable that should not be ignored when evaluating the records of the sales executives.

Because of Simpson's paradox, we need to be especially careful when drawing conclusions using aggregated data. Before drawing any conclusions about the relationship between two variables shown for a cross-tabulation – or, indeed, any type of display involving two variables (like the scatter diagram illustrated in the next section) – you should consider whether any hidden variable or variables could affect the results.

Scatter diagram and trend line

A **scatter diagram** is a graphical presentation of the relationship between two quantitative variables, and a **trend line** is a line that provides an approximation of the relationship. Consider the advertising/sales relationship for a hi-fi equipment store. On ten occasions during the past three months, the store used weekend television commercials to promote sales at its stores. The managers want to investigate whether a relationship exists between the number of commercials shown and sales at the store during the following week. Sample data for the ten weeks with sales in thousands of euros (€000s) are shown in Table 2.12.

TABLE 2.12 Sample data for the hi-fi equipment store

Week	Number of commercials	Sales in €000s
1	2	50
2	5	57
3	1	41
4	3	54
5	4	54
6	1	38
7	5	63
8	3	48
9	4	59
10	2	46

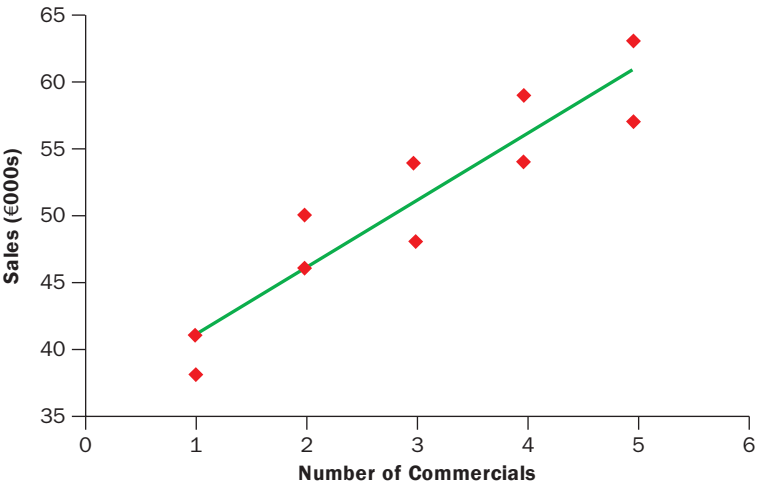

STEREO

Figure 2.8 shows the scatter diagram and the trend line* for the data in Table 2.12. The number of commercials (x) is shown on the horizontal axis and the sales (y) are shown on the vertical axis. For week 1, $x = 2$ and $y = 50$. A point with those coordinates is plotted on the scatter diagram. Similar points are plotted for the other nine weeks. Note that during two of the weeks one commercial was shown, during two of the weeks two commercials were shown and so on.

The completed scatter diagram in Figure 2.8 indicates a positive relationship between the number of commercials and sales. Higher sales are associated with a higher number of commercials. The relationship is not perfect in that all points are not on a straight line. However, the general pattern of the points and the trend line suggest that the overall relationship is positive.

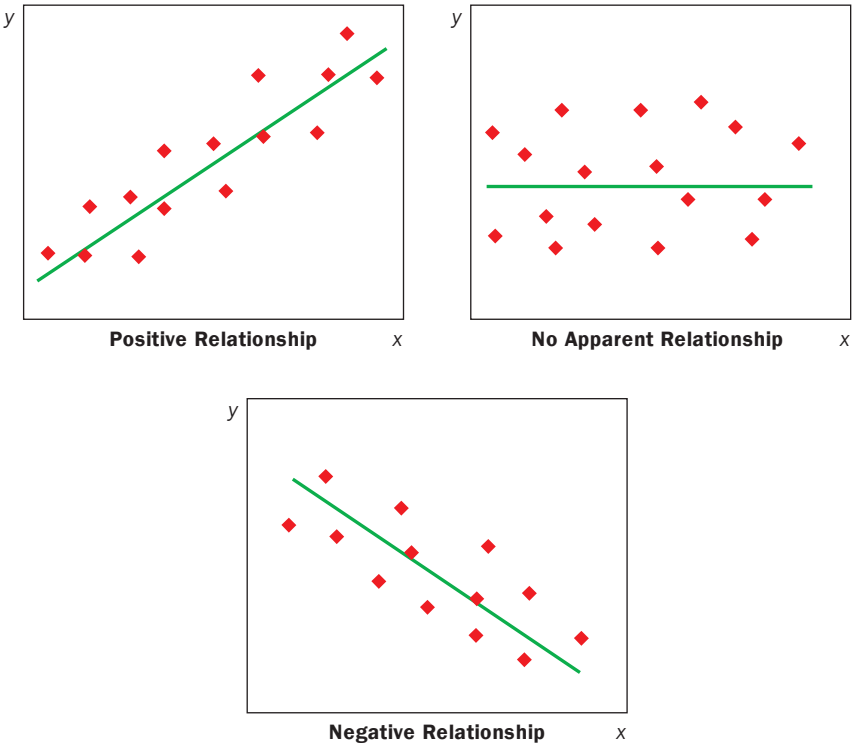
Some general scatter diagram patterns and the types of relationships they suggest are shown in Figure 2.9. The top left panel depicts a positive relationship similar to the one for the number of commercials and sales example. In the top right panel, the scatter diagram shows no apparent relationship between the variables. The bottom panel depicts a negative relationship where y tends to decrease as x increases.

FIGURE 2.8
Scatter diagram and trend line for the hi-fi equipment store



*The equation of the trend line is $y = 4.95x + 36.15$. The slope of the trend line is 4.95 and the y-intercept (the point where the line intersects the y axis) is 36.15. We will discuss in detail the interpretation of the slope and y-intercept for a linear trend line in Chapter 14 when we study simple linear regression.

FIGURE 2.9
Types of relationships depicted by scatter diagrams



EXERCISES

Methods

20. The following data are for 30 observations involving two qualitative variables, *X* and *Y*. The categories for *X* are A, B and C; the categories for *Y* are 1 and 2.

Observation	*X*	*Y*	*Observation*	*X*	*Y*
1	A	1	16	B	2
2	B	1	17	C	1
3	B	1	18	B	1
4	C	2	19	C	1
5	B	1	20	B	1
6	C	2	21	C	2
7	B	1	22	B	1
8	C	2	23	C	2
9	A	1	24	A	1
10	B	1	25	B	1
11	A	1	26	C	2
12	B	1	27	C	2
13	C	2	28	A	1
14	C	2	29	B	1
15	C	2	30	B	2


CROSSTAB

COMPLETE SOLUTIONS

a. Construct a cross-tabulation for the data, with *X* as the row variable and *Y* as the column variable.
b. Calculate the row percentages.
c. Calculate the column percentages.
d. What is the relationship, if any, between *X* and *Y*?

21. The following 20 observations are for two quantitative variables.

SCATTER

Observation	*X*	*Y*	*Observation*	*X*	*Y*
1	−22	22	11	−37	48
2	−33	49	12	34	−29
3	2	8	13	9	−18
4	29	−16	14	−33	31
5	−13	10	15	20	−16
6	21	−28	16	−3	14
7	−13	27	17	−15	18
8	−23	35	18	12	17
9	14	−5	19	−20	−11
10	3	−3	20	−7	−22

a. Construct a scatter diagram for the relationship between *X* and *Y*.
b. What is the relationship, if any, between *X* and *Y*?

Applications

22. Recently, management at Oak Tree Golf Course received a few complaints about the condition of the greens. Several players complained that the greens are too fast. Rather than react to the comments of just a few, the Golf Association conducted a survey of 100 male and 100 female golfers. The survey results are summarized here.

COMPLETE SOLUTIONS

	Male golfers *Greens condition*			*Female golfers* *Greens condition*	
Handicap	*Too fast*	*Fine*	*Handicap*	*Too fast*	*Fine*
Under 15	10	40	*Under 15*	1	9
15 or more	25	25	*15 or more*	39	51

a. Combine these two cross-tabulations into one with 'male', 'female' as the row labels and the column labels 'too fast' and 'fine'. Which group shows the highest percentage saying that the greens are too fast?
b. Refer to the initial cross-tabulations. For those players with low handicaps (better players), which group (male or female) shows the highest percentage saying the greens are too fast?
c. Refer to the initial cross-tabulations. For those players with higher handicaps, which group (male or female) shows the highest percentage saying the greens are too fast?
d. What conclusions can you draw about the preferences of men and women concerning the speed of the greens? Are the conclusions you draw from part (a) as compared with parts (b) and (c) consistent? Explain any apparent inconsistencies.

23. The file 'House Sales' on the online platform contains data for a sample of 50 houses advertised for sale in a regional UK newspaper. The first five rows of data are shown for illustration below.

Price	*Location*	*House type*	*Bedrooms*	*Reception rooms*	*Bedrooms + receptions*	*Garage capacity*
234 995	Town	Detached	4	2	6	1
319 000	Town	Detached	4	2	6	1
154 995	Town	Semi-detached	2	1	3	0
349 950	Village	Detached	4	2	6	2
244 995	Town	Detached	3	2	5	1

HOUSE SALES

a. Prepare a cross-tabulation using sale price (rows) and house type (columns). Use classes of 100 000–199 999, 200 000–299 999, etc. for sale price.
b. Compute row percentages and comment on any relationship between the variables.

24. Refer to the data in Exercise 23.

a. Prepare a cross-tabulation using number of bedrooms and house type.
b. Prepare a frequency distribution for number of bedrooms.
c. Prepare a frequency distribution for house type.
d. How has the cross-tabulation helped in preparing the frequency distributions in parts (b) and (c)?

25. The file 'OECD 2012' on the online platform contains data for 33 countries taken from the website of the Organization for Economic Cooperation & Development in mid-2012. The two variables are the Gini coefficient for each country and the percentage of children in the country estimated to be living in poverty. The Gini coefficient is a widely used measure of income inequality. It varies between 0 and 1, with higher coefficients indicating more inequality. The first five rows of data are shown for illustration below.

Country	*Child poverty (%)*	*Income inequality*
Australia	14.0	0.336
Austria	7.9	0.261
Belgium	11.3	0.259
Canada	15.1	0.324
Czech Republic	8.4	0.256

OECD 2012

a. Prepare a scatter diagram using the data on child poverty and income inequality.
b. Comment on the relationship, if any, between the variables.

ONLINE RESOURCES

For the data files, online summary, additional questions and answers, and software section for Chapter 2, go to the accompanying online platform.

SUMMARY

A set of data, even if modest in size, is often difficult to interpret directly in the form in which it is gathered. Tabular and graphical methods provide procedures for organizing and summarizing data so that patterns are revealed and the data are more easily interpreted.

Figure 2.10 shows the tabular and graphical methods presented in this chapter.

Frequency distributions, relative frequency distributions, percentage frequency distributions, bar charts and pie charts were presented as tabular and graphical procedures for summarizing qualitative data. Frequency distributions, relative frequency distributions, percentage frequency distributions, dot plots, histograms, cumulative frequency distributions, cumulative relative frequency distributions, cumulative percentage frequency distributions and cumulative distribution plots (ogives) were presented as ways of summarizing quantitative data. A stem-and-leaf display provides an exploratory data analysis technique that can be used to summarize quantitative data.

Cross-tabulation was presented as a tabular method for summarizing data for two variables. An example of Simpson's paradox was set out, to illustrate the care that must be taken when interpreting relationships between two variables using aggregated data. The scatter diagram was introduced as a graphical method for showing the relationship between two quantitative variables.

With large data sets, computer software packages are essential in constructing tabular and graphical summaries of data. The software guides on the online platform show how EXCEL, IBM SPSS and MINITAB can be used for this purpose.

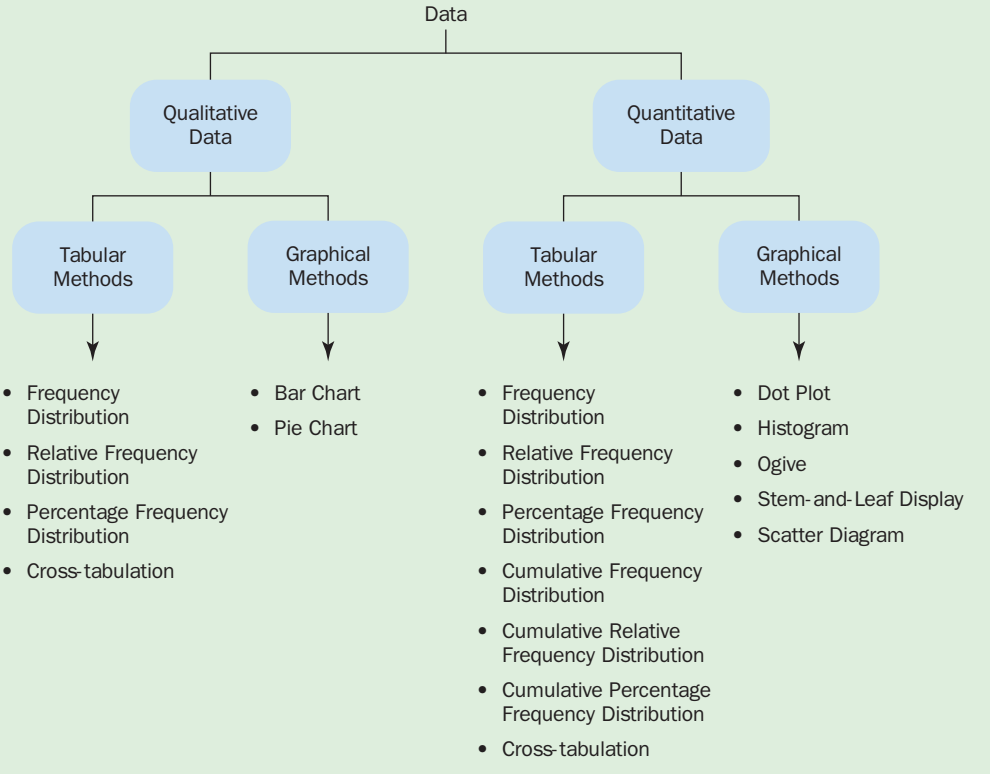


FIGURE 2.10
Tabular and graphical methods for summarizing data

KEY TERMS

Bar chart
Bar graph
Class midpoint
Cross-tabulation
Cumulative frequency distribution
Cumulative percentage frequency distribution
Cumulative relative frequency distribution
Dot plot
Exploratory data analysis
Frequency distribution
Histogram
Ogive

Percentage frequency distribution
Pie chart
Qualitative data
Quantitative data
Relative frequency distribution
Scatter diagram
Simpson's paradox
Stem-and-leaf display
Trend line

KEY FORMULAE

Relative frequency

$$\text{Relative frequency of a class} = \frac{\text{Frequency of the class}}{n} \qquad \textbf{(2.1)}$$

Approximate class width

$$\frac{\text{Largest data value} - \text{Smallest data value}}{\text{Number of classes}} \qquad \textbf{(2.2)}$$

CASE PROBLEM

In The Mode Fashion Stores

In The Mode is a chain of women's fashion stores. The chain recently ran a promotion in which discount coupons were sent to customers. Data collected for a sample of 100 in-store credit card transactions during a single day following the promotion are contained in the file 'Mode' on the online platform. A portion of the data set is shown below. A non-zero amount for the discount variable indicates that the customer brought in the promotional coupons and used them. For a very few customers, the discount amount is actually greater than the sales amount (see, for example, customer 4). In The Mode's management would like to use this sample data to learn about its customer base and to evaluate the promotion involving discount coupons.

Managerial report

Use tables and charts to help management develop a customer profile and to evaluate the promotional campaign. At a minimum, your report should include the following.

1. Percentage frequency distributions for key variables.
2. A bar chart or pie chart showing the percentage of customer purchases possibly attributable to the promotional campaign.
3. A cross-tabulation of type of customer (regular or promotional) versus sales. Comment on any similarities or differences present.
4. A scatter diagram of sales versus discount for only those customers responding to the promotion. Comment on any relationship apparent between sales and discount.
5. A scatter diagram to explore the relationship between sales and customer age.

MODE

Customer	*Method of payment*	*Items*	*Discount*	*Sales*	*Gender*	*Marital status*	*Age*
1	Visa Debit	1	0.00	39.50	Male	Married	32
2	Store Card	1	25.60	102.40	Female	Married	36
3	Store Card	1	0.00	22.50	Female	Married	32
4	Store Card	5	121.10	100.40	Female	Married	28
5	Mastercard	2	0.00	54.00	Female	Married	34
6	Mastercard	1	0.00	44.50	Female	Married	44
7	Store Card	2	19.50	78.00	Female	Married	30
8	Visa	1	0.00	22.50	Female	Married	40
9	Store Card	2	22.48	56.52	Female	Married	46
10	Store Card	1	0.00	44.50	Female	Married	36