The Open University

M249 Practical modern statistics

Introductory Unit

Introduction to statistical modelling

About this course

M249 Practical Modern Statistics uses the software packages *SPSS for Windows* (SPSS Inc.) and *WinBUGS*, and other software. This software is provided as part of the course, and its use is covered in the *Introduction to statistical modelling* and in the four computer books associated with Books 1 to 4.

Cover image courtesy of NASA. This photograph, acquired by the ASTER instrument on NASA's Terra satellite, shows an aerial view of a large alluvial fan between the Kunlun and Altun mountains in China's Xinjiang province. For more information, see NASA's Earth Observatory website at http://earthobservatory.nasa.gov.

This publication forms part of an Open University course. Details of this and other Open University courses can be obtained from the Student Registration and Enquiry Service, The Open University, PO Box 197, Milton Keynes, MK7 6BJ, United Kingdom: tel. +44 (0)870 333 4340, e-mail general-enquiries@open.ac.uk

Alternatively, you may visit the Open University website at http://www.open.ac.uk where you can learn more about the wide range of courses and packs offered at all levels by The Open University.

To purchase a selection of Open University course materials, visit the webshop at www.ouw.co.uk, or contact Open University Worldwide, Michael Young Building, Walton Hall, Milton Keynes, MK7 6AA, United Kingdom, for a brochure: tel. +44 (0)1908 858785, fax +44 (0)1908 858787, e-mail ouwenq@open.ac.uk

The Open University, Walton Hall, Milton Keynes, MK7 6AA.

First published 2006.

Edited, designed and typeset by The Open University, using the Open University TEX System.

This edition produced for Web publication by The Open University. Third party copyright images on pages 5 and 28 of the print edition are not available in this Web version.

ISBN 978 0 7492 1365 7

1.1

Contents

Study guide

This unit has two aims: first, to revise basic statistical ideas and techniques with which you are assumed to be familiar when you study Books 1 to 4 (or to provide a concise introduction to any with which you are not familiar); and secondly, to introduce SPSS, the main statistical package used in this course.

Other software is used in Book 4 and will be introduced when you study that book.

There are seven sections in this unit. Sections 1, 3, 4 and 5 do not require the use of a computer, while Sections 2, 6 and 7 are computer-based. If you have not already installed SPSS, then you will need to do so before you begin Section 2. Section 2 is longer than average, and Section 6 is shorter than average.

Instructions on how to install the course software are given in the *Software Guide*.

This unit contains both activities, which are included at various points throughout the text, and exercises. Their purposes are quite different. Activities form a central part of the text, and you should try to do them as you work through the unit. Exercises are provided to give you further practice at applying certain ideas and techniques, *if you need it*: you should not routinely try them all as you work through the unit. You may find it more helpful to try them only if you are unsure that you have understood an idea. Exercises that do not require the use of a computer are included at the end of Sections 1, 3, 4 and 5. There are no exercises at the end of Sections 2 and 6, but Section 7 consists of modelling exercises that require the use of a computer. Some of these exercises use ideas and techniques from several sections of the unit. You can use these exercises, if you wish, to help consolidate your understanding, or for further practice with SPSS. Comments on some of the computer-based activities in Sections 2 and 6 are included within the activities. Solutions to the other activities and all the exercises may be found at the back of the unit.

This unit will require seven study sessions of between $2\frac{1}{2}$ and 3 hours. The idea of a 'study session' of $2\frac{1}{2}$–3 hours has been introduced simply to help you plan your study.

One possible study pattern is as follows.

Study session 1: Section 1.

Study session 2: Section 2. You will need access to your computer for this session.

Study session 3: Section 3.

Study session 4: Section 4.

Study session 5: Section 5.

Study session 6: Section 6. You will need access to your computer for this session.

Study session 7: Consolidating your work on this unit — for example, by trying some of the modelling exercises in Section 7 — and answering the TMA question on the unit. You will need access to your computer for this session.

Introduction

In M249 *Practical Modern Statistics* you will be introduced to four topics in statistical modelling: medical statistics, time series, multivariate analysis and Bayesian statistics. Each of these topics is largely self-contained, and most of the statistical methods required will be taught where they are needed. This introductory unit includes a review of the basic statistical techniques that form the common background to the more advanced topics to be covered later. SPSS, the main statistical package used in M249, is also introduced.

The emphasis throughout this unit is on statistical modelling as an approach to deriving information on a particular topic of interest. Two topics with an environmental theme are used to motivate and link the material: levels of fish stocks in the North Sea and the Irish Sea, and air quality and asthma in Nottingham. In Section 1, methods for presenting data using graphs and numerical summaries are described. An introduction to SPSS is given in Section 2, where you will learn how to obtain graphs and numerical summaries. Some commonly used probability models are described in Section 3, while approaches to statistical inference are discussed in Section 4, including confidence intervals and significance tests. Methods for describing and analysing related variables are described in Section 5. In Section 6, you will learn how to implement some of the techniques described in Sections 3, 4 and 5 using SPSS. Finally, Section 7 consists of computer-based exercises on the material covered in Sections 1 to 6.

1 Presenting and summarizing data: the silver darlings

There are many ways of presenting data, and which method to use depends entirely on the type and amount of data available, and the purpose of the presentation. In this section, three ways of presenting data are reviewed: tables, graphs and numerical summaries. This is done in the context of several data sets relating to fish stocks around the British Isles. The data used in this section and in Section 2 were obtained in October 2004 from the website of the Department for the Environment, Food and Rural Affairs (http://www.defra.gov.uk).

In Subsection 1.1, tables, bar charts, line plots and scatterplots are discussed. Numerical summaries and histograms are reviewed in Subsection 1.2.

1.1 Presenting data

Fishing for herring, the 'silver darlings' of the title of this section, was once a mainstay of the economy of the east coast of Britain, from Great Yarmouth in East Anglia to Peterhead in Scotland. The herring industry has now largely disappeared, and has been replaced by more intensive forms of fishing, which are threatening fish stocks in many sea areas. Fish stocks are now carefully monitored. This provides information that can be used to set fishing quotas, and also to assess the impact of environmental

Example 1.1 Annual fish catch 1999

Table 1.1 shows the total annual fish catch in the North Sea, for seven fish species, measured in thousands of tonnes, for the year 1999. The key features of this table, in addition to the data, are a title describing the contents of the table (with the relevant units — in this case thousands of tonnes, which is abbreviated as 'thousand tonnes'), and short column headings. Note that the data have been rounded to the nearest thousand tonnes.

Table 1.1 Total catch (thousand tonnes) for seven fish species, North Sea, 1999

Fish species	Catch
Cod	96
Herring	372
Haddock	112
Whiting	59
Sole	23
Plaice	81
Saithe	114

Tables are ideal for conveying detailed numerical information. (Large tables are usually stored on a computer as databases or spreadsheets.) However, to illustrate a particular point, a graph might be better than a table. For example, it is clear from Table 1.1 that the herring catch in 1999 was much greater than that for sole. However, the relative size of the different catches may be conveyed more effectively using a suitable diagram.

For the data in Table 1.1, a suitable diagram is a **bar chart**, in which the 1999 catch for each species is represented by a bar, the length of the bar indicating the size of the catch. A bar chart with vertical bars is shown in Figure 1.1(a).

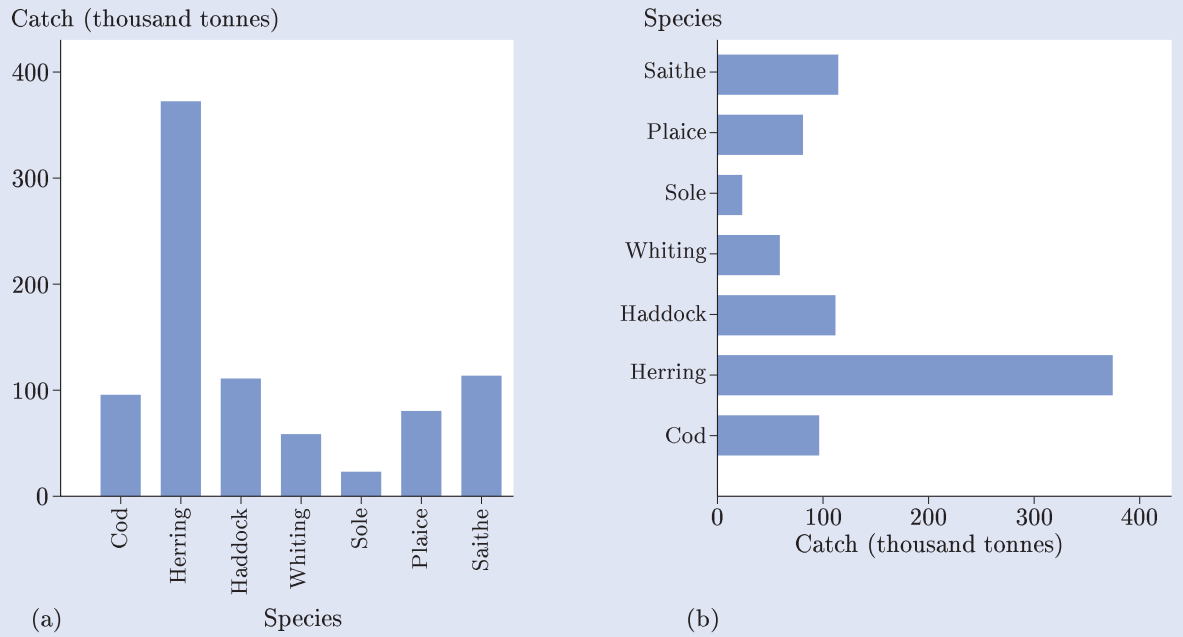

Figure 1.1 Total catch for seven fish species, North Sea, 1999

The bar chart in Figure 1.1(a) shows at a glance that in 1999 the herring catch far outstripped the catches for the other fish species.

Bar charts may also be drawn with horizontal bars. A horizontal bar chart of the data in Table 1.1 is shown in Figure 1.1(b). Horizontal bar charts are sometimes more convenient than vertical bar charts, when the labels for the bars are long, or when there is a large number of bars, as the bar labels may be easier to read. ♦

Bar charts can be used to represent changes over time when there are only a few time points. This is illustrated in Example 1.2.

Example 1.2 Variation in the fish catch, 1979–99

Table 1.2 shows the annual North Sea catch for the seven species of fish listed in Table 1.1, for the years 1979, 1989 and 1999.

Table 1.2 Annual North Sea catch (thousand tonnes)

	1979	1989	1999
Cod	270	140	96
Herring	25	788	372
Haddock	146	109	112
Whiting	244	124	59
Sole	23	22	23
Plaice	145	170	81
Saithe	136	118	114

An issue of interest, particularly to biologists and to people involved in the fishing industry, is the variation in the catch over time, for different species. This variation can be conveyed using a **comparative bar chart**, such as that shown in Figure 1.2.

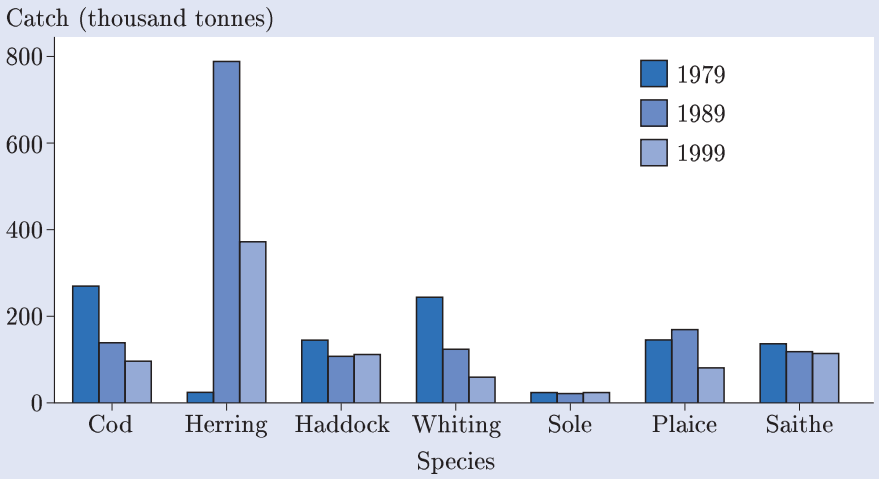

Figure 1.2 Comparative bar chart for annual catch of seven fish species

This bar chart is similar to the one in Figure 1.1(a), except that now three bars are drawn side-by-side for each fish species, representing the catches for 1979, 1989 and 1999. ♦

Activity 1.1 Trends in fish catches

(a) Use Figure 1.2 to identify a general trend in the fish catch over time for the fish species represented.

(b) Are there any exceptions to this general trend?

When there are only a few time points, a bar chart is fine for showing trends. However, to obtain a more complete picture of changes over time, more time points must be used, but then a bar chart will be too cluttered to be of much use. In such circumstances, a **line plot** is used.

Statistical techniques for the analysis of data consisting of observations collected at regular time intervals are described in Book 2 *Time series*.

Example 1.3 Annual herring catch

In Activity 1.1, you saw that the North Sea herring catch increased by a very large amount between 1979 and 1989. A line plot of the total North Sea herring catch (in thousands of tonnes) for each year between 1963 and 1999 is shown in Figure 1.3.

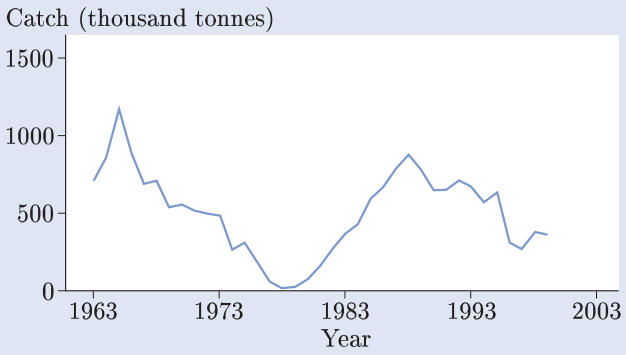

Figure 1.3 Annual catch of North Sea herring

This line plot gives a more complete picture of the variation in the herring catch than does the bar chart in Figure 1.2. In particular, it shows that there was a big drop in the annual herring catch in the late 1970s, followed by a peak in the late 1980s. ♦

A measure of mature fish stocks — that is, of the quantity of mature fish in the sea — is given by the biomass. The biomass is the total mass of mature fish, and is measured in thousands of tonnes. A line plot of the estimated herring biomass in the North Sea, between 1963 and 2003, is shown in Figure 1.4, together with the line plot of the annual herring catch from Figure 1.3.

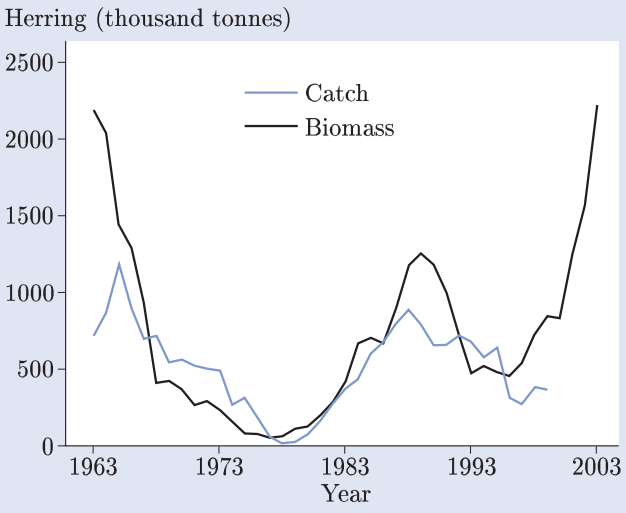

Figure 1.4 Biomass and annual catch of North Sea herring

Activity 1.2 The impact of over-fishing

In the 1970s herring stocks in the North Sea were seriously depleted by over-fishing.

(a) What features of Figure 1.4 indicate that there was a problem with over-fishing for herring in the 1970s?

(b) Fishing for herring in the North Sea was severely restricted between 1978 and 1982. What does Figure 1.4 suggest about the impact of the restrictions?

A line plot is particularly useful for representing data ordered in time. To display the relationship between two variables, neither of which is time, a **scatterplot** can be used.

Example 1.4 Herring biomass and new recruits

Two variables are commonly used to monitor fish stocks: the biomass and the number of new recruits. New recruits are young fish who become of age to be fished. Clearly, the two variables are likely to be related: the more mature fish there are, the more new fish they will produce. In Figure 1.5, the estimated number of newly recruited herring (in billions) is plotted against the herring biomass (in thousands of tonnes) in the North Sea, for each year between 1963 and 2003.

The age at which fish are deemed to be old enough to be fished varies from species to species.

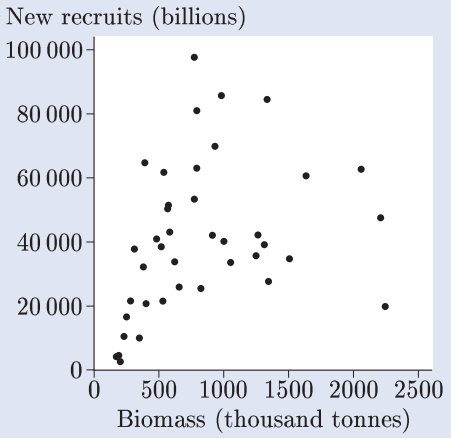

Figure 1.5 North Sea herring: new recruits and biomass ♦

Activity 1.3 The new silver darlings

(a) Briefly describe the relationship between herring new recruits and biomass in Figure 1.5.

(b) How does the variability in the number of new recruits change with the biomass?

(c) Identify any possible outliers (that is, any observations that do not appear to fit the overall pattern) in the scatterplot.

1.2 Describing samples of data

In order to describe a sample of data it is useful to begin by classifying the data as either *numerical* or *categorical*. **Numerical** data are numbers; **categorical** data are categories. For example, the variable 'Fish species' in Table 1.1 is a categorical variable, taking the values cod, herring, haddock, and so on. Sometimes, categories may be represented by numbers. For example, for the variable 'Sex' (taking values Male and Female), Male could be coded as 1, Female as 2. But this does not make Sex a numerical variable: the numbers 1 and 2 are just labels. The category Male could equally well be coded as 2 and the category Female as 1.

A distinction is drawn between two different kinds of numerical data: *discrete* data and *continuous* data. **Discrete** data arise when variables are restricted to taking particular values — for example, counts of fish $(0, 1, 2, \ldots)$. The herring biomass, on the other hand, is a **continuous** variable because it can take any value in a continuous range of values. In some instances, it is reasonable to treat discrete variables as if they were continuous. For example, in Example 1.4, the annual number of new herring recruits in the North Sea is a discrete variable. But it can reasonably be treated as if it were continuous, because it can take a great many different values, and the exact number of herring is not important.

The distribution of a sample of categorical observations may be represented by a bar chart, as in Figure 1.1. A bar chart can also be used to represent discrete numerical data. The simplest way to represent the distribution of a sample of observations on a continuous variable is using a **histogram**. A histogram of the annual North Sea herring catch between 1963 and 1999 is shown in Figure 1.6(a).

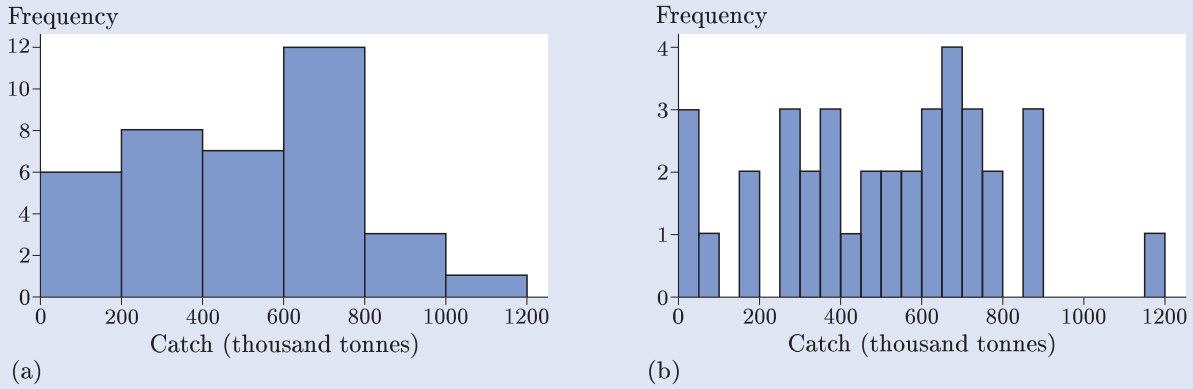

Figure 1.6 Two histograms of annual herring catch, 1963–99

In Figure 1.6(a), the data are grouped into intervals, or **bins**: 0–200, 200–400, and so on. If an observation lies exactly on a boundary it is placed in the bin immediately to the left of the boundary. For example, a catch of exactly 200 thousand tonnes is placed in the bin 0–200. The observations in each interval are represented by a vertical bar, the height of the bar being equal to the number of observations in the interval, that is, the **frequency** of the observations. For example, there were 12 observations between 600 000 and 800 000 tonnes, so the height of the bar for the bin 600–800 is 12. A difference between bar charts and histograms is that in bar charts gaps are left between the bars, while in histograms the bars are contiguous (unless there is an interval with no observations).

A histogram gives an impression of the range of the data and the overall shape of its distribution. However, it is important to remember that changing the width of the bins or the boundaries separating the bins can alter the appearance of the distribution, as illustrated in Figure 1.6. The interval width in Figure 1.6(a) is 200 whereas in Figure 1.6(b) it is 50. The peak of the distribution is less apparent in Figure 1.6(b) than it is in Figure 1.6(a). Several of the bins in Figure 1.6(b) are empty, so there are gaps in this histogram. In this case, perhaps Figure 1.6(a) gives a better impression of the shape of the distribution than does

Figure 1.6(b). It is not easy to give general rules about how many bins should be used. For very large data sets, it may be appropriate to use a large number of bins in order to obtain as much information as possible about the shape of the distribution. On the other hand, given a small data set, even as few as five bins may be too many. However, a rough guide is to begin by choosing between 5 and 20 bins, then to adjust the number of bins up or down if this seems desirable.

Numerical summaries complement graphical displays such as histograms: commonly, both graphical displays and numerical summaries are used to represent a sample of data. As is often the case in statistics, there is a choice of numerical summaries that can be used. The numerical summaries reviewed here are in two groups: measures of location and measures of dispersion.

In Section 2, a further summary, the skewness, is discussed.

Measures of location describe the 'average' or 'typical' value of a sample. They include the **mean**, **median** and **mode**, which are defined in the following box.

Measures of location

Let $x_1, x_2, \ldots, x_n$ denote a sample of n data values. The **mean** of the sample, which is denoted $\overline{x}$, is the arithmetic average of the data values:

$$\overline{x} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

The expression $\sum_{i=1}^{n} x_i$ means $x_1 + x_2 + \cdots + x_n$.

The **median** m of a sample of data with an odd number of values is the middle value of the data set when the values are placed in order of increasing size. If the sample size is even, the median is halfway between the two middle values.

For categorical data, the **mode** is the most frequently occurring (or **modal**) category. The term mode is also used to describe a clear peak in a histogram or a bar chart of a set of numerical data.

Example 1.5 Mercury contamination in plaice

In Subsection 1.1, you saw that over-fishing can have a large impact on fish stocks. Also of concern, for the health of both fish and humans, are the levels of pollution from sewage or effluent from industry. Contamination of various pollutants in fish is therefore carefully monitored.

Table 1.3 contains the average concentration of mercury contamination in plaice caught in the North Sea and the Irish Sea for various years between 1984 and 1993. The contamination is measured in mg/kg wet weight.

Table 1.3 Average concentration of mercury contamination (measured in mg/kg) in plaice

Year	North Sea	Irish Sea
1984	0.06	0.11
1985	0.05	0.09
1986	0.04	0.11
1987	0.05	0.11
1988	0.06	0.12
1989	0.05	0.10
1990	0.05	–
1991	0.05	–
1992	–	0.10
1993	0.05	0.09

There are some missing values in Table 1.3. Nevertheless, it seems clear that there is no obvious trend in the contamination levels between 1984 and 1993.

Consider the data for the North Sea. The mean concentration of mercury contamination in plaice over the decade 1984–93 is

$$\begin{aligned}\overline{x} &= \tfrac{1}{9}(0.06 + 0.05 + 0.04 + 0.05 + 0.06 + 0.05 + 0.05 + 0.05 + 0.05)\\ &= 0.05111\ldots\\ &\simeq 0.051.\end{aligned}$$

To obtain the median, the values must first be arranged in order of increasing size, as follows.

0.04 0.05 0.05 0.05 0.05 0.05 0.05 0.06 0.06

For an odd number of values, the median is the middle value. There are nine values, so the middle value is the fifth value, which is 0.05. So the median m is 0.05.

A histogram of the mercury concentration in North Sea plaice is given in Figure 1.7. This shows that the mode lies in the interval 0.045 to 0.055. ♦

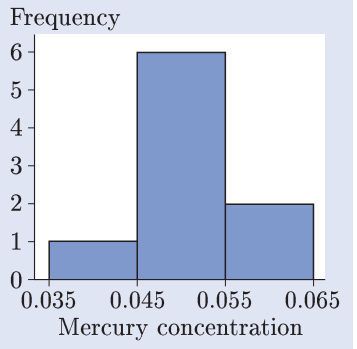

Figure 1.7 Average mercury concentration in North Sea plaice, 1984–93

Activity 1.4 *Mode and median for the fish catch*

(a) Use either Table 1.2 or Figure 1.2 to identify the modal species of fish caught in the North Sea in each of the years 1979, 1989 and 1999.

(b) Use the histogram in Figure 1.6(a) to identify the interval that includes the median annual herring catch for the 37 years 1963–99.

Measures of dispersion describe the variation within a sample around its average value. As with measures of location, several measures of dispersion are commonly used in statistics. The measures used in this course are the **standard deviation** and the **variance**, which are defined for numerical data only. These are defined in the following box.

Measures of dispersion

Let $x_1, x_2, \ldots, x_n$ denote a sample of n data values, with sample mean $\overline{x}$. The **standard deviation** of the sample, denoted s, is given by

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

The quantity s^2, the square of the standard deviation, is known as the **variance** of the sample.

Note that the sum in the expression for the sample standard deviation is divided by $n-1$ rather than n. For a sample of size 1, the sample standard deviation is undefined.

Example 1.6 *Variation in mercury levels*

The mean of the mercury concentration levels in North Sea plaice in Table 1.3 is $0.05111\ldots \simeq 0.051$. So the variance is

See Example 1.5.

$$\begin{aligned} s^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \\ &\simeq \frac{1}{9-1}\left((0.06 - 0.05111)^2 + (0.05 - 0.05111)^2 + \cdots + (0.05 - 0.05111)^2\right) \\ &= 0.00003611\ldots \simeq 0.000036. \end{aligned}$$

Hence the standard deviation is

$$\begin{aligned} s &\simeq \sqrt{0.00003611\ldots} \\ &= 0.006009 \simeq 0.0060. \end{aligned}$$ ♦

Notice that, in Example 1.6, four significant figures were retained for the mean when calculating the variance and the standard deviation. This was done in order to avoid introducing rounding errors.

Activity 1.5 Mercury contamination in Irish Sea plaice

Use the data in Table 1.3 to calculate the following summary measures for the mercury concentrations in Irish Sea plaice.

(a) The mean and the median.

(b) The variance and the standard deviation.

Note that the measures of location and dispersion discussed in this subsection all relate to a *sample* of data values. Corresponding measures relating to an entire *population* will be described in Section 3. To avoid confusion with their population counterparts, the numerical summaries described here are sometimes called *sample* summaries: sample mean, sample median, sample standard deviation, and sample variance.

Summary of Section 1

In this section, several types of graphs have been reviewed: bar charts, line plots, scatterplots and histograms have been discussed. Numerical summaries have also been reviewed: measures of location, such as the mean, median and mode, and measures of spread, such as the standard deviation and the variance, have been defined.

Exercise on Section 1

Exercise 1.1 Differences in pollution

For seven of the years listed in Table 1.3, the concentration of mercury contamination in plaice was measured both in the North Sea and in the Irish Sea. One approach to comparing contamination levels in the two sea areas is to examine the differences between the contamination levels in the two areas in these years.

(a) Suggest an appropriate graph for displaying these differences.

(b) Calculate the difference between the mercury contamination levels in plaice (Irish Sea minus North Sea) for each of the seven years in which values were obtained in both areas, and arrange them in order of increasing size.

(c) Calculate the mean and the median of the differences.

(d) Calculate the variance and the standard deviation of the differences.

(e) What might you conclude about the differences between mercury contamination levels in plaice in the Irish Sea and the North Sea?

2 Introducing SPSS: North Sea cod

In this section, the statistical software package SPSS is introduced. If you have not yet installed the course software on your computer, then do so now. Instructions are given in the *Software Guide*.

SPSS is used in Books 1, 2 and 3.

In Subsection 2.1, you will familiarize yourself with SPSS. In Subsection 2.2, you will learn how to print output, or paste it into another document. The use of SPSS to obtain line plots and scatterplots is described in Subsection 2.3, and histograms and numerical summaries in Subsection 2.4.

For clarity of presentation, bold-face type has been used for file names throughout the course. The names of menus and items in menus are also printed in bold-face when referred to in the text, as are options and the names of fields and buttons in dialogue boxes. When you are asked to use the mouse to click on an item, you should assume that this refers to the left-hand mouse button. If you need the right-hand mouse button this will be stated explicitly.

This section is organized around some data sets on stocks of cod in the North Sea. Cod stocks in the North Sea have been declining for some time. There is concern about the sustainability of these stocks, particularly following the collapse in the early 1990s of the once plentiful cod stocks off the coast of Newfoundland in Canada due to over-fishing.

2.1 Navigating SPSS

Activity 2.1 Getting started

Run SPSS now: click on the **Start** button, move the mouse pointer to **Programs** (or **All Programs** — this depends on the version of *Windows* you are using), then to **SPSS for Windows**, and click on **SPSS xx for Windows** (where xx is the version number).

The SPSS opening screen contains the **SPSS Data Editor** window and the **SPSS xx for Windows** dialogue box (which is uppermost). You will not need this dialogue box. So check the box labelled **Don't show this dialog in the future** (by clicking on it). Then close the dialogue box by clicking on the button marked **x** at the right-hand end of the title bar. The **SPSS Data Editor** window shown in Figure 2.1 will remain.

If you would prefer the window to be larger, then maximize it (by clicking on the maximize button, which is next to the close button in the title bar).

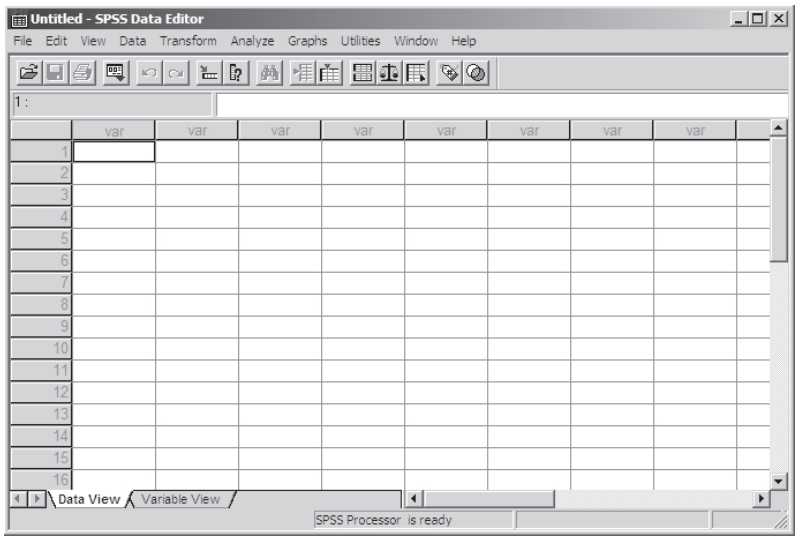

Figure 2.1 The **SPSS Data Editor** window

As with many *Windows*-based software packages, there is a menu bar at the top of the window. Below this is a toolbar containing a number of buttons. The main part of the window contains two tabbed panels, named **Data View** and **Variable View**; these are discussed further in Activity 2.2. Initially **Data View** is uppermost.

Click on **File** in the menu bar. Some of the items in the menu appear in bold, meaning that they are currently available. Others are in faint text, indicating that they are not currently available — they are disabled. For example, **Save** is disabled (because there is nothing yet to save). An arrowhead pointing to the right on a menu item indicates the existence of a submenu. Before moving on to the next activity, spend a few minutes exploring the menus and their submenus. Note the different types of facilities available in the menus.

By the way, you can exit from SPSS at any time by clicking on **File**, and choosing **Exit** from the **File** menu (by clicking on it).

Comment

The roles of the menus may be summarized as follows.

- ◇ The **File** menu is used for importing and exporting or printing data.
- ◇ The **Edit** menu contains commands for editing files.
- ◇ The **View** menu enables you to control the appearance of the software.
- ◇ The **Data** menu is used to organize data files.
- ◇ The **Transform** menu allows you to define new variables from existing variables.
- ◇ The **Analyze** menu contains the main statistical routines.
- ◇ The **Graphs** menu provides a range of graphical tools.
- ◇ The **Utilities** menu provides access to the command language.
- ◇ The **Window** menu enables you to activate a particular window.
- ◇ Finally, the **Help** menu provides access to help.

Activity 2.2 Opening a data file

In SPSS, there is little you can do without first opening a data file (or creating a new data file). You are asked to open a data file in this activity. In SPSS, data are stored in files with the extension **.sav**. All the data files for M249 are located within the **M249 Data Files** folder within **My Documents**. The files required for this unit are stored in the **Introduction** subfolder of the **M249 Data Files** folder.

You will learn how to create a data file in *Computer Book 1*.

Data on the estimated stocks of cod in different sea areas are saved in the data file **codareas.sav**. Open this data file, using **Data...** from the **Open** submenu of the **File** menu, as follows.

- ◇ Click on **File**, move the mouse pointer to **Open**, then choose **Data...** from the **Open** submenu (by clicking on it). The **Open File** dialogue box will open.

The main panel shows the folders and files contained in the default SPSS directory, whose name appears in the **Look in** field at the top of the dialogue box. Navigate to the folder where the M249 files are stored, as follows.

- ◇ Click on the **My Documents** icon to the left of the main panel. The list of folders in **My Documents** will appear in the main panel.
- ◇ In the main panel, double-click on the folder **M249 Data Files**.
- ◇ Now double-click on the **Introduction** subfolder.

The files in the **Introduction** folder will be displayed as shown in Figure 2.2.

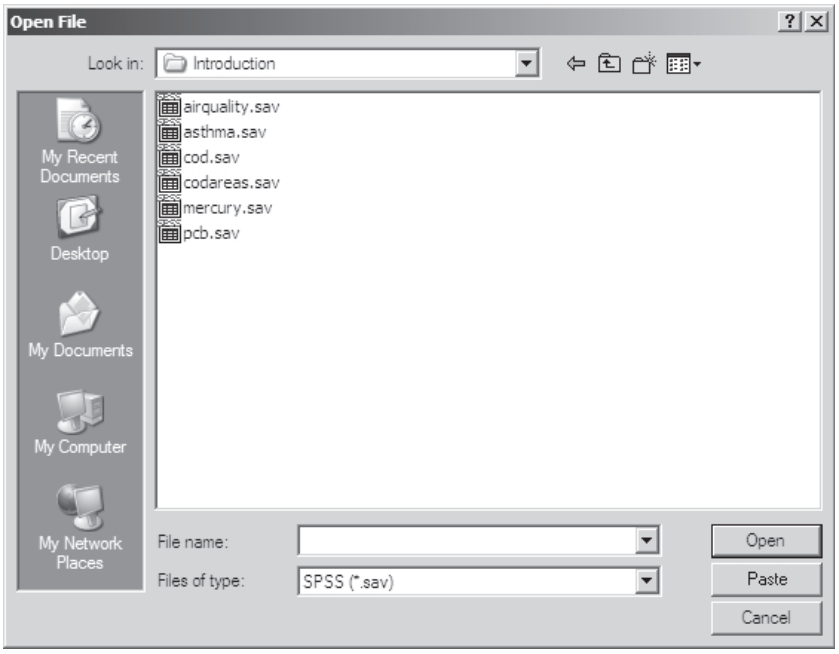

Figure 2.2 The **Open File** dialogue box

The extension **.sav** may not be displayed in the list of file names. Whether or not it is displayed depends on your computer's current settings (not on SPSS).

◇ Open the file **codareas.sav** by double-clicking on it. The data will appear in the **SPSS Data Editor**.

Alternatively, you can open a file by clicking on its name to select it, then on **Open**; or you can type its name in the **File name** field, then click on **Open**.

The file contains two variables, which appear in the **Data View** panel as columns named `area` and `biomass`. The variable `area` describes a sea area. For simplicity, names have been assigned to these. West of Scotland, for example, refers to the sea off that coast; the Celtic Sea includes the western part of the English Channel and the sea area to the south-west of Ireland. The variable `biomass` contains the estimated biomass of cod in 2002, in thousands of tonnes.

The biomass is the total mass of mature fish. This was defined just after Example 1.3.

In Activity 2.1, we noted that the **Data Editor** contains two tabbed panels, named **Data View** and **Variable View**. The tabs are located in the lower left-hand corner of the **Data Editor** window. **Data View** displays the data, whereas **Variable View** displays information about how the variables are formatted. Initially the **Data View** panel is uppermost. Click on the **Variable View** tab to see the **Variable View** panel. Generally, it is better to keep **Data View** uppermost, so that you can refer to the data. Click on the **Data View** tab so that **Data View** is once again uppermost.

Note that if you make any changes to a data file, whether changes to the data in **Data View**, or to the data formats in **Variable View**, you will be prompted to save the data file when you exit from SPSS. If you wish to do so, you should choose a file name different from that of the original file, so that you do not overwrite the original file. Now exit from SPSS.

Click on **Exit** in the **File** menu.

Activity 2.3 Producing a bar chart

In this activity you will obtain a bar chart for the cod biomass in the seven sea areas. You will need this bar chart in Activities 2.4 and 2.5, so try to do those activities immediately after this one.

Run SPSS now.

(a) You will need the data file **codareas.sav**. You could open it as described in Activity 2.2, but the following is a quicker way to open a data file that you have used recently.

- ◇ Move the mouse pointer to **Recently Used Data** in the **File** menu. A list of the data files you have used recently will appear.
- ◇ Click on **codareas.sav**, and SPSS will open the file.

(b) Bar charts are produced using **Bar...** from the **Graphs** menu. Obtain a bar chart showing the cod biomass in each of the seven sea areas, as follows.

- ◇ Choose **Bar...** from the **Graphs** menu (by clicking on it).

The **Bar Charts** dialogue box will open, as shown in Figure 2.3.

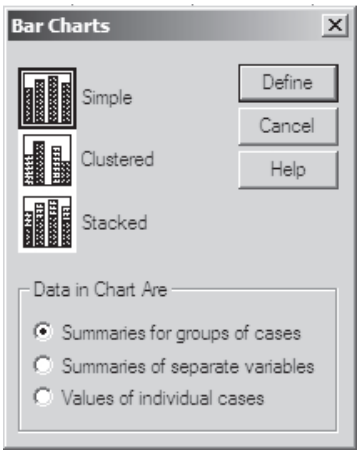

Figure 2.3 The **Bar Charts** dialogue box

This dialogue box requires you to choose the type of bar chart required and to indicate the format in which the data are stored.

- ◇ A bar chart for a single variable is required, so select **Simple** (by clicking on the corresponding bar chart).
- ◇ The heights of the bars are in the variable `biomass` (so they do not need to be calculated). In the **Data in Chart Are** area of the dialogue box, select **Values of individual cases** (by clicking on it or on its radio button).
- ◇ Click on the **Define** button. The **Define Simple Bar: Values of Individual Cases** dialogue box will open, as shown in Figure 2.4.

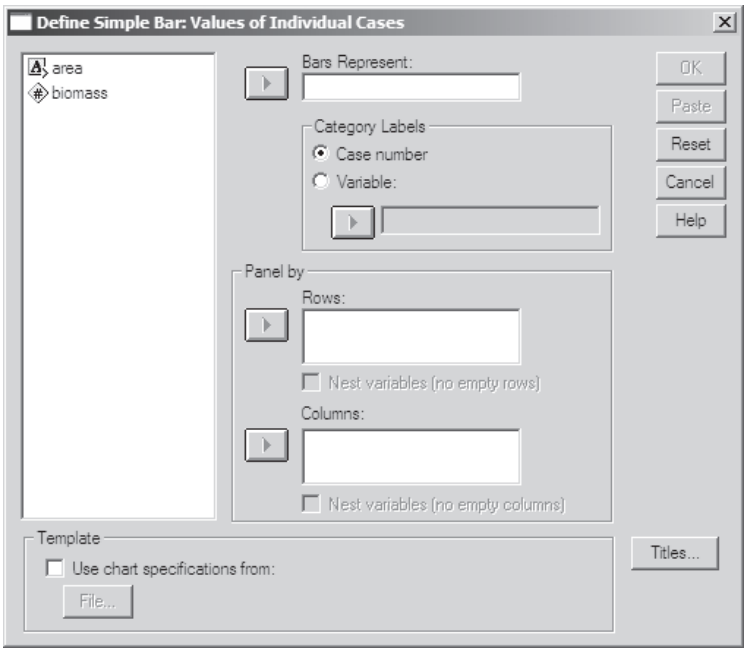

Figure 2.4 The **Define Simple Bar: Values of Individual Cases** dialogue box

This dialogue box is used to specify the variables that are to be used, and to annotate the bar chart. In this case, a bar chart showing the biomass in each of the sea areas is required. So the bars will represent the biomass, and the labels on the bars will be the sea areas. The variables `area` and `biomass` are listed in the panel on the left-hand side of the dialogue box.

- ◇ Click on `biomass` to select it.
- ◇ Click on the arrow to the left of the **Bars Represent** field and `biomass` will be entered in the field. (Notice that the direction of the arrow changes. You can remove `biomass` from the field by clicking on the arrow a second time.)

- ◇ In the **Category Labels** area, click on **Variable** (or on its radio button).
- ◇ Click on `area` (in the panel on the left-hand side of the dialogue box) to select it.
- ◇ Click on the arrow to the left of the **Variable** field to enter `area` in the field.

The method just described is the standard way to enter variables in fields in SPSS. From now on, we will refer to entering variables more briefly — for example, 'Enter `biomass` in the **Bars Represent** field'.

Titles and subtitles are added to bar charts using the **Titles** dialogue box, which is obtained using the **Titles...** button in the bottom right-hand corner of the **Define Simple Bar: Values of Individual Cases** dialogue box. Add a title to the bar chart, as follows.

- ◇ Click on **Titles...** to open the **Titles** dialogue box.
- ◇ Type a suitable title in the **Line 1** field of the **Title** area — for example, `Cod biomass by sea area`.
- ◇ Click on **Continue** to close the **Titles** dialogue box, then click on **OK** in the **Define Simple Bar: Values of Individual Cases** dialogue box.

The **SPSS Viewer** window will open, as shown in Figure 2.5.

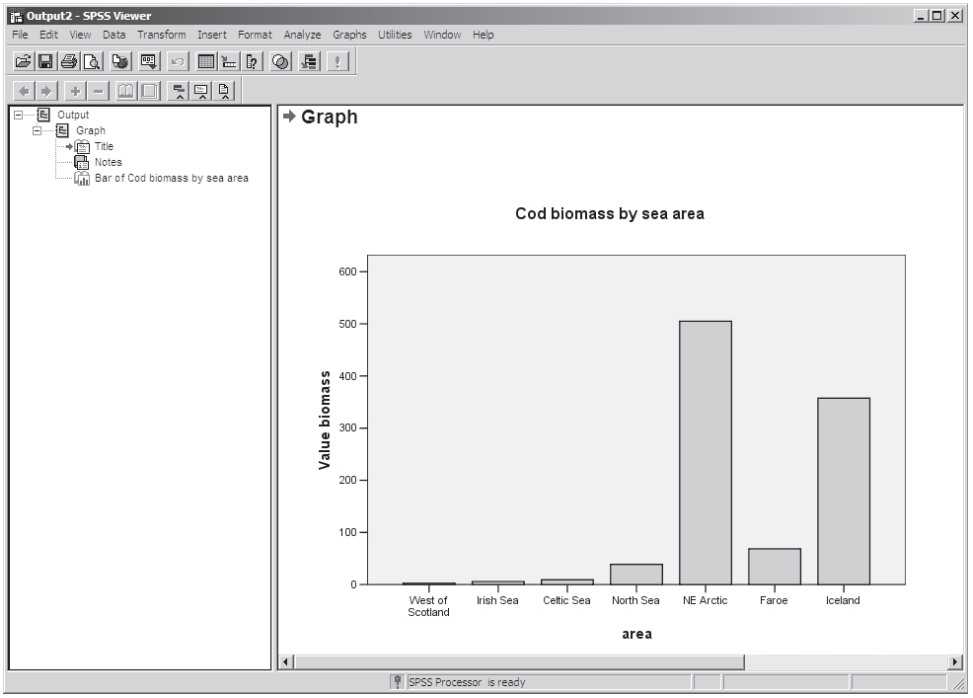

Figure 2.5 The **SPSS Viewer** window

All SPSS output appears in the **SPSS Viewer** window. The menu bar includes all the menus available in the **Data Editor**, plus two others — **Insert** and **Format**. The **SPSS Viewer** window has two panels. The right-hand panel contains the bar chart. This shows that the cod stocks in the North East Arctic and Iceland sea areas far outstripped those in the coastal sea areas of the British Isles in 2002. You will be using the left-hand panel of the **SPSS Viewer** window in Activity 2.7.

You may need to maximize the SPSS Viewer window in order to see all of the bar chart.

Activity 2.4 Editing a graph using Chart Editor

In this activity, you will learn how to change the appearance of the bar chart you produced in Activity 2.3.

◇ Place the mouse pointer anywhere on the bar chart and double-click. The **Chart Editor** window will open.

In general, within the **Chart Editor**, to alter the appearance of an item, you must first select it by placing the mouse pointer on it and clicking. Once selected, the item can be edited.

For example, the vertical axis is labelled `Value biomass`. A better label would be `biomass` (`thousand tonnes`). Make this change, as follows.

◇ Place the mouse pointer on the vertical axis label `Value biomass` and click once to select the label.

When an item is selected, it is surrounded by a coloured border.

◇ To edit the label, click a second time and the text of the label will be displayed horizontally.

◇ Delete the unwanted text and type in the new label.

◇ Press **Enter** and the new label will appear on the bar chart in the **Chart Editor** window.

You can make other changes if you wish. For example, to change the colour of the bars, double-click on one of the bars. The **Properties** dialogue box will open. Click on the **Fill & Border** tab, select your preferred colour (by clicking on it), then click on **Apply** and finally on **Close**.

If you wish to change the colour of a single bar, double-click on the bar. After the **Properties** dialogue box opens, click on the bar whose colour you wish to change (so that only this bar is selected). Then select your preferred colour, click on **Apply** and then on **Close**.

If you have time, spend a few minutes exploring the **Chart Editor**.

Once you have finished editing the chart, close the **Chart Editor** and the edited bar chart will appear in the **SPSS Viewer** window.

You can exit from the **Chart Editor** either by clicking on **Close** within the **File** menu of the **Chart Editor** or by clicking on the button marked **x** at the right-hand end of the title bar.

Activity 2.5 Saving output

SPSS output can be saved in a file: the file extension required is **.spo**. Save your output from Activity 2.4 in a file named **cod1.spo**, as follows.

◇ Choose **Save As...** from the **File** menu within the **SPSS Viewer** window to obtain the **Save As** dialogue box. (Note the folder name in the **Save in** field: the output file will be placed in this folder.)

◇ Enter `cod1` in the **File name** field and check that the **Save as type** field reads `Viewer Files (*.spo)`. If it does not, then select this option.

◇ Click on **Save**. The contents of the **SPSS Viewer** window will be saved in a file named **cod1.spo**.

◇ Now exit from SPSS.

2.2 Printing and pasting output

In this subsection, instructions are given for printing output or pasting it into a word-processor document. If the output you wish to print is saved in a file that is not already open, then you must first open the file, as described in Activity 2.6.

Activity 2.6 Opening an output file

Run SPSS now. Suppose that you wish to print the bar chart that you created in Activities 2.3 and 2.4, and then saved in the output file **cod1.spo** in Activity 2.5. There are two ways to open this file. Open the file using the following quick method.

◇ Click on **File**, move the mouse pointer to **Recently Used Files** in the **File** menu, and choose **cod1.spo** from the list that is displayed.

The **SPSS Viewer** window will open in exactly the same state as when you saved it. However, note that the data are not available: if you needed them, you would have to load them separately by opening the data file.

You will not need the data in this subsection.

This quick method only works for files that have been used recently. If you wish to open an output file that has not been used recently, then you should proceed as follows.

◇ Choose **Output...** from the **Open** submenu of the **File** menu. The **Open File** dialogue box will open.

◇ If necessary, navigate to the folder where the file is located. A list of output files in this folder will be displayed. Note that only names of output files (with the file extension **.spo**) are displayed when you use **Output...** from the **Open** submenu of **File**.

◇ Double-click on the name of the output file that you wish to open. The **SPSS Viewer** window will open in the state in which it was saved.

Activity 2.7 Selecting output for printing and export

In this activity, you will learn how to select output, for example a graph, in readiness for printing it, pasting it into a word-processor document, or saving it for future use.

Look at the panel on the left-hand side of the **SPSS Viewer** window. This shows the path structure of the **SPSS Viewer** window, with a record of the output you have generated. (At this point there is not much output. Being able to see the path structure is useful for keeping track of where you are in the **SPSS Viewer** window when you have undertaken several analyses.) Click on `Title`. A short red arrow will appear to the left of the word `Title`, and the panel on the right-hand side of the **SPSS Viewer** will scroll to the corresponding position (if required).

The `Notes` item is hidden and will not be used in this course. Items can be hidden or shown using the closed book and open book icons on the second toolbar.

To print or export output, you must first select the item you require. For example, to select the bar chart, click on it (in either the right-hand panel or the left-hand panel of the **SPSS Viewer**). A box enclosing the bar chart will appear on the right-hand panel. The box indicates that the bar chart has been selected.

You are now ready to print the selection, or paste it into a word-processor document. The instructions for doing this are given following this activity. You should read them now, then try printing and pasting the bar chart you have selected.

The instructions for printing and pasting output have been grouped below for ease of reference.

Printing output from the SPSS Viewer window

These instructions assume that SPSS is running and that the **SPSS Viewer** window is open.

- ◇ Select the item in the **SPSS Viewer** window that is to be printed.
- ◇ Choose **Print...** from the **File** menu (by clicking on it) to obtain the **Print** dialogue box.
- ◇ In the **Print range** area, click on **Selection** or on its radio button. (Warning: If you select **All visible output**, the entire contents of the right-hand panel of the **SPSS Viewer** window will be printed. You are advised not to select **All visible output** when printing, as this sometimes uses a lot of paper.)
- ◇ Click on **OK**.

Selecting an item for printing is described in Activity 2.7.

Pasting output from the SPSS Viewer window into a word processor document

These instructions assume that both SPSS and your word processor are running. The **SPSS Viewer** window and the document in which you wish to insert SPSS output should both be open.

These instructions work for Microsoft Word and many other word processors.

- ◇ Select the item in the **SPSS Viewer** window that is to be pasted into the word processor document.
- ◇ Choose **Copy** from the **Edit** menu (by clicking on it).
- ◇ Switch to your word processor (by clicking on the button corresponding to the word processor on the task bar).
- ◇ Place the cursor at the position in your document where you wish to insert the SPSS output.
- ◇ Finally, choose **Paste** from the **Edit** menu of your word processor (by clicking on it). The item you selected will be inserted in your document.

Alternatively, place the mouse pointer on your selection, click the right-hand mouse button, and choose **Copy** from the menu that is displayed (or press **Ctrl**+**C**).

Alternatively, press **Ctrl**+**V**.

2.3 Line plots and scatterplots

By the beginning of the 21st century, cod stocks in the North Sea had become severely depleted owing to over-fishing. In this subsection, line plots and scatterplots will be used to investigate the relationships between, and changes over time in, the catch, stocks and new recruits of cod in the North Sea. You will learn how to use SPSS to obtain such plots.

The data that will be used throughout this subsection are in the file **cod.sav**. Open this file now: a reminder of how to do this is given in the margin. From now on, instead of repeating detailed instructions, a reminder such as this one will often be given in the margin. In this case, the reminder indicates that you should choose **Data...** from the **Open** submenu of **File**. (Later, when an operation has been done several times, neither instructions nor a reminder will be given.)

Use **File** > **Open** > **Data...** as described in Activity 2.2.

Activity 2.8 Line plots in SPSS

There are four variables in the file **cod.sav**: `year`, `biomass`, `recruits` and `catch`. The data are the year, the spawning stock biomass (in thousands of tonnes), the estimated numbers of new recruits (in millions) and the annual catch (in thousands of tonnes) for North Sea cod.

Data are available on the catch for each year from 1963 to 1999, and on the biomass and recruits for each year from 1963 to 2003. Scroll down to the end of the data set: you will see that the last four values of `catch` are missing, as indicated by the dots in the cells for 2000 to 2003.

How has the North Sea cod catch varied over time? This can be investigated using a line plot of the annual catch by year. Line plots, which are called *line charts* in SPSS, are produced using **Line...** from the **Graphs** menu. Obtain a line plot of the North Sea cod catch, as follows.

- ◇ Choose **Line...** from the **Graphs** menu to obtain the **Line Charts** dialogue box. This is very similar to the **Bar Charts** dialogue box.
- ◇ A single plot is required, so select **Simple** (by clicking on it).
- ◇ The variable to be plotted is `catch`, and so does not need to be calculated. So, in the **Data in Chart Are** area, select **Values of Individual Cases** (by clicking on it or on its radio button).
- ◇ Click on **Define**. The **Define Simple Line: Values of Individual Cases** dialogue box will open.

This dialogue box is used to enter the variables to be plotted and to specify the title of the plot.

- ◇ Enter the variable `catch` in the **Line Represents** field.
- ◇ In the **Category Labels** area, select **Variable** (by clicking on it or its radio button) and enter the variable `year` in its field.
- ◇ Click on **Titles...** to open the **Titles** dialogue box.
- ◇ Enter a suitable title in the **Line 1** field of the **Title** area — for example, `Annual catch of North Sea cod`.
- ◇ Click on **Continue** to close the **Titles** dialogue box.
- ◇ Click on **OK**.

Entering variables was described in Activity 2.3.

The line plot will appear in the **SPSS Viewer** window. If you wish to edit the plot — for example, you might wish to change the vertical axis label from `Value catch` to `thousand tonnes` — then double-click on the graph to open the **Chart Editor**, and proceed as described in Activity 2.4. With this change of label, the graph will be as shown in Figure 2.6.

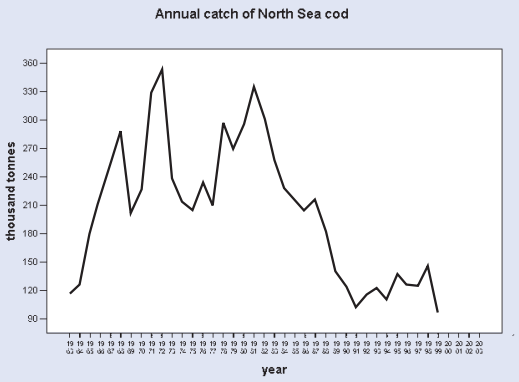

Figure 2.6 Annual catch of North Sea cod

The line plot shows that the annual catch declined substantially after 1980, and has remained at low levels since 1990. The catch was also low in the early 1960s.

Do not close the file **cod.sav**. You will need it in Activity 2.9.

Activity 2.9 *Multiple line plots*

The annual catch may be influenced by factors other than stock levels. A picture of the changes in the annual catch and the annual stock levels, and the relationship between them, can be obtained by producing line plots of the annual catch and the annual biomass for North Sea cod on a single diagram — that is, by producing a *multiple line plot*.

- ◇ Obtain the **Line Charts** dialogue box. Use **Graphs** > **Line...** .
- ◇ Since you wish to plot two lines on the same diagram, select **Multiple**.
- ◇ In the **Data in Chart Are** area, select **Values of Individual Cases**.
- ◇ Click on **Define**. The **Define Multiple Line: Values of Individual Cases** dialogue box will open.

Now proceed as follows. The procedure is similar to that described in Activity 2.3.

- ◇ Enter both the variables `catch` and `biomass` in the **Lines Represent** field.
- ◇ Enter `year` in the **Variable** field of the **Category Labels** area.
- ◇ Also specify a suitable title — for example, `North Sea cod: annual catch and biomass`. Use **Titles...** .
- ◇ Click on **Continue** (to close the **Titles** dialogue box), then on **OK** to produce the graph in the **SPSS Viewer** window.
- ◇ Place the mouse pointer anywhere on the graph and double-click to open the **Chart Editor**.

Now use the **Chart Editor** to edit the graph, as follows. The use of **Chart Editor** was discussed briefly in Activity 2.4.

- ◇ Replace the vertical axis label `Value` by the label `thousand tonnes`.

Now alter the labelling of the horizontal axis, as follows.

- ◇ Place the mouse pointer on one of the ticks on the horizontal axis and double-click to open the **Properties** dialogue box. (Alternatively, click on the large **X** in the toolbar of the **Chart Editor**.)
- ◇ Click on the **Labels & Ticks** tab. This panel offers a range of options for altering the labelling of the axis.
- ◇ In the **Major Increment Labels** area, click on the down arrow on the right of the **Label orientation** box, and select `Automatic` from the drop-down list that appears.
- ◇ Click on **Apply** and then on **Close**.
- ◇ Close the **Chart Editor**.

This will produce a graph in the **SPSS Viewer** window, as shown in Figure 2.7.

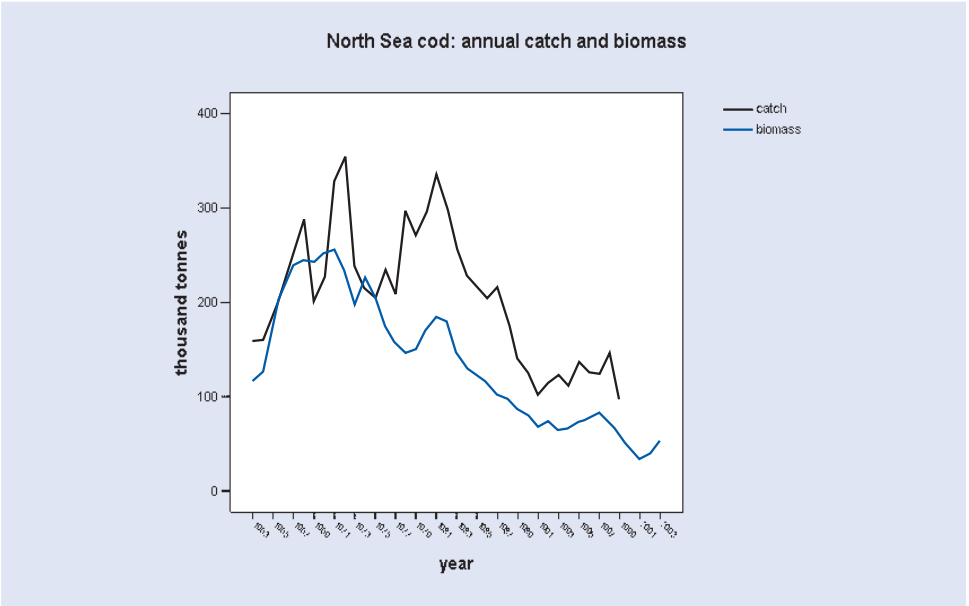

Figure 2.7 Annual North Sea cod catch and biomass

Figure 2.7 shows that the cod biomass began to decline in the early 1970s, before the annual catch began to decline, and that since then the annual catch has been greater than the biomass. This suggests that there is indeed a problem with over-fishing for cod in the North Sea.

Do not close the data file **cod.sav**. You will need it in Activity 2.10.

Activity 2.10 *New recruits of North Sea cod*

In this activity, you will investigate the relationship between the number of new cod recruits and the annual cod biomass by obtaining a scatterplot. Use the data in **cod.sav** and **Scatter/Dot...** from the **Graphs** menu to produce a scatterplot, as follows.

- ◇ Choose **Scatter/Dot...** from the **Graphs** menu. The **Scatter/Dot** dialogue box will open.
- ◇ A single scatterplot is required, so select **Simple Scatter** (by clicking on the corresponding graph).
- ◇ Click on **Define**. The **Simple Scatterplot** dialogue box will open.
- ◇ Enter the variable `recruits` in the **Y Axis** field, and the variable `biomass` in the **X Axis** field. Leave the other fields empty.
- ◇ Click on **Titles...** and enter a suitable title — for example, `North Sea cod: recruits and biomass`.
- ◇ Click on **Continue** to close the **Titles** dialogue box.
- ◇ Click on **OK**.

The scatterplot will appear in the **SPSS Viewer** window. This scatterplot can be edited using the **Chart Editor**. For example, in Figure 2.8, the points are plotted with a black fill.

Instructions for using **Chart Editor** are given in Activities 2.4 and 2.9.

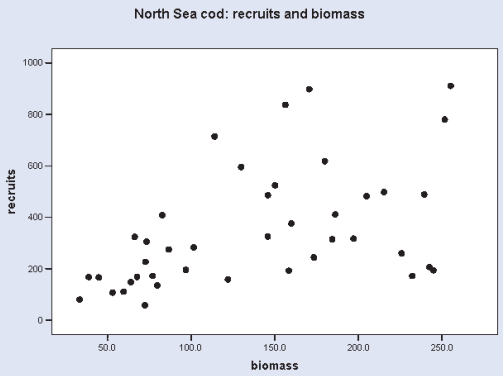

Figure 2.8 A scatterplot of recruits against biomass

For the herring data of Example 1.4, you saw that the number of recruits increases as the biomass increases, and that the variability in the number of new recruits increases with biomass. The scatterplot in Figure 2.8 suggests that this is also true for North Sea cod.

If you wish to save the graphs you have produced in this subsection, then save them now in an output file, before beginning Subsection 2.4.

Use **File** > **Save As...** within the **SPSS Viewer** window. See Activity 2.5.

2.4 Histograms and numerical summaries

In this subsection you will learn how to obtain histograms and numerical summaries in SPSS. The data used relate to levels of PCB contamination in North Sea cod. PCB is short for polychlorinated biphenyl. PCBs are pollutants that accumulate in the environment, and have been associated with a range of adverse health effects including cancer.

The data are in the data file **pcb.sav**. You will need this file for the two activities in this subsection.

Activity 2.11 PCBs in North Sea cod

Use **File** > **Open** > **Data...**.

Open the data file **pcb.sav**. (Note that in SPSS you can have only one data file open at a time, so if **cod.sav** is still open, opening **pcb.sav** will automatically close it.) There are two variables, `year` and `pcb`. The variable `pcb` gives the annual average PCB concentration in North Sea cod, measured in standard units, for most years between 1985 and 1994. Note that there is one missing value — for the year 1992.

Histograms are produced using **Histogram...** from the **Graphs** menu. Obtain a histogram of the PCB concentrations, as follows.

- ◇ Choose **Histogram...** from the **Graphs** menu. The **Histogram** dialogue box will open.
- ◇ Enter the variable `pcb` in the **Variable** field.
- ◇ Click on **Titles...** and enter a suitable title — for example, `PCB levels in North Sea cod, 1985-94`.
- ◇ Click on **Continue** to close the **Titles** dialogue box.
- ◇ Click on **OK**.

The **SPSS Viewer** window will open, displaying the histogram shown in Figure 2.9(a).

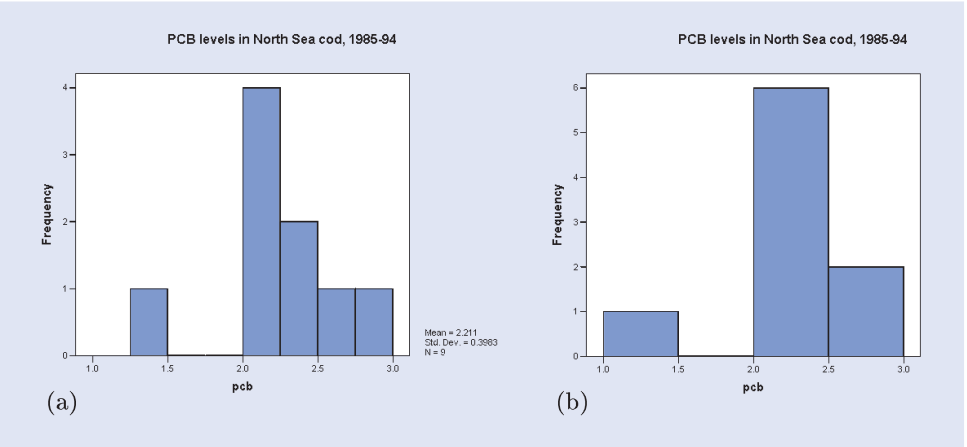

Figure 2.9 Two histograms of PCB levels in North Sea cod

Note that the sample mean, the sample standard deviation and the number of observations are given at the bottom right-hand side of the histogram.

The bin size and the boundaries of the bins on the default histogram can be changed using the **Chart Editor**. Edit the default histogram to obtain a histogram similar to the one in Figure 2.9(b), as follows.

- ◇ Place the mouse pointer on the graph and double-click to open the **Chart Editor**.
- ◇ Within the **Chart Editor**, double-click on one of the bars of the histogram. The **Properties** dialogue box will open.
- ◇ If necessary, click on the **Histogram Options** tab to bring the **Histogram Options** panel uppermost.

First, change the intervals (or bins) used to plot the histogram, so that the first bin starts at 1 and the bins have width 0.5, as follows.

- ◇ In the **Anchor First Bin** area, select **Custom value for anchor**, and change the value in its field to `1`. This sets the lower limit of the first interval (or bin) in the histogram.

- ◇ In the **Bin Sizes** area, select **Custom**. You can specify either the number of intervals or their width.
- ◇ Select **Interval width**, and change the value in its field to `0.5`.
- ◇ Click on **Apply** and then on **Close**.

Next, remove the numerical summaries, as follows.

- ◇ Place the mouse pointer on the numerical summaries next to the bottom right-hand corner of the histogram, and select them (by clicking on them).
- ◇ Delete the numerical summaries by pressing the delete key on your keyboard.

Now close the **Chart Editor**, and the histogram will be displayed in the **SPSS Viewer**, as shown in Figure 2.9(b).

Do not close the file **pcb.sav**. You will need it in Activity 2.12.

Activity 2.12 Numerical summaries

In Activity 2.11, you saw that SPSS calculates the sample mean and the sample standard deviation when drawing a histogram. There are several ways of obtaining numerical summaries directly in SPSS. One of these is described in this activity.

The data file **pcb.sav**, which you used in Activity 2.11, should still be open. If not, then open it now. Numerical summaries may be obtained using **Frequencies...** from the **Descriptive Statistics** submenu of **Analyze**. Obtain numerical summaries for the PCB concentrations, as follows.

- ◇ Click on **Analyze**, move the mouse pointer to **Descriptive Statistics** and choose **Frequencies...** from the **Descriptive Statistics** submenu. The **Frequencies** dialogue box will open.
- ◇ Enter the variable `pcb` into the **Variable(s)** field.
- ◇ Click on the **Statistics...** button. The **Frequencies: Statistics** dialogue box will open.

Summary statistics will be displayed if their check boxes are ticked. Make the following selections (by clicking on them or on their check boxes).

- ◇ In the **Central Tendency** area, select **Mean** and **Median**.
- ◇ In the **Dispersion** area, select **Std. deviation** and **Variance**.
- ◇ In the **Distribution** area, select **Skewness**.
- ◇ Click on **Continue** to return to the **Frequencies** dialogue box.
- ◇ In the **Frequencies** dialogue box, deselect **Display frequency tables** by clicking on it or on its check box. (These tables can be huge for large data sets.)
- ◇ Click on **OK**.

Measures of central tendency are measures of location.

Skewness will be explained shortly.

The following table will appear in the **SPSS Viewer** window.

Statistics

pcb

N	Valid	9
	Missing	1
Mean		2.211
Median		2.200
Std. Deviation		.3983
Variance		.159
Skewness		-.471
Std. Error of Skewness		.717

Notice that SPSS has reported that there are nine values and one missing value. The mean is 2.211 and the median is 2.200. The standard deviation is 0.3983 and the variance is 0.159.

You might like to check for yourself that the variance is the square of the standard deviation.

The (sample) **skewness** is a measure of departure from symmetry. If data are symmetrically distributed around the median, then the skewness is zero. If there is a long tail of values to the right of the median, then the data are said to be right-skew, or positively skewed, and the skewness is positive. Similarly, if there is a long tail to the left of the median, then the data are left-skew or negatively skewed. For the PCB data, the skewness is -0.471, indicating that the data are negatively skewed, but the skewness is not strong.

Also listed is the standard error of the skewness, which was not requested; you should ignore this. In common with many statistical packages, SPSS often gives more output than is strictly necessary (or required).

SPSS has many other features for presenting and summarizing data. For example, some numerical summaries can be obtained using **Descriptives...** from the **Descriptive Statistics** submenu of **Analyze**. If you have time, you might like to explore this facility.

Summary of Section 2

In this section, the statistical package SPSS has been introduced. Some of the facilities available within the **Data Editor** and **SPSS Viewer** windows have been described. You have learned how to print SPSS output or paste it into a word-processor document. Some of the facilities for producing the types of graphs and for calculating the summary statistics that were reviewed in Section 1 have been described, and applied to data on fish stocks.

3 Populations and models: health effects of air pollution

In Sections 1 and 2, the use of graphical and numerical summaries to represent variation was discussed. The variables considered, such as annual herring catch and PCB concentration in cod, are subject to random variation: they are called **random variables**. If a variable is continuous, it is said to be a **continuous random variable**. If it is discrete, it is said to be a **discrete random variable**. Random variables are generally represented by capital letters $X, Y, \ldots$, to distinguish them from the numerical values they take in particular samples of data, which are represented by lower case letters $x, y, \ldots$.

In this section, random variation is described using probability models. Typically, a probability model for a random variable involves a rule giving the probability with which each possible value of the variable will arise. The rule (usually a mathematical formula) may depend on parameters, which may be estimated from data.

In Subsection 3.1, some of the properties of random variables are described. Some models for continuous random variables and discrete random variables are discussed in Subsections 3.2 and 3.3, respectively. Some of the data used in this section relate to air quality. During the 1950s, London was infamous for its smog — a toxic combination of smoke and fog. The great smog of December 1952killed many thousand people; this disaster eventually led to the introduction of the Clean Air Acts which instituted smokeless zones and controlled industrial pollution. Although air quality has improved since the 1950s, other forms of air pollution, such as emissions from cars, have come to the fore. Smogs still occur: the London smog of 1991 is believed to have killed well over a hundred people.

Even in the absence of smog, air quality has an impact on health. Over recent decades, there has been a steady rise in the incidence of asthma in children. It has been suggested that this increase might be related to air quality, but so far the evidence for this is inconclusive. However, the relationship between air pollution and health remains an important topic of research.

In this section, two sets of data are used — one on air quality in central Nottingham, and one on admissions to a Nottingham hospital for asthma. The data on air quality were collected using an automatic monitoring device between 1 January 2000 and 30 June 2004. Air quality is described by the concentrations of several different pollutants. The data on asthma admissions to a hospital were collected over the same period as the air quality data.

These data sets are also used in Sections 4 and 5.

The air pollution data were obtained from the website of the Air Quality Archive (http://www.airquality.co.uk/archive/index.php) in September 2004. The asthma data were provided by Dr Richard Hubbard and Dr Joe West, University of Nottingham Medical School.

3.1 Samples and populations

In statistics, a key distinction is drawn between a *population* and a *sample* from that population. This distinction is illustrated in Example 3.1.

Example 3.1 Particulate matter in the air

Particulate matter comprises small particles of pollutants suspended in the air — smoke particles, for example. The PM_{10} level is the concentration of particles less than $10\,\mu$m in diameter, measured in $\mu g\,m^{-3}$. In this example, the logarithm of the PM_{10} level at a location in central Nottingham is considered.

$1\,\mu$m (1 micrometre) is one millionth of a metre. $1\,\mu g\,m^{-3}$ (microgram per cubic metre) is one millionth of a gram per cubic metre.

The average daily $\log PM_{10}$ level fluctuates from day to day around some average value. It is a continuous random variable, X say. Clearly, it is not possible to gather together all possible measurements on X that might conceivably occur. However, the distribution of X can be described approximately using a sample of values of X obtained on n different days — $x_1, x_2, \ldots, x_n$ say. The histogram in Figure 3.1, which is based on a sample of 1472 average daily $\log PM_{10}$ levels in central Nottingham, gives an approximate idea of the shape of the distribution of X.

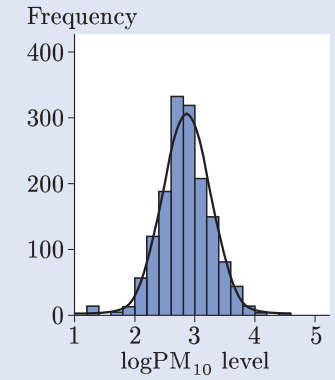

Figure 3.1 $\log PM_{10}$ levels in central Nottingham

Superimposed on the histogram is a smooth curve. This curve describes a probability model for X — that is, for the population of all values of X, not just those that were collected in the sample. If this probability model is correct then, as the sample size increases and the width of the bars in the histogram is reduced, the outline of the histogram should follow the curve with increasing accuracy. ♦

A probability model for a continuous random variable X is specified by the **probability density function** (or **p.d.f.**) of the random variable. The p.d.f. is defined for all possible values x of X, that is for all values x in the **range** of X. The p.d.f. $f(x)$, which cannot be negative, defines a curve. The total area under the curve defined by $f(x)$ is 1.

A graph of the p.d.f. corresponding to the probability model drawn in Figure 3.1 is shown in Figure 3.2. This is a scaled version of the curve in Figure 3.1, the scale being chosen so that the area under the curve is 1.

The probability model for a discrete random variable X is specified by the **probability mass function** (or **p.m.f.**) of the random variable:

$$p(x) = P(X = x).$$

The p.m.f. is defined for all values x in the range of X. It takes values between 0 and 1 ($0 < p(x) \leq 1$), and the sum of all its values is 1 (that is, $\sum p(x) = 1$).

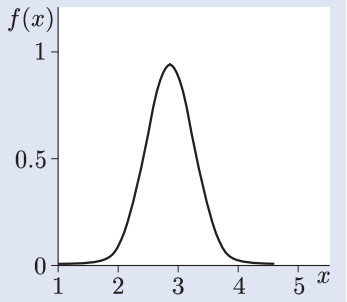

Figure 3.2 The p.d.f. for the probability model for $\log PM_{10}$ levels

Example 3.2 *Daily asthma admissions*

The number of persons admitted to a Nottingham hospital for asthma during the course of one day is a random variable X, say. Since X takes the discrete values $0, 1, 2, \ldots$, it is a discrete random variable. The bar chart in Figure 3.3(a) shows the distribution of a sample of values of X, collected on 1643 days.

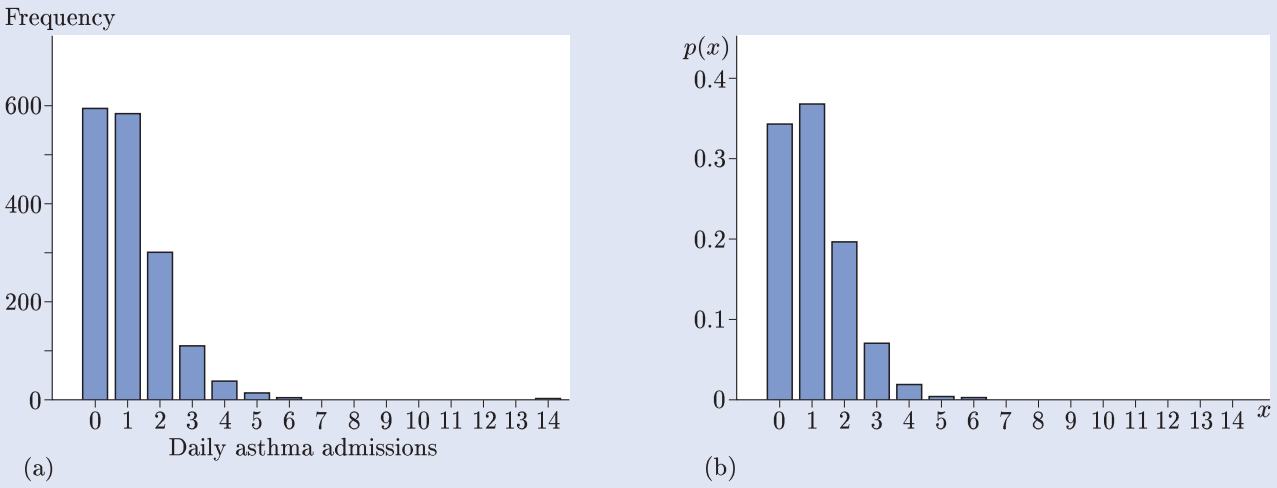

Figure 3.3 Daily admissions for asthma: (a) the data (b) a model

The p.m.f. of a probability model for X is shown in Figure 3.3(b). The vertical scales in Figure 3.3(a) and Figure 3.3(b) are different: the heights of the bars in Figure 3.3(a) sum to 1643, the number of observations, whereas the heights of the bars in Figure 3.3(b) sum to 1. However, the shapes of the two plots are similar. If the probability model describes the variation in X correctly, then any differences in shape between Figure 3.3(a) and Figure 3.3(b) are due to chance effects in the particular sample represented in Figure 3.3(a). ♦

Activity 3.1 *A probability model for asthma admissions*

Several values of the p.m.f. for the probability model in Figure 3.3(b) are given in Table 3.1.

This probability model is discussed in Subsection 3.3.

Table 3.1 A probability model for asthma admissions

x	0	1	2	3	4	...
$p(x)$	0.342	0.367	0.197	0.070	0.019	

(a) Calculate the value of each of the probabilities $P(X \geq 5)$, $P(X \leq 2)$ and $P(X > 2)$.

(b) According to this probability model, on what percentage of days might you expect there to be at least one admission to hospital for asthma?

In Section 1, numerical summaries such as the mean, the median, and the standard deviation were introduced. These are all sample quantities, that is, values calculated from a sample. Corresponding to these sample quantities are population summaries. The population mean, variance and standard deviation are defined in the following box.

The population mean, variance and standard deviation

The **mean** μ and the **variance** σ^2 of a discrete random variable X with probability mass function $p(x)$ are given by

$$\mu = E(X) = \sum_x xp(x),$$

$$\sigma^2 = V(X) = E\left[(X-\mu)^2\right] = \sum_x (x-\mu)^2 p(x),$$

where the sums are taken over all values x in the range of X.

The symbol μ is a Greek letter pronounced 'mew'. The Greek letter σ is pronounced 'sigma'.

The **mean** μ and the **variance** σ^2 of a continuous random variable X with probability density function $f(x)$ are given by

$$\mu = E(X) = \int_X xf(x)\,dx,$$

$$\sigma^2 = V(X) = E\left[(X-\mu)^2\right] = \int_X (x-\mu)^2 f(x)\,dx,$$

where the integrals are taken over all values x in the range of X.

The notation $\int_X \ldots dx$ represents an integral. This notation has been included as you may meet it elsewhere. No knowledge of integrals, or calculus, is required in M249.

For both continuous and discrete random variables X, the **standard deviation** of X is σ, the square root of the variance.

The notation $E(X)$ is read 'the expectation of X', or 'the expected value of X'. The population mean is also called the **expectation**, or **expected value**.

In Subsection 1.2 the sample median of a data set was defined to be the middle value (or halfway between the two middle values) when the values are placed in order of increasing size. So, roughly speaking, about half of the values are below the median and about half are above the median. The population median may be defined in an analogous way. More generally, it is convenient to describe random variables in terms of their **quantiles**, a particular example of which is the median. In this course, you will only need the quantiles of *continuous* random variables.

The quantiles of a continuous random variable

If X is a continuous random variable with probability density function $f(x)$, and $0 \leq \alpha \leq 1$, then the α-quantile of X is the value q_α such that

$$\alpha = P(X \leq q_\alpha).$$

The (population) **median** of X is the 0.5-quantile of X. The **lower quartile** of X is the 0.25-quantile. The **upper quartile** of X is the 0.75-quantile.

The symbol α is a Greek letter pronounced 'alpha'.

For a continuous random variable X, the probability $P(X \leq x)$ is the area under the curve of the probability density function to the left of x. This is illustrated in Figure 3.4(a).

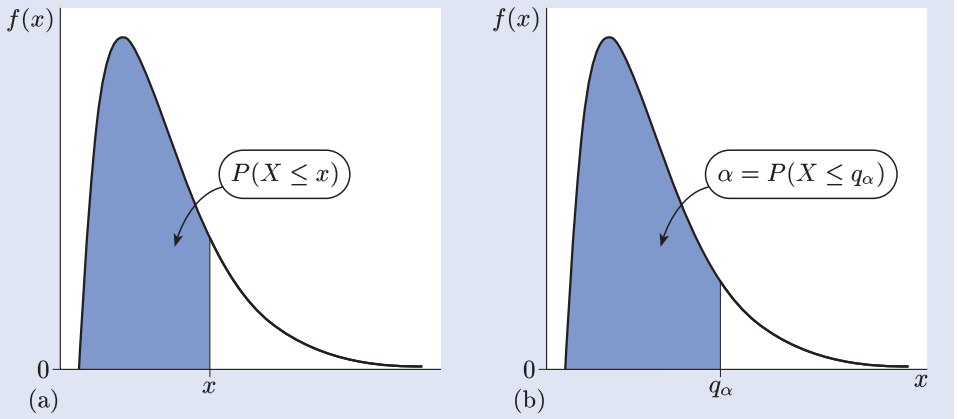

Figure 3.4 (a) The probability $P(X \leq x)$ for a random variable X with p.d.f. $f(x)$ (b) The α-quantile of X

So if $P(X \leq x) = \alpha$, then $x = q_\alpha$, the α-quantile of X. That is, the area to the left of q_α, the α-quantile of X, is α. This is illustrated in Figure 3.4(b).

Example 3.3 Calculations with the quantiles of the logPM$_{10}$ levels

The median, the lower quartile, and the upper quartile for the logPM$_{10}$ levels, based on the probability model shown in Figure 3.2, are shown in Figure 3.5.

The lower quartile $q_{0.25}$ is approximately 2.59. Thus the probability that the logPM$_{10}$ level is 2.59 or lower is 0.25. Similarly, the upper quartile $q_{0.75}$ is 3.13, so the probability that the logPM$_{10}$ level is 3.13 or lower is 0.75.

These quantiles can be used to obtain other probabilities. For example, the probability that the logPM$_{10}$ level lies above 3.13 is

$$\begin{aligned} P(X > q_{0.75}) &= 1 - P(X \leq q_{0.75}) \\ &= 1 - 0.75 \\ &= 0.25. \end{aligned}$$

Similarly, the probability that the logPM$_{10}$ level lies between 2.59 and 3.13 is

$$\begin{aligned} P(q_{0.25} < X \leq q_{0.75}) &= P(X \leq q_{0.75}) - P(X \leq q_{0.25}) \\ &= 0.75 - 0.25 \\ &= 0.5. \end{aligned}$$

So the logPM$_{10}$ level lies between 2.59 and 3.13 on about one day out of every two. ♦

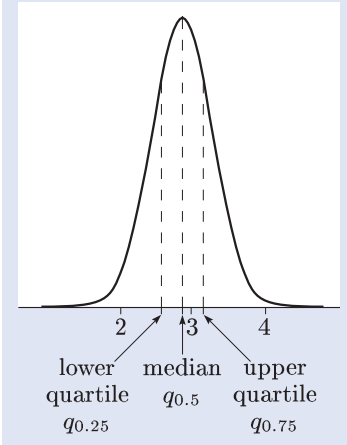

Figure 3.5 The median and quartiles of the model for logPM$_{10}$ levels

Activity 3.2 Population quantiles

The p.d.f. of a continuous random variable X, with three quantiles q_A, q_B and q_C marked, is shown in Figure 3.6.

(a) The three quantiles marked on Figure 3.6 are the 0.2-quantile, the 0.5-quantile, and the 0.9-quantile. Which of these quantiles is q_A? Which is q_B, and which is q_C?

(b) Mark on Figure 3.6 the approximate locations of the lower and upper quartiles of X.

(c) The range of X is to be partitioned into three intervals, in such a way that, for each interval, the probability that X takes a value in the interval is $\frac{1}{3}$. Which quantiles should be used to specify the boundaries of the middle interval?

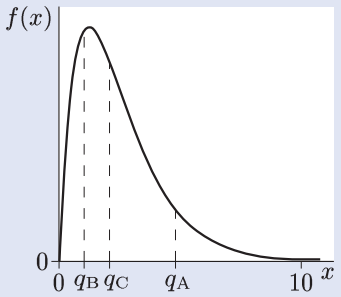

Figure 3.6 Quantiles of a continuous random variable

Finally, for a discrete random variable, the **mode** is the value that has the highest probability of occurring, if there is just one such value. For a continuous random variable, a **mode** corresponds to a local maximum of the p.d.f. For example, the mode of the random variable with the p.d.f. shown in Figure 3.6 is 1.25. (If a p.d.f. has more than one maximum point, then the random variable has more than one mode.)

3.2 Probability models for continuous random variables

Probability models for random variables often involve one or more parameters. Different values of these parameters give different p.d.f.s. A collection of p.d.f.s, indexed in this way by one or more parameters, is called a **family** of probability models. Three families of models for continuous random variables are reviewed briefly in this subsection: the families of normal, exponential and continuous uniform distributions.

Normal distributions

Perhaps the most commonly used family of probability models for continuous random variables is the family of **normal distributions**. This family is indexed by two parameters, the mean μ and the variance σ^2 (or alternatively, the standard deviation σ). A random variable X with such a probability distribution is said to be **normally distributed**, and this is written $X \sim N(\mu, \sigma^2)$.

The p.d.f. of a normal distribution is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right), \quad -\infty < x < \infty.$$

You do not need to remember this expression.

The p.d.f.s of three normal distributions are shown in Figure 3.7.

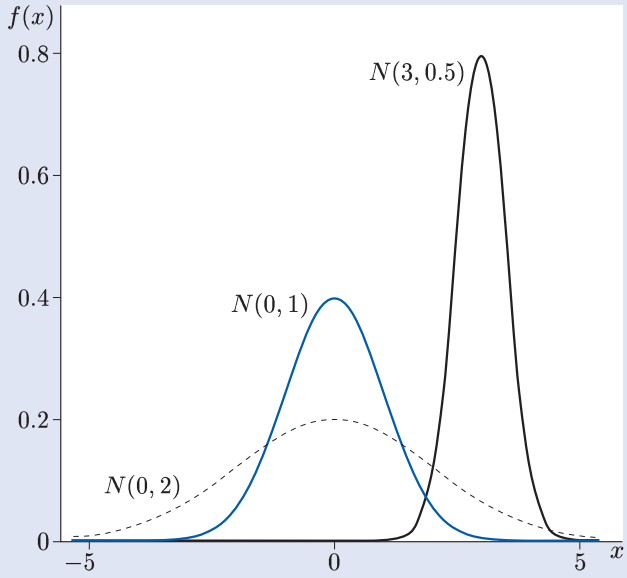

Figure 3.7 The p.d.f.s of three normal distributions

Two key features of a normal distribution are that it is symmetric with a single mode at the mean μ. Changing the value of the mean μ moves the position of the mode to the left or right along the horizontal axis, while changing the value of the standard deviation σ changes the spread of the distribution. A normal distribution is often a sensible probability model given data on a continuous random variable that are clustered symmetrically around a central peak.

There are other distributions that may be sensible for such variables.

Example 3.4 Normal distributions

The probability model suggested for the logarithms of the PM_{10} levels in Example 3.1 is a normal distribution. (See Figure 3.2.) The mean and standard deviation of this distribution were estimated from the data: the mean was taken to be 2.86, and the standard deviation 0.402, so the model used was $N(2.86, 0.402^2)$. ♦

All normal distributions share the same basic shape. So, for instance, whatever the values of the parameters μ and σ^2, values more than 1.96 standard deviations away from the mean arise with probability 0.05. Hence a reference normal distribution, the **standard normal** distribution, which has mean $\mu = 0$ and standard deviation $\sigma = 1$, is often used, and selected quantiles are tabulated in books of statistical tables. The letter Z is used to denote the standard normal random variable: $Z \sim N(0, 1)$.

Exponential distributions

A family of distributions that have a completely different shape from the normal distributions is the family of **exponential** distributions. This family is indexed by a single parameter, the **rate** λ. The p.d.f. of an exponential distribution is given by

The Greek letter λ is pronounced 'lambda'.

$$f(x) = \lambda e^{-\lambda x}, \quad x \geq 0.$$

The p.d.f.s of three exponential distributions are shown in Figure 3.8.

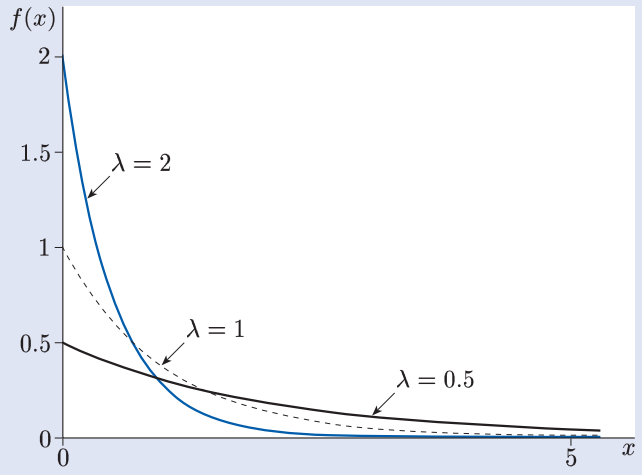

Figure 3.8 The p.d.f.s of three exponential distributions

Note that, whatever the value of the parameter λ, the p.d.f. has its maximum at $x = 0$, and hence the mode is zero.

When a random variable X has an exponential distribution with parameter λ, this is written $X \sim M(\lambda)$. Note that a random variable X with an exponential distribution takes only non-negative values.

When events occur randomly in time, an exponential distribution often provides a suitable probability model for the time intervals between successive events.

The letter M stands for Markov, after Andrei Andreyevich Markov (1856–1922), a Russian mathematician after whom several random processes are named, including Markov chains which you will meet in Book 4 *Bayesian statistics*.

Example 3.5 Duration of hospital stays

The numbers of daily admissions for asthma to a Nottingham hospital were discussed in Example 3.2. There were 1762 admissions. Each person admitted to hospital remained there until he or she was discharged. A histogram (with interval width one day) of the number of days between admission and discharge is shown in Figure 3.9. The histogram is not symmetric, so a normal model is not appropriate for these data. Most hospital stays are of short duration, and for durations longer than two days the frequency declines rapidly as the duration increases. So perhaps an exponential model would be more appropriate. ♦

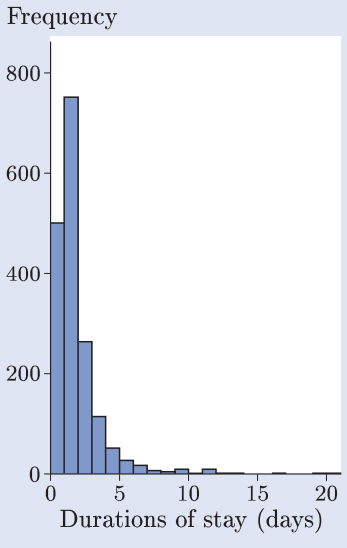

Figure 3.9 Times between admission and discharge

The mean and variance of a random variable X with an exponential distribution with parameter λ are given by

$$E(X) = \frac{1}{\lambda}, \quad V(X) = \frac{1}{\lambda^2}.$$

Thus the mean and standard deviation of X are equal. This fact can be used to help decide whether an exponential distribution is a suitable probability model. This is illustrated in Activity 3.3.

Activity 3.3 Is an exponential model appropriate?

For the hospital stays data represented in Figure 3.9, the mean and standard deviation are 1.885 and 1.850, respectively. Use this information, and the shape of the histogram in Figure 3.9, to identify one reason why an exponential model might be appropriate for these data, and one reason why an exponential model might not be appropriate.

Continuous uniform distributions

The third family of probability models reviewed in this subsection is the family of continuous uniform distributions. A random variable X, defined on an interval $[a, b]$, is said to have the **continuous uniform** distribution if its p.d.f. is $f(x) = 1/(b-a)$, for $a \leq x \leq b$. This is written $X \sim U(a, b)$. All values of X in the interval $[a, b]$ are equally likely.

Activity 3.4 Adults with asthma

It is claimed that the age at admission to hospital for asthma of adults aged between 20 and 50 years is uniformly distributed: $U(20, 50)$.

(a) Sketch the p.d.f. of this distribution.

(b) Data on the exact age of patients aged between 20 and 50 years who are admitted for asthma are available. Given the data, what type of graph might you use to investigate the claim that the uniform probability model is appropriate for describing the variation in the age of such patients?

The families of probability models reviewed in this subsection are summarized in the following box.

Probability models for continuous random variables

The random variable X is normally distributed with mean μ and variance σ^2 if the p.d.f. of X is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right), \quad -\infty < x < \infty.$$

This is written $X \sim N(\mu, \sigma^2)$. The normal distribution with mean 0 and variance 1 is called the **standard normal distribution**. The letter Z is used to denote the standard normal random variable: $Z \sim N(0, 1)$.

If the random variable X has an **exponential distribution** with parameter λ, its p.d.f. is given by

$$f(x) = \lambda e^{-\lambda x}, \quad x \geq 0.$$

This is written $X \sim M(\lambda)$. Its mean is $1/\lambda$ and its variance is $1/\lambda^2$.

A random variable X with the **continuous uniform distribution** on the interval $[a, b]$ has p.d.f. given by

$$f(x) = \frac{1}{b-a}, \quad a \leq x \leq b.$$

This is written $X \sim U(a, b)$.

3.3 Probability models for discrete random variables

The continuous uniform distribution has a discrete counterpart, the **discrete uniform** distribution. The family of discrete uniform distributions is indexed by a single parameter n. The discrete uniform distribution with parameter n is a model for a random variable X that can take values on a discrete set of n points, which are usually labelled $1, 2, \ldots, n$ for convenience. The p.m.f. of X is

$$p(x) = P(X = x) = \frac{1}{n}, \quad x = 1, 2, \ldots, n.$$

So the n possible outcomes are all equally likely.

Often there are just two possible outcomes. For example, a child might or might not be admitted to hospital with asthma on a particular day; and a person admitted to hospital with asthma will be male or female. In this case, the random variable X is said to be **binary**, and the outcomes are often labelled 0 and 1.

The discrete uniform distribution on $\{0, 1\}$ is very restrictive, as it implies that both outcomes have probability 0.5. A more general and more useful probability model for a binary random variable X is the **Bernoulli** distribution. The family of Bernoulli distributions is indexed by a single parameter p, where $p = P(X = 1)$; p is called the **Bernoulli probability** or **success probability**. If a random variable X has a Bernoulli distribution with parameter p, this is written $X \sim \text{Bernoulli}(p)$. The mean and variance of X are given by

$$E(X) = p, \quad V(X) = p(1 - p).$$

$\{0, 1\}$ is mathematical notation for the set containing the numbers 0 and 1.

The Bernoulli model is useful primarily as a building block for other models. The most important of these models is the **binomial distribution**. This is introduced in Example 3.6.

Example 3.6 Gender and asthma

Information on the gender distribution of persons with asthma can help to inform strategies for controlling the disease. Records of the hospital admissions for asthma described in Example 3.2 include data on the gender of persons with asthma. Over the data collection period, 1762 persons were admitted for asthma, and records on their gender were available for 1761 persons.

The gender of a person admitted to hospital for asthma can be represented by a binary random variable X with a Bernoulli(p) distribution as follows: $X = 1$ if the case is female, $X = 0$ if the case is male, and $p = P(X = 1)$ is the probability that the case is female.

Let X_i denote the gender of the ith person admitted. Then the total number of females among a sample of 1761 admissions is a random variable R, where

$$R = X_1 + X_2 + \cdots + X_{1761}.$$

We could just as well have defined a random variable X by $X = 1$ if a person is male and $X = 0$ if a person is female. Then R would be the number of males in a sample of 1761 admissions.

Clearly, R can take any of the values $0, 1, 2, \ldots, 1761$. It is a discrete random variable with range $\{0, 1, 2, \ldots, 1761\}$. However, its distribution is not uniform since the values are not all equally likely. For example, if an asthma case is equally likely to be male or female, so that $P(X = 1) = P(X = 0) = 0.5$, then you would expect about half the admissions to be female. So, for example, the values $R = 880$ or $R = 881$ are more likely than the value $R = 0$.

Provided that the gender of one person admitted does not influence the gender of another (which is a reasonable assumption in this case), the random variable R has the binomial distribution with parameters 1761 and p; this is written $R \sim B(1761, p)$.

Of the asthma cases admitted to the Nottingham hospital, 907 were female and 854 were male. So a sensible estimate of the success probability p is $907/1761 \simeq 0.515$. ♦

The term **Bernoulli trial** is used to describe a single statistical experiment for which there are just two possible outcomes. So, in Example 3.6, whether or not a patient admitted to hospital for asthma is male or female is a Bernoulli trial. In Example 3.6, it was assumed that the outcome of one Bernoulli trial did not influence the outcome of another; that is, it was assumed that the Bernoulli trials were **independent**. The corresponding Bernoulli random variables are said to be **independent**.

In general, a random variable X is said to have a **binomial distribution with parameters n and p**, written $X \sim B(n, p)$, if it is the sum of n independent Bernoulli random variables each with success probability p. The random variable X is discrete, and takes values in $\{0, 1, 2, \ldots, n\}$. Its p.m.f. is

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, 2, \ldots, n.$$

The bracketed term at the front is read 'n C x', or alternatively as 'n choose x'. It is defined as follows:

$$\binom{n}{x} = \frac{n!}{x!(n-x)!},$$

where $x! = 1 \times 2 \times \cdots \times x, \quad 0! = 1.$

$x!$ is read 'x factorial'. Much use will be made of the binomial model in Book 1 *Medical statistics*, but you will not be required to calculate binomial probabilities.

The binomial distribution, $B(n, p)$, provides a probability model for the total number of successes in a sequence of n independent Bernoulli trials, in which the probability of success in a single trial is p.

The mean and variance of $X \sim B(n, p)$ are

$$E(X) = np, \qquad V(X) = np(1-p).$$

The p.m.f.s of three binomial distributions are shown in Figure 3.10.

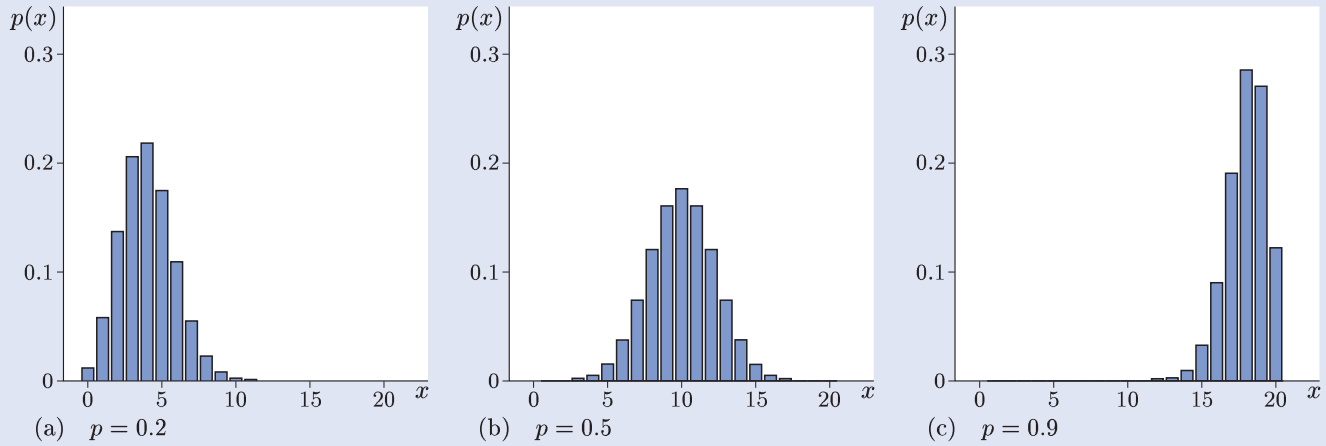

Figure 3.10 Three binomial models with $n = 20$

A binomial distribution has a single mode. It is right-skew when the parameter $p < 0.5$, left-skew when $p > 0.5$, and symmetric when $p = 0.5$. The binomial model is very commonly used to represent data on the number of individuals with a particular attribute in a sample of given size.

Activity 3.5 Severity of asthma cases

One way to classify the severity of cases of asthma admitted to hospital is by their length of stay. For example, a case might be regarded as severe if one or more days in hospital are required. Of the 1762 admissions for asthma at the Nottingham hospital, 500 were admitted for less than a day.

(a) Identify a suitable model for R, the number of patients admitted for less than a day in a sample of 1762 admitted for asthma. Use the data for the Nottingham hospital to estimate the values of the parameters.

(b) Calculate the mean and variance of R using the model you specified in part (a).

Both the discrete uniform distribution and the binomial distribution have a finite range — $\{1, \ldots, n\}$ for the discrete uniform model and $\{0, 1, \ldots, n\}$ for the binomial model. The distribution described in Example 3.7 has an unbounded range $\{0, 1, 2, \ldots\}$.

Example 3.7 Daily number of asthma admissions

In the Nottingham hospital at which the data used in this section were collected, 1762 persons were admitted on 1643 days, so on average 1.072 persons were admitted per day. However, the mean number of admissions per day does not provide any information about the variation in the number of admissions from day to day. Was just one person admitted on most days? Or were all 1762 persons admitted on the same day?

A bar chart of the data is shown in Figure 3.11.

On most days, the number of admissions was 0, 1 or 2, but there were some days with more than two admissions. On one day, 14 patients were admitted (this is barely visible on Figure 3.11). This is the sort of pattern that might be expected if all individuals in a large population have the same low probability of being admitted to hospital with asthma. There is no pre-determined maximum number of admissions, but numbers very much larger than the mean are exceptional, though not impossible. ♦

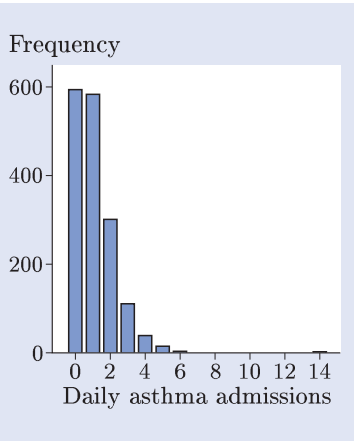

Figure 3.11 Daily admissions for asthma

The characteristics of the distribution described in Example 3.7 are typical of a member of a family of distributions called **Poisson distributions**. This family of distributions is indexed by a single parameter μ. The parameter μ of a Poisson distribution is the mean of the distribution. If a random variable X has the Poisson distribution with mean μ, this is written $X \sim \text{Poisson}(\mu)$. The p.m.f. of X is

$$p(x) = \frac{\mu^x e^{-\mu}}{x!}, \quad x = 0, 1, 2, \ldots .$$

The mean and variance of X are given by

$$E(X) = \mu, \quad V(X) = \mu.$$

So both the mean and variance of a Poisson distribution are equal to μ. The Poisson model is often used to represent counts of independently occurring events.

Activity 3.6 A model for admissions

In Table 3.1, a probability model for the number of asthma admissions per day was suggested. This model was based on the Poisson distribution with mean 1.072.

(a) For $X \sim \text{Poisson}(1.072)$, calculate the value of $p(x)$ for $x = 0, 1, 2$. Check that these values are the same as those given in Table 3.1.

(b) The variance of the number of admissions is 1.285. Explain whether or not, in your view, the Poisson model is adequate.

The main properties of the three discrete distributions described in this subsection are summarized in the following box.

Probability models for discrete random variables

If a random variable X has a **binomial distribution** with parameters n and p, then its p.m.f. is given by

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, 2, \ldots, n.$$

This is written $X \sim B(n, p)$. The mean of X is np and the variance is $np(1-p)$.

If a random variable X has a **Poisson distribution** with parameter μ, then its p.m.f. is given by

$$p(x) = \frac{\mu^x e^{-\mu}}{x!}, \quad x = 0, 1, 2, \ldots .$$

This is written $X \sim \text{Poisson}(\mu)$. The mean and variance of X are both equal to μ.

If a random variable X has the **discrete uniform distribution** on $\{1, 2, \ldots, n\}$, then its p.m.f. is

$$p(x) = \frac{1}{n}, \quad x = 1, 2, \ldots, n.$$

Summary of Section 3

In this section, the distinction between populations and samples has been discussed. Population summaries, including the mean, variance and standard deviation, have been reviewed. Quantiles for continuous random variables have been defined, including the median and the quartiles. A range of commonly used probability models for continuous and discrete data have been presented, including the normal, exponential, continuous uniform, discrete uniform, Bernoulli, binomial and Poisson distributions.

Exercises on Section 3

Exercise 3.1 Calculating probabilities

The p.m.f. of a discrete random variable X is given in Table 3.2. Calculate the values of the probabilities $P(X \leq 2)$ and $P(X > 0)$.

Table 3.2 The p.m.f. of X

x	0	1	2	3
$p(x)$	0.1	0.2	0.4	0.3

Exercise 3.2 Quantiles of continuous random variables

A continuous random variable X has lower quartile 2.3, median 4.6 and upper quartile 6.2. Decide whether each of the following statements is true or false, giving reasons for your answers.

(a) $P(X > 6.2) = 0.25$.

(b) $q_{0.1} < 4.6$.

(c) $q_{0.8} \leq 2.3$.

Exercise 3.3 *Choosing a probability model*

Figure 3.12 shows histograms of observations on two continuous random variables X and Y defined on the interval $[0, 20]$.

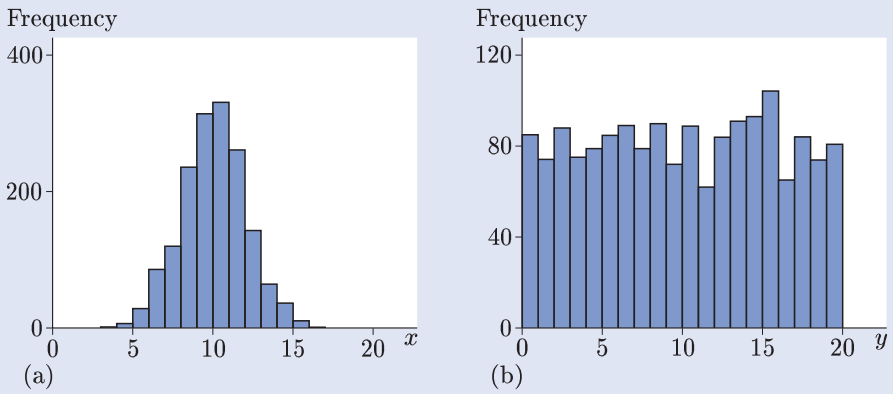

Figure 3.12 Two histograms: (a) observations on X, (b) observations on Y

Suggest a plausible probability model for each variable, giving reasons for your choice in each case.

4 From samples to populations: asthma and air quality

Researchers usually seek to draw general conclusions about a particular topic on the basis of observations. In statistical terms, they are interested in populations, rather than samples. For example, ecologists might be interested in understanding the relationship between fishing and fish stocks in the North Sea, not just in describing what happened in the particular years for which data have been collected; and public health doctors might be interested in whether there is a link between air pollution and asthma, not just between air quality and asthma hospital admissions in a particular area.

The process of drawing conclusions about populations on the basis of samples of data from those populations is called **statistical inference**, and is the subject of this section. In Subsection 4.1, the *central limit theorem* is discussed briefly. This is one of the fundamental results of statistical theory. Given a large sample of data, it can be used to infer information about the population from which the sample is drawn. This result underpins the methods described in Subsections 4.2 and 4.3. In Subsection 4.2, the central limit theorem is used to obtain a plausible range of values for a population parameter, given a sample of data; this range of values is a *confidence interval* for the parameter. *Significance tests* are used to evaluate the evidence against a particular hypothesis. In Subsection 4.3, a test based on the central limit theorem is used to illustrate the ideas involved in significance testing.

The approach to statistical inference reviewed here is a classical, or frequentist, approach. In Book 4 *Bayesian statistics* you will learn about a different approach to statistical inference.

4.1 Samples and estimates

Much of statistical inference proceeds by taking averages of all the values in a sample. Averages have simple statistical properties which can be exploited to make inferences about population parameters. The central limit theorem concerns the means of large samples. Before stating the theorem, some notation and terminology will be introduced.

In order to distinguish between an estimate and the population parameter being estimated, an estimate is denoted by writing a 'hat' over the population parameter. For example, if μ is a population mean, then $\widehat{\mu}$ is used to denote an estimate of μ; if p is a proportion or probability, then $\widehat{p}$ is used to denote an estimate of p. For obvious reasons, this notation is called the **hat notation**.

$\widehat{\mu}$ is read 'μ-hat' and $\widehat{p}$ is read 'p-hat'.

More generally, suppose that θ is a population parameter of interest, and that $\widehat{\theta}$ is an estimate of θ obtained from a sample of size n (say). The estimate $\widehat{\theta}$ is calculated using some procedure or estimating formula. (For example, if the parameter is the population mean, then $\widehat{\theta}$ is the sample mean, which is calculated by adding together the values in the sample and dividing by the sample size.) Different samples of size n will in general lead to different estimates, so the estimating formula is a random variable. This estimating formula is called an **estimator** for θ. The hat notation is also used for an estimator. So $\widehat{\theta}$ is used to denote both a random variable expressing an estimating formula and an estimate obtained from a particular sample of data.

Now consider a population with mean μ and variance σ^2. Suppose that a sample of size n is taken from this population and that the sample values are selected independently. An estimate of μ is given by the sample mean $\overline{x}$: $\widehat{\mu} = \overline{x}$. This estimate is an observation on the estimator $\widehat{\mu} = \overline{X}$. Since the estimator $\widehat{\mu}$ is a random variable, it has a distribution. This distribution is called the **sampling distribution** of the mean; it is the distribution of the means of all possible samples of size n from the population. The mean and variance of this distribution are given by

$$E(\widehat{\mu}) = \mu, \quad V(\widehat{\mu}) = \frac{\sigma^2}{n}.$$

Furthermore, it can be shown that, for large n, the probability distribution of $\widehat{\mu}$ — that is, the sampling distribution of the mean — is approximately normal. These results together comprise the central limit theorem, which is stated formally in the following box.

The central limit theorem

If n independent random observations are taken from a population with mean μ and finite variance σ^2, then for large n the distribution of their mean $\widehat{\mu}$ is approximately normal with mean μ and variance σ^2/n:

$$\widehat{\mu} \approx N\left(\mu, \frac{\sigma^2}{n}\right).$$

The symbol $\approx$ is read 'has approximately the same distribution as'.

The central limit theorem underpins all the statistical methods described in Subsections 4.2 and 4.3. Note that it applies to the mean of both discrete and continuous random variables.

The standard deviation of the sampling distribution of the mean, which is equal to $\sigma/\sqrt{n}$, is called the **standard error** of $\widehat{\mu}$. The standard error may be estimated by substituting the sample standard deviation s in the expression $\sigma/\sqrt{n}$. So the estimated standard error is $s/\sqrt{n}$.

Activity 4.1 Daily asthma admissions

For the data on the number of admissions for asthma on each of 1643 days discussed in Example 3.7, the sample variance is 1.285.

(a) Estimate the standard error of the mean daily number of admissions.

(b) Figure 4.1 shows two graphs. One represents the probability distribution of the daily number of admissions for asthma; the other represents the sampling distribution of the mean number of daily admissions for samples of size 1643. Which graph represents the probability distribution, and which represents the sampling distribution of the mean? Explain your answer.

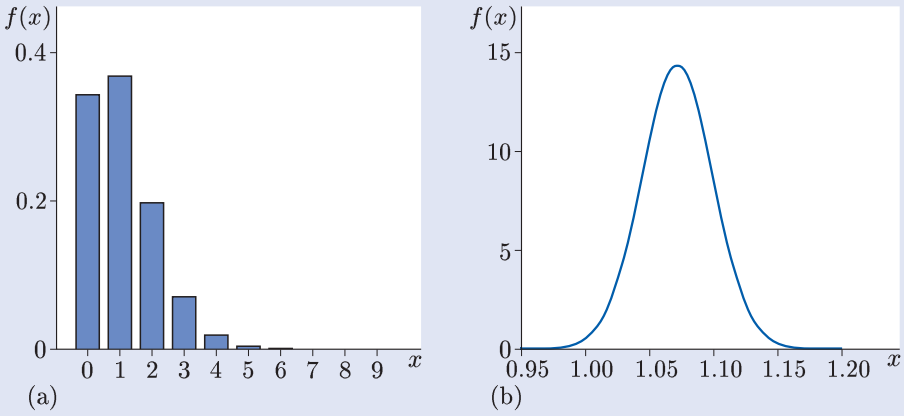

Figure 4.1 Two distributions relating to asthma admissions

4.2 Confidence intervals

Different samples lead to different estimates, so parameter estimates are subject to **sampling error**. The word 'error' here refers to the difference between the estimate and the true value, and does not signify that a mistake has been made. The central limit theorem makes it possible to quantify the likely size of the sampling error.

Example 4.1 Ozone levels

Ozone is a form of oxygen that forms a thin layer at high altitude. This layer shields the Earth's surface from ultraviolet sunlight. However, at ground level, ozone is a toxic pollutant: it is an important constituent of smog.

The average concentration of ozone (in parts per billion) in Nottingham was measured on each of 1594 days. A histogram of these values is shown in Figure 4.2.

Note that the histogram does not suggest a normal, exponential or uniform probability model. However, the central limit theorem applies for large samples whatever the distribution of the population from which the data are drawn. Since the sample size is large in this case ($n = 1594$), the central limit theorem may be used.

The mean concentration is 15.171, with sample standard deviation 8.2773. Owing to random fluctuations, it is unlikely that the population mean is exactly 15.171, the value of the sample mean. A measure of the likely discrepancy between the sample mean and the population mean is provided by the standard error of the mean $\sigma/\sqrt{n}$.

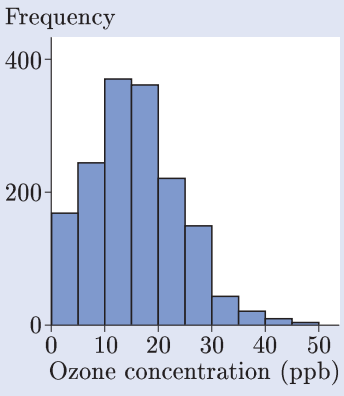

Figure 4.2 A histogram of daily average ozone concentrations in Nottingham

The estimated standard error of the mean is obtained by substituting the sample standard deviation s for σ in the formula for the standard error:

$$\frac{s}{\sqrt{n}} \simeq \frac{8.2773}{\sqrt{1594}} \simeq 0.2073.$$

Since this is quite small compared to the range of the data, this suggests that the sample mean gives a good estimate of the population mean.

Note that, in reporting results, it is good practice to round the numbers appropriately. The ozone concentrations were reported as whole numbers. So reporting the mean ozone concentration as 15.171 probably conveys a spurious impression of accuracy. On the other hand, rounding too much loses information. There are no hard and fast rules about how much rounding should be used. In this case, a reasonable compromise is to keep one decimal place, and report the mean as 15.2. However, note that full accuracy should be kept in intermediate calculations in order to avoid introducing rounding error. ♦

Given a sample of data, one way to represent the uncertainty in an estimate of a mean μ is to obtain a *confidence interval* for μ. A confidence interval for μ is a range of plausible values for μ, which stretches from some value μ^- below the estimate $\widehat{\mu}$, to some value μ^+ above $\widehat{\mu}$. But how wide should the interval be?

The estimated standard error of the mean provides an indication of the uncertainty with which the population mean μ is estimated. So a reasonable approach is to make the width of the interval proportional to the estimated standard error: the larger the standard error is, the greater is the uncertainty surrounding the estimate $\widehat{\mu}$, and the wider is the confidence interval for μ. The constant of proportionality reflects a quantity called the *confidence level*, which is expressed as a percentage: the higher the confidence level, the wider the interval. Thus, for example, a 99% confidence interval — that is, one with confidence level 99% — will be wider than a 90% confidence interval. By the central limit theorem, for large samples, the sampling distribution of the mean is approximately normal, so approximate confidence intervals can be obtained by using quantiles of the standard normal distribution for the constant of proportionality. Quantiles of the standard normal distribution are traditionally denoted z. The calculation of approximate confidence intervals for the population mean, based on large samples, is summarized in the following box.

Large-sample confidence intervals for the population mean

Given a sufficiently large sample (of size n) from a population with mean μ, an **approximate** $100(1-\alpha)\%$ **confidence interval** for the mean μ is given by (μ^-, μ^+), where the end-points are calculated from the sample mean $\widehat{\mu}$, the estimated standard error of the mean $s/\sqrt{n}$, and the $(1-\alpha/2)$-quantile of the standard normal distribution, which is denoted z, as follows:

$$\mu^- = \widehat{\mu} - z\frac{s}{\sqrt{n}}, \quad \mu^+ = \widehat{\mu} + z\frac{s}{\sqrt{n}}.$$

The confidence interval (μ^-, μ^+) is also called a ***z*-interval**. The end-points are called **confidence limits**.

Quantiles of the standard normal distribution may be found using a table of quantiles such as the one given in the *Handbook*. Part of that table is reproduced here as Table 4.1.

For example, to find the quantile required for a 95% confidence interval, proceed as follows. For $100(1-\alpha) = 95$, $\alpha = 0.05$, so $1 - \alpha/2 = 1 - 0.05/2 = 0.975$, and hence the 0.975-quantile is required. From Table 4.1, this is 1.960. Thus, for a 95% confidence interval, $z = 1.960$.

Table 4.1 Selected quantiles of the standard normal distribution

α	q_α
0.800	0.8416
0.850	1.036
0.900	1.282
0.950	1.645
0.975	1.960
0.990	2.326
0.995	2.576

Example 4.2 *Ozone levels — a confidence interval*

An approximate 95% confidence interval for the mean daily ozone concentration is calculated using $\widehat{\mu} = 15.171$, $s/\sqrt{n} = 0.2073$, and $z = 1.96$ (the 0.975-quantile of the standard normal distribution). Thus

$$\mu^- = \widehat{\mu} - z\frac{s}{\sqrt{n}} \simeq 15.171 - 1.96 \times 0.2073 \simeq 14.765,$$
$$\mu^+ = \widehat{\mu} + z\frac{s}{\sqrt{n}} \simeq 15.171 + 1.96 \times 0.2073 \simeq 15.577.$$

Note that full numerical accuracy should be retained throughout these calculations: any rounding should be carried out at the end.

Hence an approximate 95% confidence interval for μ, the mean daily ozone concentration, is $(14.8, 15.6)$. ♦

It follows from the central limit theorem that if samples of size n were repeatedly and independently obtained, and the $100(1-\alpha)\%$ confidence interval (μ^-, μ^+) was calculated on each occasion, then the percentage of intervals containing the true value μ would be approximately $100(1-\alpha)\%$, the approximation improving as n increases. This is known as the **repeated experiments** interpretation of confidence intervals.

A limitation of the repeated experiments interpretation is that it does not help in interpreting the confidence interval you have actually calculated, which either contains the true mean, or does not (and you do not know which is the case). In practice, statisticians use the **plausible range** interpretation of confidence intervals: a confidence interval represents a range of values of μ that are plausible at the 95% confidence level, given the observed data. The plausible range interpretation is described in Example 4.3.

Example 4.3 *A plausible range for the mean daily ozone concentration*

In Example 4.1, the sample mean for the daily ozone concentrations was found to be about 15.2 parts per billion, and in Example 4.2, you saw that a 95% confidence interval for the mean daily ozone concentration is $(14.8, 15.6)$. The plausible range interpretation of this confidence interval is as follows.

If the true value of μ were 15.6 or greater, then the probability of observing a sample mean less than or equal to 15.2 would be 0.025 or less. Similarly, if the true value of μ were 14.8 or less, then the probability of observing a sample mean greater than or equal to 15.2 would also be 0.025 or less.

Thus values of μ outside the confidence interval are implausible, since they would require the data configuration to be unlikely. ♦

The population mean is not the only parameter for which z-intervals can be calculated. Whenever the central limit theorem applies, z-intervals can be used, so z-intervals are very versatile. Also, although they are approximate, the accuracy of the approximation improves as the sample size increases.

In general, for a sufficiently large sample, an approximate $100(1-\alpha)\%$ z-interval for a parameter θ is denoted (θ^-, θ^+) and is given by

$$(\theta^-, \theta^+) = \left(\widehat{\theta} - z\widehat{\sigma}, \widehat{\theta} + z\widehat{\sigma}\right),$$

where $\widehat{\theta}$ is the sample estimate of θ, $\widehat{\sigma}$ is the estimated standard error of the estimator $\widehat{\theta}$, and z is the $(1-\alpha/2)$-quantile of the standard normal distribution.

For example, suppose that it is required to calculate a confidence interval for a proportion p, from a sample of size n. In this case, the parameter is p, its estimate is the sample proportion $\widehat{p}$, and the standard error of $\widehat{p}$ may be estimated by

$$\widehat{\sigma} = \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}.$$

So an approximate $100(1-\alpha)\%$ confidence interval for p is given by the z-interval

$$(p^-, p^+) = \left(\widehat{p} - z\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}},\ \widehat{p} + z\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}\right),$$

You do not need to remember this formula.

where z is the $(1-\alpha/2)$-quantile of the standard normal distribution.

Activity 4.2 A confidence interval for the proportion of females among asthma cases

In Example 3.6, the gender distribution of the 1762 asthma cases admitted to hospital was discussed. Of the 1761 cases for whom gender was recorded, 907 were female. Thus the sample proportion of females is

$$\widehat{p} = \frac{907}{1761} \simeq 0.5150.$$

This is an estimate of p, the underlying proportion of females among persons admitted for asthma.

(a) Calculate an approximate 95% confidence interval for p, and summarize your results.

(b) From population statistics, it is known that the proportion of females among residents in Nottingham was 0.4976 during the period when the data were collected. Using the plausible range interpretation, comment on whether females are more or less likely than males to be admitted to hospital for asthma in Nottingham.

The calculation of z-intervals is summarized in the following box.

Large-sample confidence intervals

Given a sufficiently large sample (of size n), let $\widehat{\theta}$ be the sample estimate of the population parameter θ. Then an approximate $100(1-\alpha)\%$ **confidence interval** or ***z*-interval** for θ, which is denoted (θ^-, θ^+), is given by

$$(\theta^-, \theta^+) = \left(\widehat{\theta} - z\widehat{\sigma}, \widehat{\theta} + z\widehat{\sigma}\right),$$

where $\widehat{\sigma}$ is the estimated standard error of the estimator $\widehat{\theta}$, and z is the $(1-\alpha/2)$-quantile of the standard normal distribution.

If θ is the population mean μ, then $\widehat{\theta}$ is the sample mean and $\widehat{\sigma} = s/\sqrt{n}$, where s is the sample standard deviation.

If θ is a binomial proportion p, then $\widehat{\theta}$ is the sample proportion $\widehat{p}$ and $\widehat{\sigma} = \sqrt{\widehat{p}(1-\widehat{p})/n}$.

4.3 Testing hypotheses

A confidence interval quantifies the uncertainty of an estimate due to sampling error. Sometimes, however, scientific questions are formulated as hypotheses about the value or values that a particular parameter may take. There are several related approaches to testing hypotheses. The approach reviewed here is called **significance testing**. You will meet several significance tests in the course. These will be discussed in detail as they are required. In this subsection, data on the lengths of stay for patients admitted to hospital for asthma will be used to illustrate the steps involved in carrying out a significance test.

The purpose of a significance test is to evaluate the strength of the evidence against a **null hypothesis**, denoted H_0. It is sometimes convenient in constructing the test to specify an **alternative hypothesis**, denoted H_1. In M249, the null and alternative hypotheses will be of the form

$$H_0 : \theta = \theta_0, \quad H_1 : \theta \neq \theta_0,$$

where θ_0 denotes some particular value of the parameter θ. Alternative hypotheses of the form $H_1 : \theta \neq \theta_0$ are called **two-sided**, as they allow either $\theta < \theta_0$ or $\theta > \theta_0$. All the significance tests in this course involve two-sided alternative hypotheses. So, on most occasions, the alternative hypothesis will not be stated explicitly, though for clarity it will be stated in this unit.

You may have encountered alternative hypotheses of the form $H_1 : \theta > \theta_0$. These are called one-sided alternative hypotheses.

Example 4.4 Mean duration of hospital stay: the hypotheses

Suppose that an evaluation of the cost of treating asthma in hospital is based on the assumption that the mean length of stay for patients admitted for asthma is two days. Is this assumption valid for patients admitted to the Nottingham hospital of Section 3?

Data on the length of hospital stay in Nottingham were described in Example 3.5 and Activity 3.3. For these data, the average length of stay was 1.885 days, which is less than 2. However, it could be hypothesized that the difference between this sample mean and 2 is due to random variation, and that the underlying mean length of stay is indeed 2.

A significance test is required to test this hypothesis. Let μ denote the mean duration of stay for a patient admitted to hospital for asthma in Nottingham. Then the null and alternative hypotheses are as follows:

$$H_0 : \mu = 2, \quad H_1 : \mu \neq 2. \quad \blacklozenge$$

After setting out the null and alternative hypotheses, the next step is to identify a suitable test statistic, and obtain its **null distribution**, that is, its distribution if the null hypothesis is true.

Example 4.5 Mean duration of hospital stay: the test statistic

An appropriate test statistic for the significance test discussed in Example 4.4 is $\widehat{\mu}$, the mean length of stay. Under the null hypothesis, $\widehat{\mu}$ has mean 2. By the central limit theorem, the approximate distribution of $\widehat{\mu}$ is $N(2, \sigma^2/1762)$, where σ is the population standard deviation of the durations. The sample standard deviation, s, is 1.850. Substituting s for σ gives the estimated standard error:

$$\frac{s}{\sqrt{n}} \simeq \frac{1.850}{\sqrt{1762}} \simeq 0.04407.$$

Thus, if the null hypothesis is true, the distribution of the test statistic is approximately $N(2, 0.04407^2)$.

Note that, in reporting the results, the mean length of stay could be rounded to 1.9 days, since the original data were rounded to the nearest day. Similarly, the standard deviation 1.850 could be rounded to 1.9. However, full accuracy should be retained throughout the calculations. $\blacklozenge$

The observed value of the test statistic is then computed for the sample, and all values at least as extreme as the observed value (in relation to the null hypothesis) are identified. Then the probability of these values under the null hypothesis is calculated. This probability is called the **significance probability**, or ***p* value**, for the test.

Example 4.6 Mean duration of hospital stay: the p value

The value of the mean under the null hypothesis is 2. The observed value of the test statistic is the sample mean, namely 1.885 days. Thus the difference between the observed value and the value under the null hypothesis is 0.115 in magnitude. Since $\mu = 2$ under the null hypothesis, values lying at least 0.115 below or above 2 are at least as extreme as the value observed. These 'at least as extreme' values comprise the two tails of the null distribution: the lower tail is $\widehat{\mu} \leq 1.885$ and the upper tail is $\widehat{\mu} \geq 2.115$. The null distribution, the observed value of the test statistic and the 'at least as extreme' tail regions are shown in Figure 4.3.

The significance probability, or p value, is the probability that the test statistic is at least as extreme as the value observed, if the null hypothesis is true. This is given by

$$p = P(\widehat{\mu} \leq 1.885) + P(\widehat{\mu} \geq 2.115),$$

where $\widehat{\mu} \approx N(2, 0.04407^2)$. This probability, calculated using a computer, is approximately 0.00907. Thus the p value is about 0.009. ♦

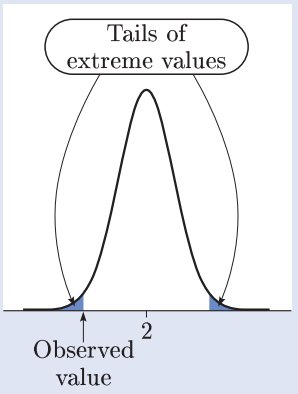

Figure 4.3 The null distribution of the test statistic, showing the observed value and the 'at least as extreme' tails

The final step in a significance test is to interpret the p value. There are no hard and fast rules on how to do this, but Table 4.2 sets out a rough guide, which will be used throughout this course.

Table 4.2 Interpreting p values

Significance probability p	Rough interpretation
$p > 0.10$	little evidence against H_0
$0.10 \geq p > 0.05$	weak evidence against H_0
$0.05 \geq p > 0.01$	moderate evidence against H_0
$p \leq 0.01$	strong evidence against H_0

Example 4.7 Mean duration of hospital stay: conclusion

Following the guide in Table 4.2, our conclusion is as follows.

The p value is 0.009, so $p < 0.01$. So there is strong evidence against the hypothesis that the mean length of stay is two days for patients admitted to the Nottingham hospital for asthma. Since the observed mean duration of hospital stay is less than 2, this suggests that the mean duration is less than two days. ♦

The steps involved in conducting a significance test are set out in the following box.

Significance testing

1 Determine the null hypothesis H_0 and the alternative hypothesis H_1.
2 Choose a suitable test statistic and determine the null distribution of the test statistic.
3 Calculate the observed value of the test statistic and identify the values that are at least as extreme as the observed value in relation to H_0.
4 Calculate the significance probability p.
5 Interpret the significance probability and report the results.

Activity 4.3 Gender and asthma cases

In Activity 4.2 you calculated a confidence interval for the proportion p of hospital admissions for asthma who are female, and compared this to the proportion of females in the population, which is 0.4976.

It is required to test the null hypothesis that the proportion of hospital admissions for asthma who were female is the same as the proportion of females in the population, using the sample of size 1761 patients described in Activity 4.2, 907 of whom were female.

(a) State the null and alternative hypotheses for the test.

(b) An appropriate test statistic is X, the number of females admitted to hospital with asthma, out of 1761. Specify the null distribution of X.

(c) The significance probability for the test is 0.146. Interpret this significance probability. Are females more likely or less likely than males to be admitted to hospital for asthma?

Summary of Section 4

In this section, the hat notation for estimates and estimators has been introduced. The sampling distribution of the mean and the standard error of the mean have been defined, and the role of the central limit theorem in statistical inference has been outlined. Large-sample confidence intervals, or z-intervals, have been described, together with their interpretations in terms of repeated experiments and plausible ranges. Significance testing has been reviewed briefly. The steps involved in carrying out a test have been discussed. These include defining the null and alternative hypotheses, choosing the test statistic, determining its null distribution, and interpreting significance probabilities.

Exercises on Section 4

Exercise 4.1 Mercury contamination in plaice

In Exercise 1.1 you calculated the differences between the mercury contamination levels in plaice in the Irish Sea and the North Sea for seven years between 1984 and 1993. The mean of the seven values (Irish Sea minus North Sea) is 0.053, and the standard deviation is 0.011.

(a) Obtain a 95% z-interval for the difference between the mean mercury contamination level in plaice in the Irish Sea and the North Sea.

(b) Briefly discuss the validity of this confidence interval.

(c) A significance test of the null hypothesis that there is no difference between the mean mercury contamination levels in the two sea areas is to be undertaken using these data. State the null and alternative hypotheses for the test.

(d) The p value for the test is reported as being less than 0.001. Interpret this result. Do the mean contamination levels in plaice differ in the two sea areas?

Exercise 4.2 Asthma in young children

Of the 1762 hospital admissions for asthma described in Example 3.5, 452 were of children aged between 0 and 6 years.

(a) Obtain an approximate 95% confidence interval for p, the underlying proportion of children aged between 0 and 6 years among hospital admissions for asthma.

(b) A significance test is to be conducted of the null hypothesis that a quarter of all hospital admissions for asthma are children aged between 0 and 6 years. State the null and alternative hypotheses for the test.

(c) The significance probability for the test is 0.545. Interpret this result. Is it correct to conclude that children aged between 0 and 6 years account for more than a quarter of all hospital admissions for asthma?

5 Related variables: pollutants and people

In this section, statistical methods for describing the association, if any, between two variables are reviewed. Associations between continuous variables are discussed in Subsection 5.1 using data on air pollutants. Associations between discrete variables are discussed in Subsection 5.2 using data on hospital admissions for asthma.

5.1 Association between two continuous variables

Air quality is evaluated by measuring the concentrations (or levels) of several different pollutants. Data on the levels of particulate matter in the air in central Nottingham on each 1472 days were discussed in Subsection 3.1. In this subsection, data from the same source on the concentrations of five pollutants are discussed: carbon monoxide (CO), nitric oxide (NO), nitrogen dioxide (NO_2), ozone (O_3) and particulate matter (PM_{10}). The PM_{10} level is measured in micrograms per cubic metre ($\mu g\,m^{-3}$); the CO concentration is measured in parts per million (ppm); the others are measured in parts per billion (ppb).

See Example 3.1.

Data such as these, involving several variables, are called multivariate. In Book 3 *Multivariate analysis*, you will learn special techniques for analysing such data.

Example 5.1 Ozone and nitrogen dioxide levels

A scatterplot of the ozone (chemical formula O_3) and nitrogen dioxide (NO_2) concentrations in central Nottingham is shown in Figure 5.1.

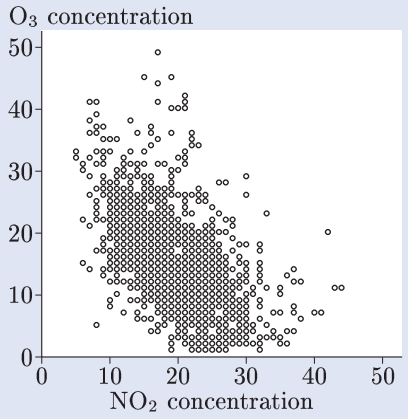

Figure 5.1 A scatterplot of O_3 and NO_2 concentrations

Each point in Figure 5.1 represents a pair of daily averages for the same day, measured in parts per billion (ppb). The pattern of the points on the scatterplot slopes downwards from left to right: ozone concentrations tend to be high when the nitrogen dioxide concentration is low, and low when the nitrogen dioxide concentration is high. ♦

In Example 5.1, knowing the value of one of the concentrations tells you something about the value of the other: the two variables are said to be **related** or **associated**. Note that the words 'related' and 'associated' do not imply that one variable directly or indirectly influences the other. They just mean that the two variables tend to vary together in some systematic way.

When two variables are related, it is often of interest to describe the way in which they are related. In general, two variables are said to be **positively related**, or **positively associated**, if one tends to be high when the other is high, and low when the other is low. When this is the case, the pattern of points in a scatterplot slopes upwards from left to right. Similarly, two variables are said to be **negatively related**, or **negatively associated**, if one tends to be high when the other is low, and low when the other is high. When this is the case, the pattern of points in a scatterplot slopes downwards from left to right. So, for example, the two variables in the scatterplot in Figure 5.1, ozone concentration and nitrogen dioxide concentration, are negatively related.

When the points on a scatterplot appear to be distributed randomly on either side of a straight line, the variables are said to be **linearly related**. From Figure 5.1, it looks as though the ozone and nitrogen dioxide concentrations might be linearly related.

Example 5.2 Ozone and nitric oxide concentrations

A scatterplot of ozone (O_3) and nitric oxide (NO) concentrations in central Nottingham is shown in Figure 5.2.

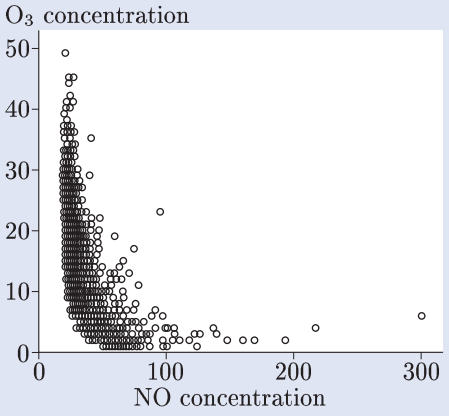

Figure 5.2 A scatterplot of O_3 and NO concentrations

The scatterplot suggests that the two variables may be negatively related: high O_3 concentrations tend to correspond to low NO concentrations, and low O_3 concentrations to high NO concentrations. However, the scatterplot has a curved shape, like the letter J written backwards. In this case the two variables are related, but the relationship is not linear. ♦

Activity 5.1 Describing relationships between continuous variables

Three scatterplots of concentrations of various pollutants in central Nottingham are shown in Figure 5.3.

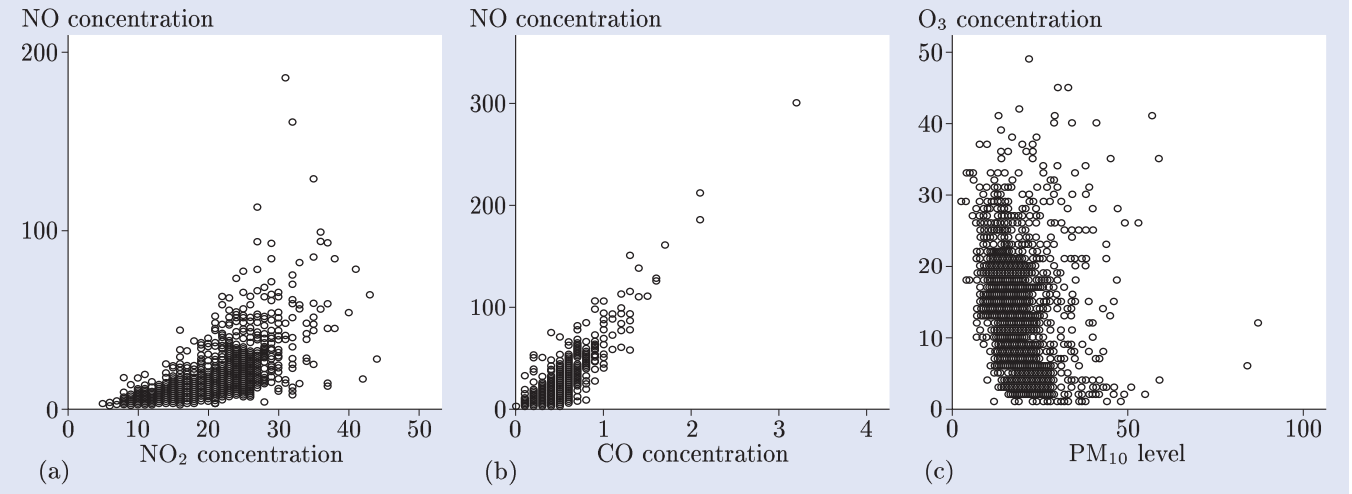

Figure 5.3 Three scatterplots of pollutant concentrations

For each of the scatterplots, describe the relationship between the variables. In particular, say whether the variables are positively or negatively associated. If the variables are related, say whether or not the relationship is linear.

When two continuous variables are linearly related, the points on a scatterplot lie on either side of a straight line. However, the amount of scatter around the line may vary. This is illustrated in Figure 5.4 for data on two pairs of random variables: X and Y, and X and Z.

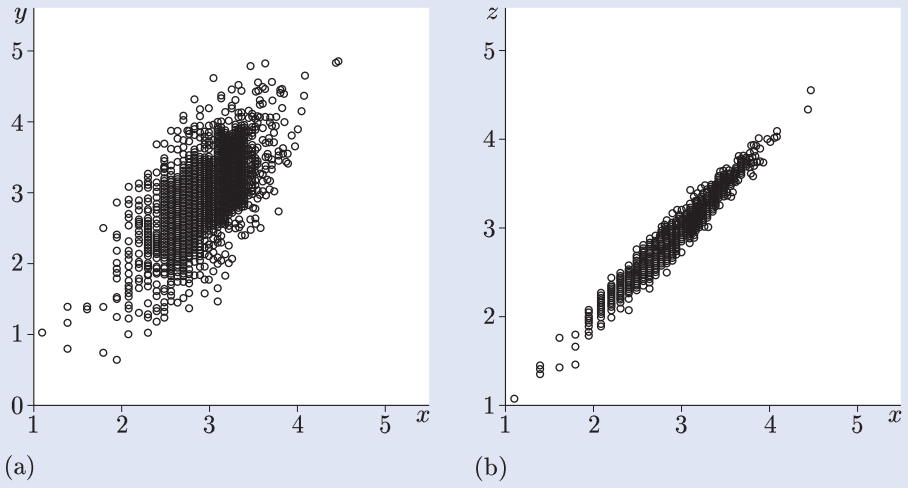

Figure 5.4 Two scatterplots of linearly related variables: (a) X and Y (b) X and Z

The amount of scatter in Figure 5.4(a) is greater than that in Figure 5.4(b). The association between X and Z, shown in Figure 5.4(b), is said to be *stronger* than that between X and Y, shown in Figure 5.4(a).

A measure of the strength of a linear association is provided by the **Pearson correlation coefficient**, which is often simply called the **correlation**. This measure is based upon a statistic called the covariance, which describes how two variables vary together (or 'co-vary').

For observations $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ on two random variables X and Y, for which the sample means of the x-values and the y-values are $\overline{x}$ and $\overline{y}$ respectively, the **sample covariance** is defined by

$$\mathrm{Cov}(x, y) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}). \tag{5.1}$$

Consider the expression $(x_i - \overline{x})$ in (5.1). It is positive for values of x_i above the mean $\overline{x}$, and negative for values below the mean. Similarly $(y_i - \overline{y})$ is positive for values of y_i above $\overline{y}$, and negative for values below $\overline{y}$.

Now suppose that the random variables X and Y are positively related. Then they will tend to take relatively large values at the same time, and relatively low values at the same time. Thus for any i, both x_i and y_i are likely to be above their respective means, or both below their means. If both are above their means, then the two terms in brackets in (5.1) will be positive for that value of i, and their product will be positive, so that this data point will contribute a positive value to the sum in (5.1). If both x_i and y_i are below their means, the two terms will be negative, so again their product will be positive, and the data point will contribute a positive value to the sum. Only when one of the values is below its mean while the other is above, will the contribution to the sum be negative. Since X and Y are positively related, x_i and y_i will tend to be either high (above their means) at the same time, or low (below their means) at the same time. Thus most terms in (5.1) will be positive, and so the covariance will be positive. This is illustrated in Figure 5.5(a).

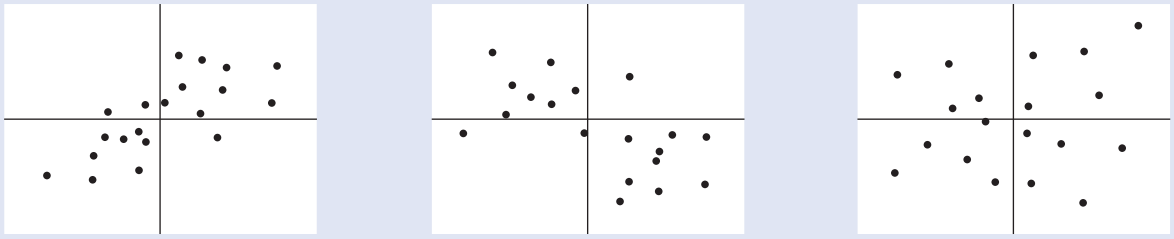

Figure 5.5 Related variables

Similarly, if X and Y are negatively related, then negative terms will dominate in the sum in (5.1) and the covariance will be negative (see Figure 5.5(b)). Finally, if X and Y are not related, a positive value of $(x_i - \overline{x})$ will sometimes be paired with a positive value of $(y_i - \overline{y})$, and sometimes with a negative value. The result is that about half of the terms in (5.1) will be positive and about half will be negative, thus producing a covariance close to zero (see Figure 5.5(c)).

However, the covariance depends on the scale on which X and Y are measured. A measure of association that does not depend on scale, is obtained if the covariance is divided by the sample standard deviations of X and Y, denoted s_x and s_y respectively. So the following expression can be used as a measure of association:

$$r = \frac{\mathrm{Cov}(x, y)}{s_x s_y}.$$

This is the Pearson correlation coefficient. It can be shown, though it will not be done here, that its value always lies between -1 and 1. Values close to $+1$ arise when the points in the scatterplot lie close to a straight line with positive slope. Values close to -1 arise when the points in the scatterplot lie close to a straight line with negative slope.

The correlation coefficient for the variables in Figure 5.4(a) is 0.642, while $r = 0.970$ for the variables in Figure 5.4(b).

Both correlations are positive, indicating positive relationships. The correlation between X and Z is greater than that between X and Y. Thus the association (or correlation) is stronger between X and Z than it is between X and Y.

Activity 5.2 Correlation between ozone and nitrogen dioxide concentrations

One of the following values is the Pearson correlation coefficient for the data shown in Figure 5.1.

-2.055 -0.558 0.607 -0.981 -0.048

Say which value this is, and explain why have you ruled out each of the other values.

5.2 Association between two discrete variables

In Section 3 and Subsection 4.3, data on the daily number of admissions to a Nottingham hospital for asthma, and on their gender and length of stay in hospital, were discussed. In this subsection, data on age and length of stay in hospital will be used to introduce some ideas concerning association between discrete random variables.

See Examples 3.2, 3.5, 3.6 and Examples 4.4 to 4.7.

Example 5.3 Age and length of hospital stay of asthma cases

When two discrete variables X and Y can take only a few distinct values, their joint distribution in a sample of data can be conveniently represented in a table of frequencies, such as the one in Table 5.1.

Table 5.1 Age and length of stay of persons admitted to hospital for asthma

	Age group (years)			
Length of stay	0–19	20–59	60+	Total
Short (0–6 days)	790	698	221	1709
Long (> 6 days)	9	14	29	52
Total	799	712	250	1761

This table is called a **contingency table**. Table 5.1 is a 2×3 table, because there are two row categories and three column categories. (The row and column totals do not count as categories.)

In Table 5.1, the age of 1761 patients admitted to hospital for asthma has been categorized in three groups: children and teenagers (aged 0–19 years), young and middle-aged adults (aged 20–59) and older adults (60+). The length of stay in hospital has been categorized in two groups: short (0 to 6 days), and long (more than 6 days). ♦

Example 5.4 Estimating probabilities from a contingency table

Let the variable X denote the age of a randomly chosen person admitted to hospital for asthma, and let Y denote the length of their stay in hospital. If you knew the probability distribution of X, you could find, for instance, the probability p_1 that a randomly chosen person admitted to hospital for asthma is aged between 20 and 59 years. The true distribution of X is unknown, but the data in Table 5.1 can be used to obtain an estimate $\widehat{p}_1$ of this probability. The total number of persons aged between 20 and 59 years among the 1761 admitted for asthma is 712, so

$$\widehat{p}_1 = \frac{712}{1761} \simeq 0.404.$$

Similarly, suppose that an estimate is required of the probability p_2 that a randomly chosen person admitted to hospital for asthma is aged between 20 and 59 years *and* remains in hospital for more than 6 days. From Table 5.1, out of 1761 persons admitted for asthma, the total number of persons who are aged between 20 and 59 years and who remain in hospital for more than 6 days is 14. Hence an estimate of p_2 is given by

$$\widehat{p}_2 = \frac{14}{1761} \simeq 0.008. \quad \blacklozenge$$

Conditional probabilities are probabilities of events of the form '$Y = y$, given $X = x$'. In mathematical notation, the 'given' in this sentence is denoted by a vertical bar. Thus

$$P(Y = y|X = x)$$

is notation for 'the probability that $Y = y$, given that $X = x$', or 'the probability that $Y = y$, conditional on $X = x$'. Conditional probabilities can also be estimated from a contingency table.

Example 5.5 Estimating a conditional probability

Suppose that it is required to estimate p_3, the conditional probability that a randomly chosen person admitted to hospital for asthma remains there for more than 6 days, given that the person is aged between 20 and 59 years. From Table 5.1, the total number of admissions in persons aged between 20 and 59 years is 712. Of these, 14 stayed in hospital for more than 6 days. Hence an estimate of p_3 is

$$\widehat{p}_3 = \frac{14}{712} \simeq 0.020.$$

Note that the probability p_3 can equivalently be described as the probability that a randomly chosen person of age between 20 and 59 years who is admitted to hospital for asthma remains there for more than 6 days. In this description, the words 'given that' do not feature. Nevertheless, the description refers to a conditional probability. $\blacklozenge$

p_3 may be written as $P(Y = \text{'long'} \mid X = \text{'20–59'})$. Strictly speaking, random variables must take numerical values, and not category labels such as '20–59 years'. This is not a problem as we can represent the categories with numbers, for example, 1 for the 0–19 age group, 2 for 20–59, 3 for 60+.

Activity 5.3 Estimating probabilities from a contingency table

State whether or not each of the following statements describes a conditional probability. Use the data in Table 5.1 to estimate each of the probabilities.

(a) The probability that a randomly chosen person admitted to hospital for asthma is aged between 0 and 19 years.

(b) The probability that a randomly chosen person admitted to hospital for asthma is aged between 0 and 19 years and remains in hospital for more than 6 days.

(c) The probability that a randomly chosen person admitted to hospital for asthma, who is aged between 0 and 19 years, remains in hospital for more than 6 days.

Example 5.6 Dependence between age and hospital stay

Using Table 5.1, an estimate of the probability p_4 that a randomly chosen person admitted to hospital for asthma stays there for more than 6 days is given by

$$\widehat{p}_4 = \frac{52}{1761} \simeq 0.030.$$

Now consider p_5, the conditional probability that a randomly chosen person admitted to hospital for asthma, who is aged 60+ years, stays there for more than 6 days. An estimate of p_5 is given by

p_5 may be written as $P(Y = \text{'long'} \mid X = \text{'60+'})$.

$$\widehat{p}_5 = \frac{29}{250} \simeq 0.116.$$

Thus the estimated conditional probability of a long hospital stay, given that the person admitted is aged 60+ years, is larger than $\widehat{p}_4$. This suggests that longer hospital stays are more likely for older patients than among patients as a whole: in other words, length of hospital stay depends on age.

The dependence of length of stay on age may also be seen from the data for the younger age groups, for which the proportions of patients with hospital stays over 6 days are lower than among all patients: 0.020 for the 20–59 year olds (as shown in Example 5.5) and 0.011 for 0–19 year olds (as you found in Activity 5.3). ♦

In Example 5.6, the dependence between two discrete random variables was investigated by examining the effect on the distribution of one variable (length of hospital stay) of conditioning on the value of the other variable (age). This idea can be used to define independence for two discrete random variables.

Independent discrete random variables

Two discrete random variables X and Y are **independent** if, for all values of x and y,

$$P(Y = y|X = x) = P(Y = y).$$

If X and Y are not independent, they are said to be **dependent**, or **related**, or **associated**.

In Example 5.6, estimates were used to compare conditional and unconditional probabilities. However, the differences between these estimated probabilities could be due to sampling variation. A formal significance test is required to evaluate the evidence against the null hypothesis of no association. The statistical analysis of associations between discrete random variables, including tests for no association, is discussed in Book 1 *Medical statistics*.

Activity 5.4 *Age and gender of hospital admissions for asthma*

Data on both age and gender were available for 1760 persons admitted to hospital for asthma. Table 5.2 gives the distribution of these persons by age group and gender.

Table 5.2 Age and gender of persons admitted to hospital for asthma

	Age group (years)			
Gender	0–19	20–59	60+	Total
Male	497	276	81	854
Female	302	435	169	906
Total	799	711	250	1760

(a) Estimate the probability that a person admitted to hospital for asthma is male.

(b) Estimate the probability that a person admitted to hospital for asthma, who is aged between 0 and 19 years, is male.

(c) What do the probabilities you estimated in parts (a) and (b) suggest about the relationship, if any, between age and gender in persons admitted to hospital for asthma?

Summary of Section 5

In this section, relationships between two continuous variables have been discussed, and the sample covariance and the Pearson correlation coefficient have been defined. The estimation of probabilities, including conditional probabilities, from contingency tables has been described. These estimates have been used to investigate dependence between discrete random variables.

Exercises on Section 5

Exercise 5.1 *Nitrogen dioxide and particulate matter*

A scatterplot of nitrogen dioxide (NO_2) concentrations and particulate matter (PM_{10}) levels in the air in central Nottingham is shown in Figure 5.6.

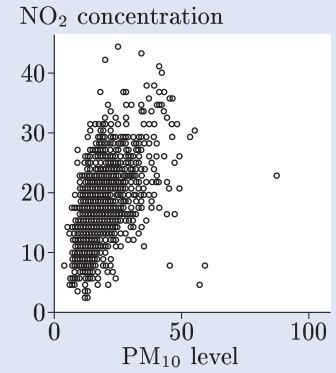

Figure 5.6 A scatterplot of NO_2 concentration and PM_{10} levels in central Nottingham

(a) Briefly describe the relationship between nitrogen dioxide concentrations and particulate matter levels.

(b) One of the following values is the correlation between the two variables.

$$-0.178 \quad 1.328 \quad 0.986 \quad 0.516 \quad 0.002$$

State which value is the correlation, and explain why you have not chosen each of the other values.

Exercise 5.2 *Hospital admissions for asthma in older people*

(a) For each of the probabilities described below, state whether or not it is a conditional probability, and use the data in Table 5.2 to estimate the probability.

(i) The probability that an individual admitted to hospital for asthma is aged 60 years or over.

(ii) The probability that an individual admitted to hospital for asthma is male and aged 60 years or over.

(iii) The probability that a male admitted to hospital for asthma is aged 60 years or over.

(b) Describe how you would use the probabilities defined in part (a) to investigate whether age and gender are associated in persons admitted to hospital for asthma.

6 Statistical modelling in SPSS: the air we breathe

Throughout this section, you will use the data on air quality and asthma that have been discussed in Sections 3, 4 and 5. The data are in the SPSS data file **airquality.sav**. This data file contains information on air quality and asthma admissions in Nottingham, obtained for 1643 successive days between 1 January 2000 and 30 June 2004. There are eight variables. The first variable, `day`, represents the day number: day 1 is 1 January 2000, day 2 is 2 January 2000, and so on. The next six variables give the daily average concentrations of six pollutants: carbon monoxide (`CO`), nitric oxide (`NO`), nitrogen dioxide (`NO2`), ozone (`O3`), particulate matter (`PM10`), and sulphur dioxide (`SO2`). `CO` is measured in parts per million (ppm), `PM10` is measured in micrograms per cubic metre ($\mu g\,m^{-3}$), and the others are measured in parts per billion (ppb). The eighth variable, `asthma`, is the number of hospital admissions for asthma on each day.

Transforming data in SPSS is described in Subsection 6.1. In Subsection 6.2, you will learn how to use SPSS to obtain confidence intervals and correlation coefficients.

6.1 Transforming variables

Transforming one variable into another is a commonly used statistical procedure. In Subsection 3.1, the PM_{10} levels were transformed by taking logarithms, and the $\log PM_{10}$ levels were modelled using a normal distribution. This transformation was done so that a standard distribution could be used in the modelling process. In this subsection, you will learn how to transform variables using SPSS.

Activity 6.1 A model for nitric oxide concentrations

Open the data file **airquality.sav**.

(a) Obtain a histogram of nitric oxide concentrations (`NO`), with the first interval starting at 0 and with interval width 5. Also obtain the mean and standard deviation of the concentrations.

Use **Graphs** > **Histogram...**. Instructions for altering the interval widths using **Chart Editor** are given in Activity 2.11.

(b) Suggest a possible model for the nitric oxide concentrations.

(c) Do you think your suggested model is appropriate for these data? Give one reason in favour of your model and one reason against it.

Activity 6.2 The log transformation

Rather than attempt to model the nitric oxide concentrations directly, an alternative approach is to transform them. In SPSS one variable can be transformed into another using **Compute...** from the **Transform** menu. Transform `NO` by taking natural logarithms, as follows.

In M249, natural logarithms are used, rather than logarithms to any other base.

◇ Choose **Compute...** from the **Transform** menu (by clicking on it). The **Compute Variable** dialogue box will open.

◇ Type the name of the variable in which the transformed data are to be stored, `logNO` say, in the **Target Variable** field.

◇ The natural log function in SPSS is called `LN`. Type `LN(NO)` in the **Numeric Expression** field.

◇ Click on **OK**.

A new variable `logNO` will be created containing the logarithms of the nitric oxide concentrations. If you wish, you can view `logNO` in the **Data Editor**.

Obtain the default histogram of `logNO`, and suggest a possible model for `logNO`.

Activity 6.3 *Finding a function*

In Activity 6.2, you were told the SPSS name for the natural logarithm function. In this activity, instructions are given for using a function when you do not know its SPSS name. The method is illustrated for the natural logarithm function using the nitric oxide concentrations. (You might like to try transforming one of the variables using a different function of your choice.)

Use **Transform** > **Compute...**.

Obtain the **Compute Variable** dialogue box and click on **Reset**. (This removes any previous entries and resets the SPSS defaults.) Type the name of the variable in which the transformed data are to be stored in the **Target Variable** field. Then select the natural logarithm function from the list of functions available in SPSS, as follows.

- ◇ In the **Function group** list, select `Arithmetic` (by clicking on it), and a list of the arithmetic functions available in SPSS will be displayed in the **Functions and Special Variables** area.
- ◇ Select `Ln` from the list of functions (by clicking on it), and a description of the function will appear in the area to the left of the list.
- ◇ Click on the vertical arrow just above this description, and `LN(?)` will be entered in the **Numeric Expression** field.
- ◇ Replace the question mark with the variable name, `NO` say, by typing `NO` (or by entering `NO` from the list of variables on the left-hand side of the dialogue box).

The dialogue box, with the relevant items highlighted, is shown in Figure 6.1.

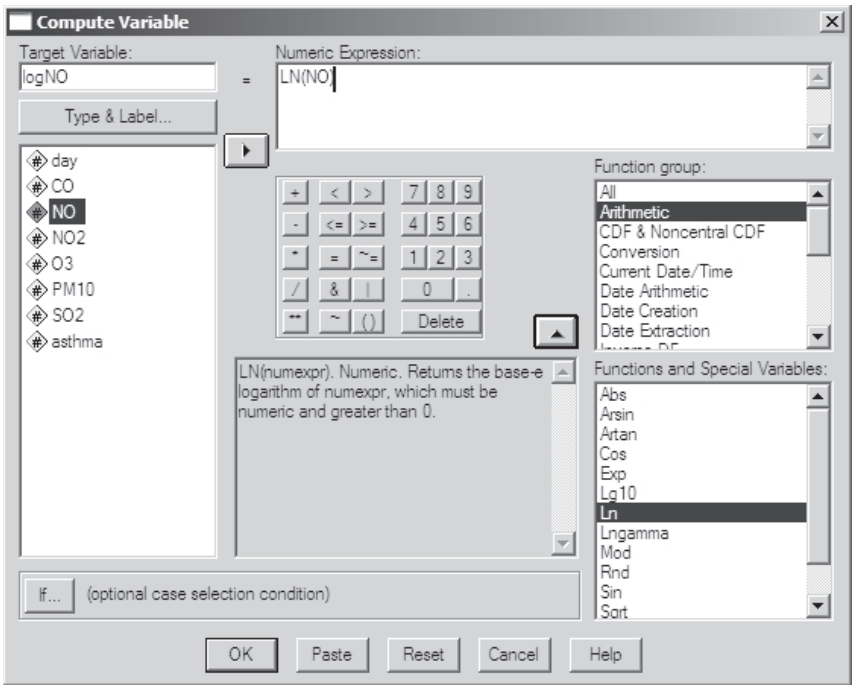

Figure 6.1 Calculating the logarithm of `NO`

When you click on **OK**, the new variable will be created.

In Section 3, the data on particulate matter levels and on daily hospital admissions for asthma were discussed separately. Is the concentration of particulate matter associated with asthma attacks? One way to investigate this is by comparing the numbers of admissions to hospital on days with high PM_{10} levels with the numbers on days with low PM_{10} levels. A reasonable approach is to classify as 'low' those PM_{10} levels that are less than or equal to the median, and to classify as 'high' those PM_{10} levels that are greater than the median. Combining values of a variable into categories in this way to form a categorical variable is called **recoding** in SPSS. You will learn how to do this in Activity 6.4.

Later, in Activity 6.6, you will compare the numbers of admissions on days with low PM_{10} levels and days with high PM_{10} levels.

Activity 6.4 Recoding the PM_{10} levels

In SPSS, recoding a variable is done using **Recode** from the **Transform** menu. In this activity you will use **Recode** to create a categorical variable named `PM10group` from the variable `PM10`. The median of `PM10` is 17. Values of `PM10` that are less than or equal to the median ('low') will form category 1, and values that are greater than the median ('high') will form category 2.

You can check the value of the median using **Analyze** > **Descriptive Statistics** > **Frequencies...**, as described in Activity 2.12.

Use **Recode** to create the variable `PM10group`, as follows.

- ◇ Choose **Into Different Variables...** from the **Recode** submenu of **Transform**. The **Recode into Different Variables** dialogue box will open.
- ◇ Enter the variable `PM10` in the **Input Variable -> Output Variable** field. This field will display `PM10 --> ?`.
- ◇ In the **Output Variable** area, type `PM10group`, the name of the new variable, in the **Name** field.
- ◇ Click on **Change** in the **Output Variable** area. The display in the **Input Variable -> Output Variable** field will change to `PM10 --> PM10group`.
- ◇ Click on the **Old and New Values...** button.

The **Recode into Different Variables: Old and New Values** dialogue box will open. This offers many recoding options. Begin by recoding values less than or equal to 17 (the median of `PM10`) as 1, as follows.

- ◇ In the **Old Value** area of the dialogue box, click on the fifth radio button down and type `17` in the **Range: Lowest through** field.
- ◇ In the **New Value** area (on the right), type `1` in the **Value** field.
- ◇ Click on **Add**.

The recoding you have defined will appear in the **Old --> New** area as `Lowest thru 17 --> 1`.

Now recode the values greater than 17 as 2, as follows.

- ◇ In the **Old Value** area, click on the sixth radio button down, and type `17.01` in the **Range: ... through highest** field.
- ◇ In the **New Value** area, type `2` in the **Value** field.
- ◇ Click on **Add**, and the text `17.01 thru Highest --> 2` will appear in the **Old --> New** area.
- ◇ Click on **Continue**, then on **OK**.

The new variable `PM10group` will appear in the **Data Editor**. Check that the new variable is correct, as follows.

Use **Analyze** > **Descriptive Statistics** > **Frequencies...**.

- ◇ Obtain the **Frequencies** dialogue box.
- ◇ Enter `PM10group` in the **Variable(s)** field. If **Display frequency tables** is not selected — that is, if there is no tick in its check box — then select it (by clicking on it or on its check box).
- ◇ Click on **OK**.

The following frequency table will appear in the **SPSS Viewer**.

PM10group

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1.00	802	48.8	52.0	52.0
	2.00	740	45.0	48.0	100.0
	Total	1542	93.9	100.0	
Missing	System	101	6.1		
Total		1643	100.0		

This table shows that PM_{10} levels were measured on 1542 days. On 802 days, the PM_{10} level was lower than or equal to 17, and on 740 days it was greater than 17.

The recoded variable `PM10group` will be required in Activity 6.6, so save the data file, which now includes `PM10group`, in a data file named **airquality2.sav**.

6.2 Confidence intervals and correlations

In this subsection, you will learn how to use SPSS to obtain confidence intervals for the mean, and to calculate correlation coefficients. You will need the air quality data again for the activities. In particular, in Activity 6.6 you will need the variable `PM10group` that you created and saved in the data file **airquality2.sav** in Activity 6.4. So if that data file is not still open, then open it now.

Activity 6.5 Mean daily number of asthma admissions

In this activity, you will obtain a confidence interval for the mean daily number of hospital admissions for asthma. A confidence interval for a population mean may be found using **Explore...** from the **Descriptive Statistics** submenu of **Analyze**. Summary statistics can also be found using **Explore...**.

There are several ways of obtaining summary statistics in SPSS: in Activity 2.12, you used **Frequencies**.

Obtain the sample mean daily number of hospital admissions for asthma, together with a 95% confidence interval for the mean, as follows.

- ◇ Choose **Explore...** from the **Descriptive Statistics** submenu of **Analyze**. The **Explore** dialogue box will open.
- ◇ Enter `asthma` in the **Dependent List** field. Leave the **Factor List** field and the **Label Cases by** field empty.
- ◇ In the **Display** area, select **Statistics**. This will limit the output to numerical summaries.
- ◇ Click on the **Statistics...** button to open the **Explore: Statistics** dialogue box. Check that **Descriptives** is checked, and that `95` appears in the **Confidence Interval for the Mean** field. (Leave the other boxes unchecked.)
- ◇ Click on **Continue**, and then on **OK**.

Two tables will appear in the **SPSS Viewer**: the `Case Processing Summary`, which gives the number of cases processed, and the following table.

Descriptives

			Statistic	Std. Error
asthma	Mean		1.07	.028
	95% Confidence Interval for Mean	Lower Bound	1.02	
		Upper Bound	1.13	
	5% Trimmed Mean		.97	
	Median		1.00	
	Variance		1.285	
	Std. Deviation		1.134	
	Minimum		0	
	Maximum		*	
	Range		14	
	Interquartile Range		2	
	Skewness		1.881	.060
	Kurtosis		10.966	.121

The sample mean and the 95% confidence interval are given at the top of this table: the mean daily number of asthma admissions is 1.07, with 95% confidence interval $(1.02, 1.13)$.

This confidence interval is a t-interval. In large samples, z-intervals and t-intervals are nearly the same. The calculation of t-intervals is not described in M249.

Activity 6.6 Particulate matter and asthma

In this activity you will obtain the sample mean number of asthma cases admitted on days with low average PM_{10} levels, and on days with high average PM_{10} levels, with 95% confidence intervals. Obtain the sample means and confidence intervals, as follows.

◇ Obtain the **Explore** dialogue box.

Use **Analyze** > **Descriptive Statistics** > **Explore...**.

◇ Enter `asthma` in the **Dependent List** field and `PM10group` in the **Factor List** field. (Leave the **Label Cases by** field empty.)

◇ In the **Display** area, select **Statistics**.

◇ Click on the **Statistics...** button to open the **Explore: Statistics** dialogue box. Check that **Descriptives** is checked, and that `95` appears in the **Confidence Interval for the Mean** field. (Leave the other boxes unchecked.)

◇ Click on **Continue**, and then on **OK**.

In the **SPSS Viewer**, locate the `Descriptives` table. Notice that there are two sub-tables, corresponding to the categories `PM10group` = 1 and `PM10group` = 2. Use these sub-tables to write down the mean number of hospital admissions for asthma on days with low average PM_{10} levels, and the mean number on days with high average PM_{10} levels. Also write down the 95% confidence intervals for the underlying means.

In your view, do these results lend support to the hypothesis that the higher the average daily PM_{10} level, the greater the number of persons admitted to hospital for asthma?

Activity 6.7 Carbon monoxide and nitric oxide

Air quality can be measured in many different ways. In the Nottingham air quality data, concentrations of several pollutants are given. Relationships between pollutants can be investigated by producing scatterplots and calculating correlation coefficients. You learned how to obtain a scatterplot in Activity 2.10. Correlation coefficients are calculated using **Bivariate...** from the **Correlate** submenu of **Analyze**. In this activity you will calculate the correlation between the carbon monoxide and nitric oxide concentrations in the air.

Obtain a scatterplot of `CO` (carbon monoxide concentration) against `NO` (nitric oxide concentration).

Use **Graphs** > **Scatter/Dot...**.

A scatterplot is shown in Figure 6.2.

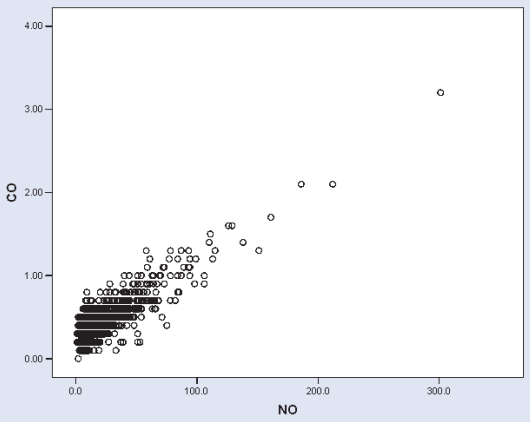

Figure 6.2 A scatterplot of `CO` against `NO`

This scatterplot indicates a rather strong linear relationship between the two variables.

Obtain the correlation between `CO` and `NO`, and carry out a significance test of the null hypothesis of zero correlation, as follows.

- ◇ Choose **Bivariate...** from the **Correlate** submenu of **Analyze**. The **Bivariate Correlations** dialogue box will open.
- ◇ Enter `CO` and `NO` in the **Variables** field.
- ◇ In the **Correlation Coefficients** area, make sure that **Pearson** is checked; and in the **Test of Significance** area, make sure that **Two-tailed** is selected.
- ◇ Deselect **Flag significant correlations** (by clicking on it or on the tick in its check box).
- ◇ Click on **OK**.

The following table will appear in the **SPSS Viewer**.

Correlations

		CO	NO
CO	Pearson Correlation	1	.837
	Sig. (2-tailed)		.000
	N	1568	1562
NO	Pearson Correlation	.837	1
	Sig. (2-tailed)	.000	
	N	1562	1606

Look at the top entries in the `NO` column. The Pearson correlation coefficient between `CO` and `NO` is 0.837; it was calculated from $N = 1562$ pairs of values. (Of course, this is the same as the correlation between `NO` and `CO`.)

SPSS has conducted a significance test of the null hypothesis of zero correlation. The significance probability for the test is given in the row labelled `Sig. (2-tailed)`. This is quoted as .000, which means that $p < 0.0005$.

Thus there is strong evidence against the null hypothesis of zero correlation, and hence strong evidence that carbon monoxide and nitrogen dioxide concentrations in the air are related.

Summary of Section 6

In this section, you have learned how to transform and recode data in SPSS, and how to obtain a confidence interval for the mean. The calculation of correlation coefficients has been described. These methods have been applied to data on air quality and asthma in Nottingham.

7 Modelling exercises

The purpose of these exercises is to give you more practice, if you feel you need it, in using SPSS to apply the methods described in this unit.

Exercise 7.1 Mercury in North Sea and Irish Sea plaice

In Activity 1.5 you compared the levels of mercury contamination in plaice in the North Sea and in the Irish Sea. In this exercise you will investigate the issue further. The data are stored in the SPSS data file **mercury.sav**.

These data are in Table 1.3.

(a) Create a variable `diff` containing the differences between the mercury contamination levels in the Irish Sea and the North Sea (Irish Sea minus North Sea) for years when data are available from both sea areas, as follows.

- ◇ Obtain the **Compute Variable** dialogue box.
- ◇ Type `diff` in the **Target Variable** field.
- ◇ Enter `Irishsea − Northsea` in the **Numeric Expression** field.
- ◇ Click on **OK**.

To enter the minus sign, click on the minus button on the calculator keypad below the **Numeric Expression** field.

How does SPSS cope with the missing values?

(b) Obtain a histogram of the differences, with bin width 0.01.

(c) Estimate the mean difference and obtain a 95% confidence interval for the mean difference in mercury concentrations.

(d) Summarize your conclusions.

Exercise 7.2 Catch and biomass of North Sea cod

It might be expected that the annual catch is positively related to the spawning stock biomass for the simple reason that the more fish there are in the sea, the more will be caught. Use the data file **cod.sav** to investigate this hypothesis as it relates to cod in the North Sea.

This data set is described in Activity 2.8. This exercise uses the variables `biomass` and `catch`.

(a) Obtain a scatterplot of `catch` against `biomass`. What do you conclude about the relationship between catch and biomass?

(b) Obtain the Pearson correlation coefficient between `catch` and `biomass`, and the p value for the significance test of the null hypothesis of zero correlation.

(c) Interpret the p value you obtained in part (b).

Exercise 7.3 Nitrogen oxides in the air

In the solution to Activity 6.2, it was suggested that a normal model may be appropriate for describing the day-to-day variation in the logarithms of average daily NO concentrations. In this exercise you will consider the distribution of nitrogen dioxide (NO_2) concentrations, and the relationship between NO_2 concentrations and nitric oxide (NO) concentrations. The data are in the data file **airquality.sav**.

(a) Create a variable named `sqrtNO2` containing the square root of `NO2`. Obtain histograms of `NO2` and `sqrtNO2`, and calculate the sample skewness of both variables.

Use **Transform** > **Compute...** and the function `Sqrt`.

(b) Briefly describe the results you obtained in part (a). Is a normal model appropriate for `NO2` or for `sqrtNO2`?

(c) Create a variable named `logNO` containing the logarithms of the nitric oxide concentrations. Obtain scatterplots of `NO` against `sqrtNO2`, and `logNO` against `sqrtNO2`.

(d) Calculate the Pearson correlation coefficients between `NO` and `sqrtNO2`, and between `logNO` and `sqrtNO2`.

(e) Which of the correlation coefficients that you calculated in part (d) is larger? Use the scatterplots you obtained in part (c) to explain why it is larger.

Exercise 7.4 *Hospital stays*

In this exercise you will examine the lengths of stay for males and females who are admitted to hospital for asthma, and how they differ. This may be explored using the data in **asthma.sav** using the variables `stay` and `sex`.

(a) Obtain an estimate of the mean length of stay, the sample standard deviation and a 95% confidence interval for the mean stay, for males and females separately.

(b) What do your findings from part (a) suggest?

(c) A significance test of the null hypothesis that the mean length of stay is the same for males and females yields the p value 0.001. What do you conclude?

Summary of Unit

A statistical analysis often begins with exploring the data using various graphical methods. The graphs reviewed in this unit include bar charts, line plots, scatterplots and histograms. Numerical summaries, including measures of location and dispersion, can also help to describe a data set. The measures reviewed include the mean, median, mode, standard deviation and variance. Several commonly used probability models have been described: the normal, exponential and uniform models for continuous variables, and the binomial, Poisson and uniform models for discrete data. A key aspect of statistical modelling is to draw inferences about populations on the basis of samples. Methods for drawing such inferences, including z-intervals and significance tests, have been discussed. The methods described are valid for large samples and depend on the central limit theorem. Techniques for describing and quantifying the association between two variables, including correlation coefficients and conditional probabilities, have been reviewed. The implementation of these statistical methods in SPSS has been described.

Learning outcomes

You have been working to develop the following skills.

- ◇ Explore a data set using appropriate graphs and numerical summaries.
- ◇ Interpret graphs and numerical summaries.
- ◇ Select an appropriate probability model for a continuous random variable or a discrete random variable.
- ◇ Calculate approximate confidence intervals for means and proportions.
- ◇ Describe the rationale underlying significance tests.
- ◇ Interpret significance probabilities.
- ◇ Interpret the Pearson correlation coefficient.
- ◇ Estimate conditional probabilities and explore dependence in contingency tables.

You have also been working to acquire the following skills in using SPSS.

- ◇ Manipulate, transform and recode data.
- ◇ Construct and customize graphs including bar charts, line plots, scatterplots and histograms.
- ◇ Calculate numerical summaries including measures of location and measures of dispersion.
- ◇ Calculate confidence intervals for the mean.
- ◇ Calculate Pearson correlation coefficients.
- ◇ Print output, and export output for use in other documents.

Solutions to Activities

Solution 1.1

(a) There is considerable variation between species, but for most species the catch declined between 1979 and 1999.

(b) For sole, the catch appears to have remained broadly constant. For herring, the catch was very much lower in 1979 than in 1989 and 1999.

Solution 1.2

(a) The biomass declined steeply during the 1970s. The annual catch exceeded the biomass between 1968 and 1977. This could indicate over-fishing, which may have endangered the ability of mature fish to replenish stocks naturally.

(b) The biomass rose again in the 1980s as the herring stocks recovered. The start of this recovery coincides with the restrictions on herring fishing. Note that such a coincidence does not in general prove a causal relation. However, in this case it is well known that over-fishing endangers fish stocks, so a causal explanation is reasonable.

Solution 1.3

(a) The number of new recruits increases with the biomass, at least for lower values of the biomass. The trend is increasing: it may be linear, or it may level out as the biomass increases.

(b) The variability in the number of recruits increases markedly with the biomass, as shown by the funnel shape of the scatterplot.

(c) The point in the bottom right-hand corner of the scatterplot corresponds to the highest biomass but a rather low number of recruits. If the relationship between the two variables were linear, then this point would be an outlier. This point and other points that might be considered outliers have been circled in Figure S.1. (You might have identified other points as possible outliers.)

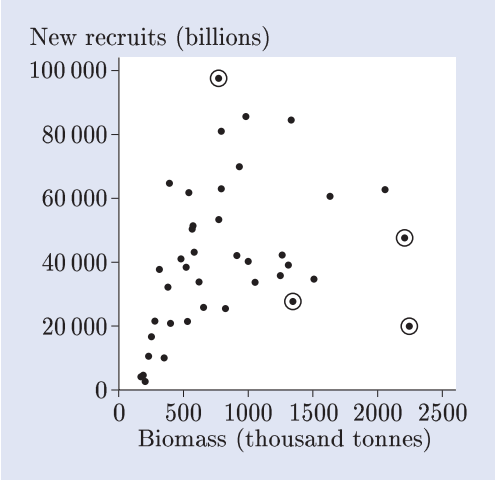

Figure S.1 Possible outliers

Solution 1.4

(a) In each year, the modal fish species is the species with the highest catch. In 1979, it was cod; in 1989 and 1999, it was herring.

(b) As there are 37 values, the median is the nineteenth value. From Figure 1.6(a), there are 6 values in the first interval (0–200), 8 in the second (200–400), and 7 in the third (400–600). So the nineteenth largest value lies in the third interval, and hence the median lies between 400 000 and 600 000 tonnes.

Solution 1.5

(a) There are eight values. The mean is

$$\begin{aligned}\overline{x} &= \tfrac{1}{8}(0.11 + 0.09 + \cdots + 0.09)\\ &= 0.10375 \simeq 0.104.\end{aligned}$$

The eight values, arranged in order of increasing size, are as follows.

0.09 0.09 0.10 0.10 0.11 0.11 0.11 0.12

The median is halfway between the two middle values, which are 0.10 and 0.11. Hence the median m is 0.105.

(b) Using the value 0.10375 for the mean leads to values of 0.0001125 for the variance and $0.0106066\ldots \simeq 0.0106$ for the standard deviation. For simplicity, the calculations are illustrated here using the rounded value of 0.104 for the mean.

The variance is given by

$$\begin{aligned}s^2 &\simeq \tfrac{1}{7}\left((0.09 - 0.104)^2 + \cdots + (0.12 - 0.104)^2\right)\\ &= 0.00011257\ldots\\ &\simeq 0.000113.\end{aligned}$$

So the standard deviation is

$$s = \sqrt{0.00011257\ldots} \simeq 0.0106.$$

Using the rounded value for the mean when calculating the variance has not resulted in a large rounding error in this case. However, in general, using a rounded value for the mean is not recommended as it will often lead to rounding errors being introduced.

Solution 3.1

(a) Since the probabilities sum to 1,

$$\begin{aligned}P(X \geq 5) &= 1 - \big(p(0) + p(1) + p(2) + p(3) + p(4)\big)\\ &= 1 - (0.342 + 0.367 + 0.197 + 0.070 + 0.019)\\ &= 0.005.\end{aligned}$$

Also,

$$P(X \leq 2) = p(0) + p(1) + p(2) = 0.906.$$

Finally,

$$P(X > 2) = 1 - P(X \leq 2) = 1 - 0.906 = 0.094.$$

(b) The probability that there is at least one admission is

$$P(X \geq 1) = 1 - p(0) = 1 - 0.342 = 0.658.$$

Hence there is at least one admission on 65.8% of days.

Solution 3.2

(a) The 0.9-quantile is q_A; q_B is the 0.2-quantile; and q_C is the 0.5-quantile (or median).

(b) The lower quartile should be between q_B and q_C, and the upper quartile between q_C and q_A. The actual values are $q_{0.25} \simeq 1.20$ and $q_{0.75} \simeq 3.37$. The quartiles are shown on Figure S.2.

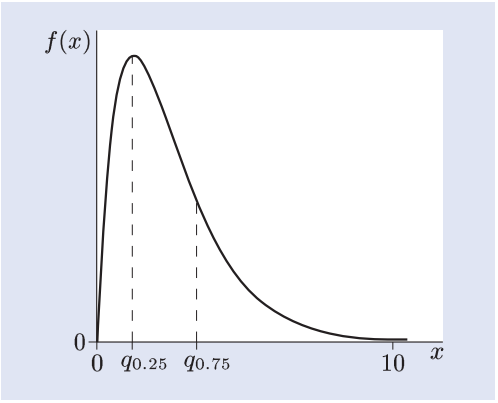

Figure S.2 The quartiles of X

(c) Let a denote the lower boundary of the middle interval, then

$$P(X \leq a) = \tfrac{1}{3} \simeq 0.333.$$

Hence a is the 0.333-quantile of X. Similarly, let b denote the upper boundary of the middle interval, then

$$\tfrac{1}{3} = P(X > b) = 1 - P(X \leq b).$$

Hence

$$P(X \leq b) = 1 - \tfrac{1}{3} \simeq 0.667.$$

So b is the 0.667-quantile of X.

Solution 3.3

For an exponential distribution, the population mean and standard deviation are equal. For this sample of size 1762, the sample mean and standard deviation are roughly equal. This supports (or at least does not rule out) the possibility that an exponential model is appropriate.

If the exponential model is appropriate, very short stays should be the most common. From Figure 3.9, this does not appear to be the case: the mode of the histogram is for stays of between 1 and 2 days.

Solution 3.4

(a) The p.d.f. is shown in Figure S.3.

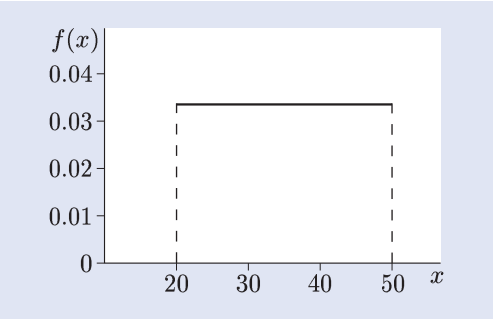

Figure S.3 The p.d.f. of the distribution $U(20, 50)$

(b) A histogram of the ages of all patients aged between 20 and 50 years admitted to hospital for asthma might be used. If the uniform model is correct, the heights of the bars should be roughly equal.

Solution 3.5

(a) Let $X = 1$ if a patient is admitted for less than a day, and $X = 0$ otherwise. Then $X \sim \text{Bernoulli}(p)$, where p is the probability that a patient will remain in hospital for less than a day. If it is assumed that whether or not the length of a patient's stay in hospital is less than a day is independent of whether or not any other patient's stay is less than a day, then an appropriate model for R is $B(1762, p)$, where

$$p = \frac{500}{1762} \simeq 0.284.$$

(b) If $R \sim B(1762, 0.284)$, then

$$E(R) = np = 1762 \times 0.284 \simeq 500,$$

$$V(R) = np(1-p) = 1762 \times 0.284 \times (1 - 0.284) \simeq 358.$$

Solution 3.6

(a) For a Poisson distribution with parameter $\mu = 1.072$,

$$p(0) = e^{-1.072} \simeq 0.342,$$

$$p(1) = \frac{1.072 \times e^{-1.072}}{1!} \simeq 0.367,$$

$$p(2) = \frac{1.072^2 \times e^{-1.072}}{2!} \simeq 0.197.$$

These are the same as the values given in Table 3.1.

(b) For a Poisson distribution, the variance is equal to the mean. In this case, the mean is 1.072 and the variance is 1.285. The variance is slightly greater than the mean. However, the difference is not great, so perhaps the Poisson model is adequate. On the other hand, in view of the large sample size, the discrepancy might indicate that the Poisson model is not adequate. The larger variance suggests that there are more days than expected with a large number of admissions.

Solution 4.1

(a) The estimated standard error of the mean is

$$\frac{s}{\sqrt{n}} \simeq \frac{\sqrt{1.285}}{\sqrt{1643}} \simeq 0.0280.$$

(b) Figure 4.1(a) shows the p.m.f. of a discrete random variable, and is similar in shape to the histogram of numbers of daily admissions in Figure 3.11. Thus it represents the p.m.f. of the distribution of the daily number of admissions for asthma. Figure 4.1(b) shows the p.d.f. of a continuous random variable, which appears to be normal. Hence it is the sampling distribution of the mean.

Solution 4.2

(a) The sample estimate is $\widehat{p} \simeq 0.5150$, so

$$\begin{aligned} p^- &= \widehat{p} - z\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} \\ &\simeq 0.5150 - 1.96 \times \sqrt{\frac{0.5150 \times (1-0.5150)}{1761}} \\ &\simeq 0.4917, \\ p^+ &= \widehat{p} + z\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} \\ &\simeq 0.5150 + 1.96 \times \sqrt{\frac{0.5150 \times (1-0.5150)}{1761}} \\ &\simeq 0.5383. \end{aligned}$$

Hence the estimated proportion is about 0.515, and an approximate 95% confidence interval for p is $(0.492, 0.538)$. (In reporting the results, three decimal places have been kept in view of the large sample size.)

(b) The value 0.4976 is included in the 95% confidence interval, and hence is plausible at the 95% confidence level. Thus it is plausible that the gender distribution of asthma cases admitted to hospital is the same as the gender distribution of the general population. This does not suggest that females are any more or less likely than males to be admitted to hospital for asthma.

Solution 4.3

(a) Let p denote the underlying proportion of hospital admissions for asthma who are female. The null and alternative hypotheses are as follows:

$$H_0 : p = 0.4976, \quad H_1 : p \neq 0.4976.$$

(b) Under the null hypothesis, the probability that a hospital admission for asthma is female is 0.4976. So, under the null hypothesis, $X \sim B(1761, 0.4976)$.

(c) The significance probability 0.146 provides little evidence against the null hypothesis. Thus there is little evidence that the proportion of hospital admissions for asthma who are female differs from the proportion of females in the population. From this it may be concluded that there is little evidence that males and females differ in their likelihood of being admitted to hospital for asthma.

Solution 5.1

Figure 5.3(a): NO and NO_2 concentrations are positively related, but the relationship is not linear.

Figure 5.3(b): NO and CO concentrations are positively related, and the relationship seems to be linear.

Figure 5.3(c): There is no clear relationship between O_3 concentrations and PM_{10} levels.

Solution 5.2

The correlation for the data in Figure 5.1 is -0.558. The value -2.055 does not lie between -1 and $+1$, so it is not a correlation. The value 0.607 is positive, whereas the variables are negatively associated. The value -0.981 indicates a very strong association, with little scatter, which is not the case. The value -0.048 is close to 0, indicating a very weak association, which is not the case in Figure 5.1.

Solution 5.3

In each case the probability is denoted p.

(a) This probability is not a conditional probability. An estimate of p is given by

$$\widehat{p} = \frac{799}{1761} \simeq 0.454.$$

(b) This probability is not a conditional probability. An estimate of p is given by

$$\widehat{p} = \frac{9}{1761} \simeq 0.005.$$

(c) This probability is a conditional probability. An estimate of p is given by

$$\widehat{p} = \frac{9}{799} \simeq 0.011.$$

The probability may be written as

$$P(Y = \text{`long'} | X = \text{`0–19'}).$$

Solution 5.4

(a) The estimated probability that a patient is male is

$$\widehat{p} = \frac{854}{1760} \simeq 0.485.$$

(b) The estimated (conditional) probability that a patient is male is

$$\widehat{p} = \frac{497}{799} \simeq 0.622.$$

(c) The estimate of the conditional probability in part (b) is larger than the estimate of the probability in part (a). This suggests that the gender distribution may depend on the age group, and hence that age and gender may be related in hospital admissions for asthma. However, it is possible that the difference between the two estimates is due to random variation. (In fact, a formal significance test indicates that it is unlikely that this is the case.)

Solution 6.1

(a) The histogram required is shown in Figure S.4.

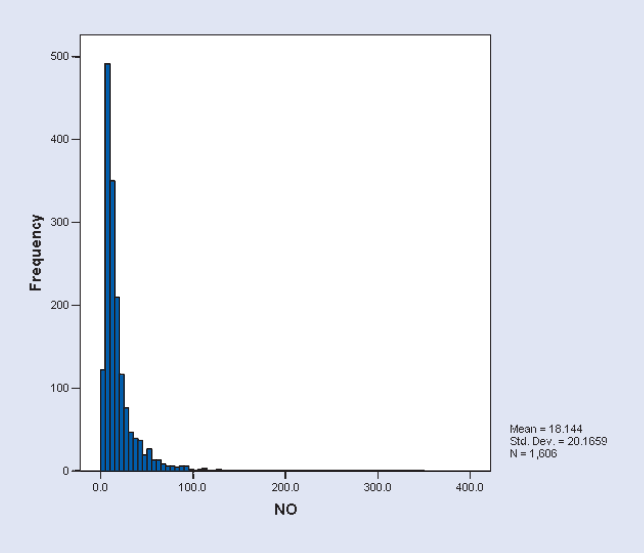

Figure S.4 A histogram of NO concentrations

The mean and standard deviation are displayed next to the histogram: the mean is $18.144 \simeq 18.1$, and the standard deviation is $20.1659 \simeq 20.2$.

(b) The histogram has a long tail to the right, and the mode is close to zero. Of the probability models described in Section 3, the exponential model is perhaps the most suitable (or least unsuitable).

(c) The mean and standard deviation of an exponential distribution are equal. The sample values you obtained in part (a) are similar, which supports the choice of an exponential model. However, the mode of an exponential distribution is zero, whereas in Figure S.4 values in the interval 0–5 are much less frequent than values in the interval 5–10. Thus the exponential distribution may not be appropriate for these data.

Solution 6.2

The default histogram is shown in Figure S.5.

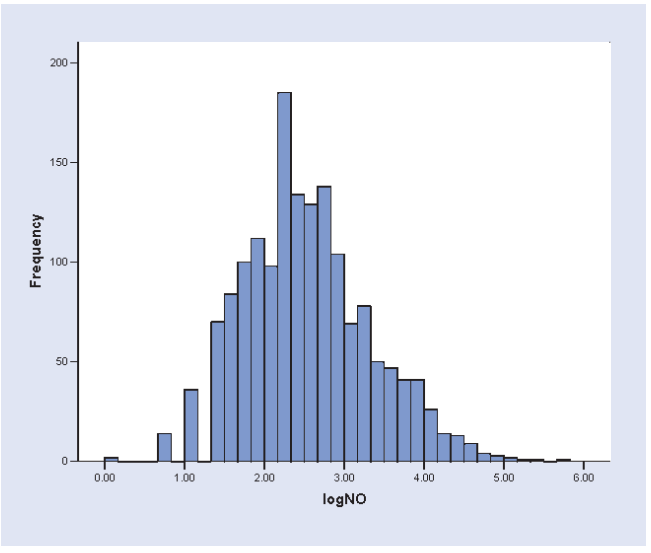

Figure S.5 The default histogram

The histogram is roughly symmetric around a single mode. A reasonable model for `logNO` is a normal distribution.

Solution 6.6

For `PM10group` $= 1$ — that is, on days when the average PM_{10} level is lower than or equal to the median — the mean number of asthma cases is 1.10, with 95% confidence interval $(1.03, 1.18)$.

For `PM10group` $= 2$ — that is, on days when the average PM_{10} level is greater than the median — the mean number of asthma cases is 1.09, with 95% confidence interval $(1.00, 1.17)$.

The mean number of asthma cases on days with high average PM_{10} levels is slightly lower than that on days with low average PM_{10} levels. This does not support the hypothesis that average daily PM_{10} levels are positively associated with asthma.

Solutions to Exercises

Solution 1.1

(a) A histogram would be appropriate for displaying the differences. Alternatively, a line plot of the difference by year could be used.

(b) The differences arranged in order of increasing size are as follows.

0.04 0.04 0.05 0.05 0.06 0.06 0.07

(c) The mean is 0.053, and the median is 0.05.

(d) The standard deviation is 0.011, and the variance is 0.00012.

(e) All the differences are positive, indicating that the mercury contamination level in plaice was higher in the Irish Sea than in the North Sea in each of the seven years. However, in view of the small sample size, the possibility that this pattern is due to chance cannot be ruled out without further analysis.

Solution 3.1

Either

$$\begin{aligned} P(X \leq 2) &= P(X = 0) + P(X = 1) + P(X = 2) \\ &= 0.1 + 0.2 + 0.4 \\ &= 0.7 \end{aligned}$$

or

$$\begin{aligned} P(X \leq 2) &= 1 - P(X = 3) \\ &= 1 - 0.3 \\ &= 0.7. \end{aligned}$$

Either

$$\begin{aligned} P(X > 0) = 1 - P(X \leq 0) &= 1 - P(X = 0) \\ &= 1 - 0.1 \\ &= 0.9 \end{aligned}$$

or

$$\begin{aligned} P(X > 0) &= P(X = 1) + P(X = 2) + P(X = 3) \\ &= 0.2 + 0.4 + 0.3 \\ &= 0.9. \end{aligned}$$

Solution 3.2

(a) Since 6.2 is the 0.75-quantile of X, $P(X \leq 6.2) = 0.75$. Hence

$$P(X > 6.2) = 1 - P(X \leq 6.2) = 1 - 0.75 = 0.25.$$

So the statement is true.

(b) Since $0.1 < 0.5$, the 0.1-quantile of X, $q_{0.1}$, is less than the 0.5-quantile of X, which is 4.6. So the statement is true.

(c) Since $0.8 > 0.25$, the 0.8-quantile of X, $q_{0.8}$, is greater than the 0.25-quantile of X, which is 2.3. So the statement is false.

Solution 3.3

The histogram for X in Figure 3.12(a) is roughly symmetric with a single clear peak. Thus a normal distribution would be an appropriate choice of probability model.

The histogram for Y in Figure 3.12(b) does not have a clear peak, and the bars are of similar height. Thus a continuous uniform distribution would be an appropriate choice of probability model.

Solution 4.1

(a) Let μ denote the (population) mean difference between the mercury contamination levels in the Irish Sea and the North Sea. For a 95% confidence interval, the 0.975-quantile of the standard normal distribution is required, namely $z = 1.96$. An approximate 95% confidence interval for μ is given by (μ^-, μ^+), where

$$\mu^- = \widehat{\mu} - z\frac{s}{\sqrt{n}} = 0.053 - 1.96 \times \frac{0.011}{\sqrt{7}} \simeq 0.045,$$

$$\mu^+ = \widehat{\mu} + z\frac{s}{\sqrt{n}} = 0.053 + 1.96 \times \frac{0.011}{\sqrt{7}} \simeq 0.061.$$

Thus the 95% z-interval for the mean difference is $(0.045, 0.061)$.

(b) The approximation involved in calculating z-intervals improves as the sample size increases. Here the sample size n is 7, which is not large. Thus the confidence level of the confidence interval calculated in part (a) may not be accurate. (A different method, the t-interval, which may be more accurate in small samples, gives the 95% confidence interval $(0.043, 0.063)$. So the z-interval is quite good even with this small sample size.)

(c) Let μ denote the mean difference between the mercury contamination levels in plaice in the two sea areas (Irish Sea − North Sea). The null and alternative hypotheses are

$$H_0 : \mu = 0, \quad H_1 : \mu \neq 0.$$

(d) The p value is less than 0.001, so there is strong evidence against the null hypothesis of zero difference between the mean contamination levels. Thus there is strong evidence that there is a difference between the mean contamination levels. Since the observed mean difference (Irish Sea − North Sea) was 0.053, this suggests that the mean mercury contamination level in plaice is higher in the Irish Sea than in the North Sea.

Solution 4.2

(a) The sample proportion of young children among hospital admissions for asthma is

$$\widehat{p} = \frac{452}{1762} \simeq 0.2565.$$

The estimated standard error of $\widehat{p}$ is

$$\widehat{\sigma} = \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} = \sqrt{\frac{0.2565 \times (1 - 0.2565)}{1762}} \simeq 0.0104.$$

For a 95% confidence interval, the 0.975-quantile of the standard normal distribution is required. So $z = 1.96$. Hence the 95% confidence limits of the z-interval are

$$p^- \simeq 0.2565 - 1.96 \times 0.0104 \simeq 0.2361,$$

$$p^+ \simeq 0.2565 + 1.96 \times 0.0104 \simeq 0.2769.$$

Thus the estimated proportion of children aged between 0 and 6 years among hospital admissions for asthma is about 0.257, with approximate 95% confidence interval $(0.236, 0.277)$.

(b) Let p denote the underlying proportion of children aged between 0 and 6 years among hospital admissions for asthma. The null and alternative hypotheses are

$$H_0 : p = 0.25, \quad H_1 : p \neq 0.25.$$

(c) The significance probability p is 0.545. This indicates that there is little evidence against the null hypothesis that the underlying proportion of children aged between 0 and 6 years among hospital admissions for asthma is 0.25. In particular, there is little evidence that the underlying proportion is more than a quarter.

Solution 5.1

(a) The two variables are positively and (roughly) linearly related.

(b) The correlation is 0.516. The first value is negative, indicating a negative relationship; the second value is greater than 1, so it is not a correlation; the third indicates a very strong relationship between the two variables, which is not the case here; and the last indicates virtually no relationship between the two variables.

Solution 5.2

(a) **(i)** This probability is not a conditional probability. An estimate of the probability is

$$\widehat{p} = \frac{250}{1760} \simeq 0.142.$$

(ii) This probability is not a conditional probability. An estimate of the probability is

$$\widehat{p} = \frac{81}{1760} \simeq 0.046.$$

(iii) This probability is a conditional probability. An estimate of the probability is

$$\widehat{p} = \frac{81}{854} \simeq 0.095.$$

(b) The probabilities estimated in parts (a)(i) and (a)(iii) are relevant for investigating a possible association between age and gender. If age and gender are unrelated, then the underlying probabilities should be equal.

Solution 7.1

(a) The **Compute Variable** dialogue box is discussed in Activity 6.1. SPSS returns a missing value whenever either the value for the Irish Sea or the value for the North Sea is missing.

(b) The histogram required is shown in Figure S.6.

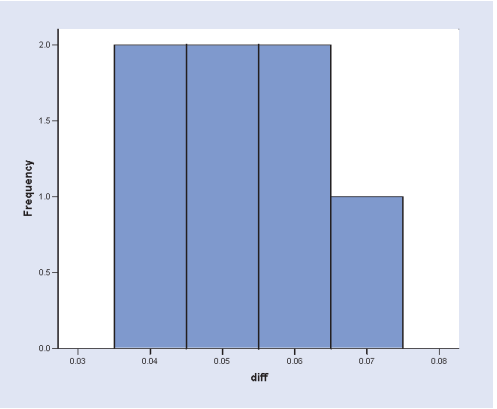

Figure S.6 A histogram of differences in mercury contamination

Altering the bin widths of a histogram is described in Activity 2.11.

(c) The estimated mean difference is 0.0529, and the 95% confidence interval is $(0.0426, 0.0631)$. These may be obtained using **Explore...**, as described in Activity 6.5.

(d) Data on average mercury concentrations in plaice were available for seven years for both the Irish Sea and the North Sea. This analysis is based on seven differences. The mean difference (Irish Sea minus North Sea) was 0.053, with 95% confidence interval $(0.043, 0.063)$. The confidence interval is located well above zero, suggesting that mercury concentrations in plaice are higher in the Irish Sea than in the North Sea. (This confidence interval is a t-interval. See the margin note at the end of Activity 6.5, and also the solution to Exercise 4.1(b).)

Solution 7.2

(a) Producing a scatterplot is described in Activity 2.10. The required scatterplot is shown in Figure S.7.

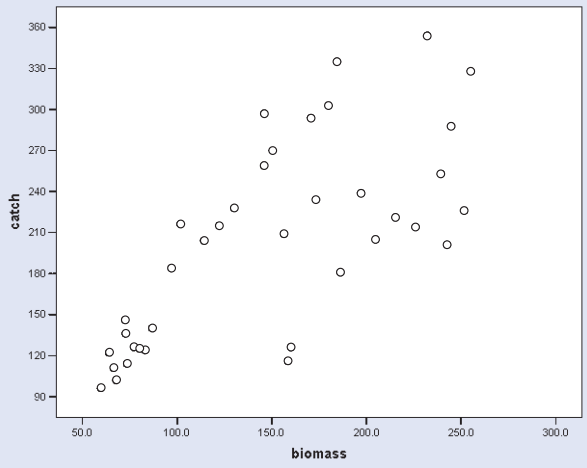

Figure S.7 A scatterplot of cod catch against biomass

The scatterplot suggests a positive relationship between catch and biomass. The relationship looks as though it may be roughly linear.

(b) Use **Bivariate Correlations** as described in Activity 6.7. The correlation coefficient is 0.714 and the p value is reported as .000, so $p < 0.0005$.

(c) There is strong evidence against the null hypothesis of zero correlation. So it appears that the cod catch and spawning stock biomass are related.

Solution 7.3

(a) Transforming variables using **Compute...** is described in Activities 6.2 and 6.3. (Enter `SQRT(NO2)` in the **Numeric Expression** field.) Obtaining histograms is described in Activity 2.11. The default histograms of nitrogen dioxide concentrations and their square roots are shown in Figures S.8 and S.9.

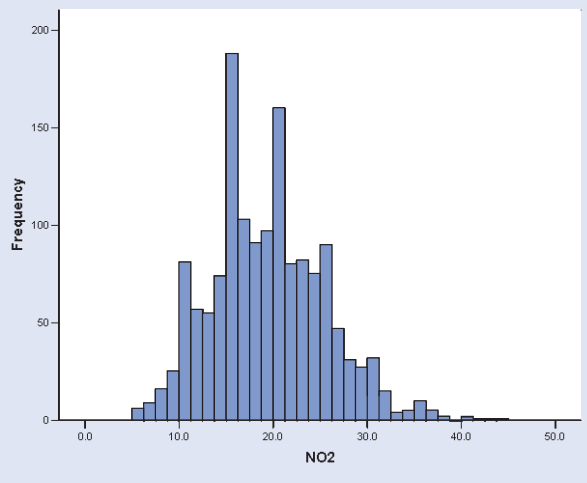

Figure S.8 A histogram of `NO2`

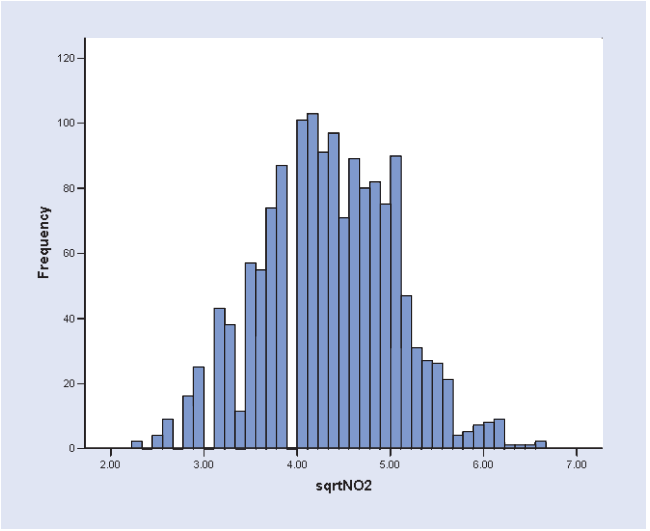

Figure S.9 A histogram of `sqrtNO2`

Calculating the skewness is described in Activity 2.12. (Use **Analyze** > **Descriptive Statistics** > **Frequencies ...**.) The skewness of `NO2` is 0.436, that of its square root is −0.028.

(b) `NO2` is slightly positively skewed, whereas its square root is roughly symmetric, with a single clear peak. A normal model appears appropriate for the square root of `NO2`.

(c) The scatterplot of `NO` against `sqrtNO2` is as shown in Figure S.10.

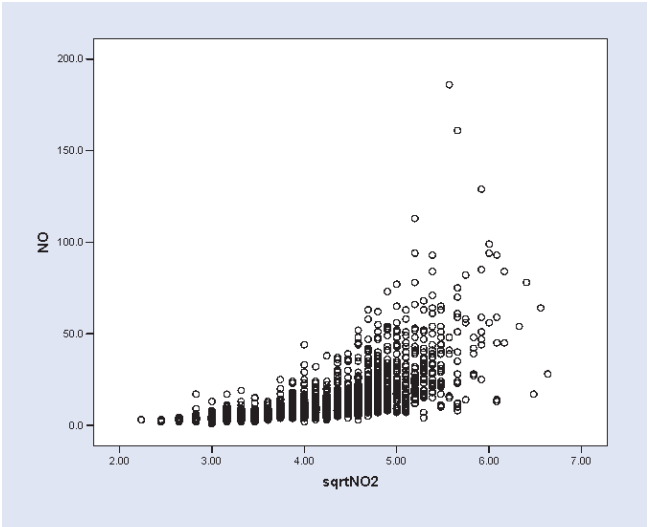

Figure S.10 A scatterplot of `NO` against `sqrtNO2`

Use **Compute...** (as described in Activity 6.2) to create a variable `logNO` containing the natural logarithm of `NO`. The scatterplot of `logNO` against `sqrtNO2` is shown in Figure S.11.

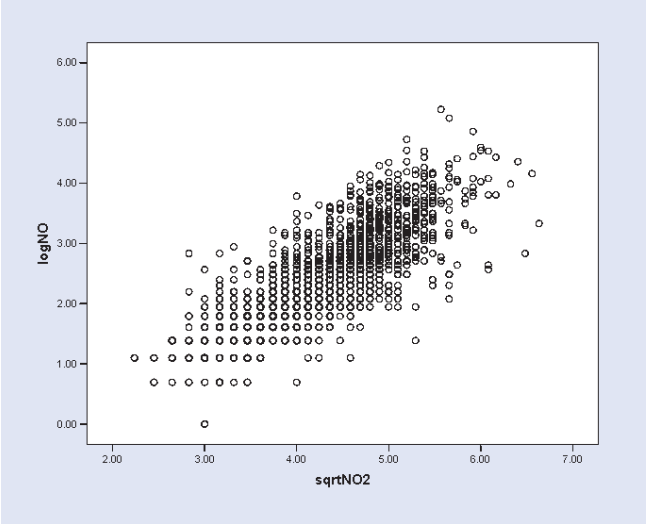

Figure S.11 A scatterplot of `logNO` against `sqrtNO2`

(The layered effect in the scatterplot is due to rounding and may be ignored.)

(d) Calculating the Pearson correlation is described in Activity 6.7. (Use **Analyze** > **Correlate** > **Bivariate...**.) The correlation between `NO` and `sqrtNO2` is 0.625, and that between `logNO` and `sqrtNO2` is 0.769.

(e) The correlation between `sqrtNO2` and `logNO` is higher than that between `sqrtNO2` and `NO` because the relationship between `sqrtNO2` and `logNO` is linear, as shown in Figure S.11, while that between `sqrtNO2` and `NO` is not (see Figure S.10). The Pearson correlation coefficient is a measure of the strength of a linear relationship between two variables.

Solution 7.4

(a) Use **Explore** as described in Activity 6.6. The (unrounded) statistics, as given by SPSS, are as follows.

	Mean	SD	95% CI
Males	1.738	1.6208	(1.629, 1.847)
Females	2.025	2.0334	(1.892, 2.157)

(b) The estimated mean stay is longer for women than for men, and the length of stay also appears more variable for women than for men. The 95% confidence intervals do not overlap, suggesting that the difference is not attributable to chance. However, to assess the null hypothesis that men and women have the same mean length of stay, a significance test is required.

(c) A p value of 0.001 provides strong evidence against the null hypothesis that the underlying mean lengths of stay are the same, and hence provides strong evidence that they are different. Examination of the sample means suggests that the mean length of stay is greater for women than for men.

Index