CHAPTER 3

Displaying the data

3.1 **Introduction**
3.2 **Frequencies, frequency distributions and histograms**
Frequencies (categorical variables)
Frequency distributions (numerical variables)
Histograms
Frequency polygon
Frequency distribution of the population
Shapes of frequency distributions
3.3 **Cumulative frequency distributions, quantiles and percentiles**
Cumulative frequency distributions
Median and quartiles
Quantiles and percentiles
3.4 **Displaying the association between two variables**
Cross tabulations
Scatter plots
3.5 **Displaying time trends**

3.1 INTRODUCTION

With ready access to statistical software, there is a temptation to jump straight into complex analyses. This should be avoided. An essential first step of an analysis is to summarize and display the data. The familiarity with the data gained through doing this is invaluable in developing an appropriate analysis plan (see Chapter 38). These initial displays are also valuable in identifying **outliers** (unusual values of a variable) and revealing possible errors in the data, which should be checked and, if necessary, corrected.

This chapter describes simple tabular and graphical techniques for displaying the distribution of values taken by a single variable, and for displaying the association between the values of two variables. Diagrams and tables should always be clearly labelled and self-explanatory; it should not be necessary to refer to the text to understand them. At the same time they should not be cluttered with too much detail, and they must not be misleading.

3.2 FREQUENCIES, FREQUENCY DISTRIBUTIONS AND HISTOGRAMS

Frequencies (categorical variables)

Summarizing categorical variables is straightforward, the main task being to count the number of observations in each category. These counts are called **frequencies**. They are often also presented as **relative frequencies**; that is as proportions or percentages of the total number of individuals. For example, Table 3.1 summarizes the method of delivery recorded for 600 births in a hospital. The

Table 3.1 Method of delivery of 600 babies born in a hospital.

Method of delivery	No. of births	Percentage
Normal	478	79.7
Forceps	65	10.8
Caesarean section	57	9.5
Total	600	100.0

variable of interest is the method of delivery, a categorical variable with three categories: normal delivery, forceps delivery, and caesarean section.

Frequencies and relative frequencies are commonly illustrated by a **bar chart** (also known as a **bar diagram**) or by a **pie chart**. In a bar chart the lengths of the bars are drawn proportional to the frequencies, as shown in Figure 3.1. Alternatively the bars may be drawn proportional to the percentages in each category; the shape is not changed, only the labelling of the scale. In either case, for ease of reading it is helpful to write the actual frequency and/or percentage to the right of the bar. In a pie chart (see Figure 3.2), the circle is divided so that the areas of the sectors are proportional to the frequencies, or equivalently to the percentages.

Frequency distributions (numerical variables)

If there are more than about 20 observations, a useful first step in summarizing a numerical (quantitative) variable is to form a **frequency distribution**. This is a table showing the number of observations at different values or within certain ranges. For a discrete variable the frequencies may be tabulated either for each value of the variable or for groups of values. With continuous variables, groups have to be formed. An example is given in Table 3.2, where haemoglobin has been measured

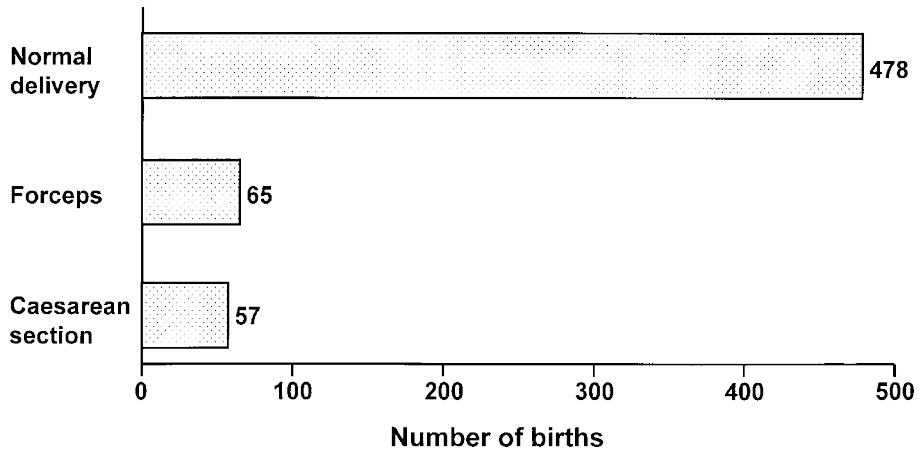


Fig. 3.1 Bar chart showing method of delivery of 600 babies born in a hospital.

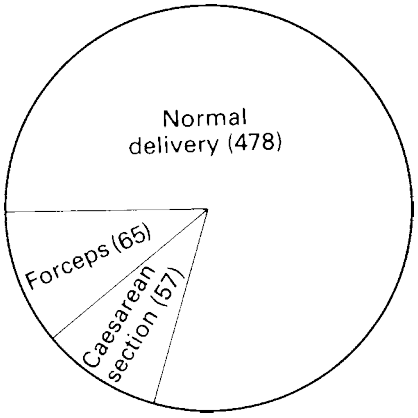


Fig. 3.2 Pie chart showing method of delivery of 600 babies born in a hospital.

to the nearest 0.1 g/100 ml and the group 11–, for example, contains all measurements between 11.0 and 11.9 g/100 ml inclusive.

When forming a frequency distribution, the first things to do are to count the number of observations and to identify the lowest and highest values. Then decide

Table 3.2 Haemoglobin levels in g/100 ml for 70 women.

(a) Raw data with the highest and lowest values underlined.

10.2	13.7	10.4	14.9	11.5	12.0	11.0
13.3	12.9	12.1	9.4	13.2	10.8	11.7
10.6	10.5	13.7	11.8	14.1	10.3	13.6
12.1	12.9	11.4	12.7	10.6	11.4	11.9
9.3	13.5	14.6	11.2	11.7	10.9	10.4
12.0	12.9	11.1	8.8	10.2	11.6	12.5
13.4	12.1	10.9	11.3	14.7	10.8	13.3
11.9	11.4	12.5	13.0	11.6	13.1	9.7
11.2	15.1	10.7	12.9	13.4	12.3	11.0
14.6	11.1	13.5	10.9	13.1	11.8	12.2

(b) Frequency distribution.

Haemoglobin (g/100 ml)	No. of women	Percentage
8–	1	1.4
9–	3	4.3
10–	14	20.0
11–	19	27.1
12–	14	20.0
13–	13	18.6
14–	5	7.1
15–15.9	1	1.4
Total	70	100.0

whether the data should be grouped and, if so, what grouping interval should be used. As a rough guide one should aim for 5–20 groups, depending on the number of observations. If the interval chosen for grouping the data is too wide, too much detail will be lost, while if it is too narrow the table will be unwieldy. The starting points of the groups should be round numbers and, whenever possible, all the intervals should be of the same width. There should be no gaps between groups. The table should be labelled so that it is clear what happens to observations that fall on the boundaries.

For example, in Table 3.2 there are 70 haemoglobin measurements. The lowest value is 8.8 and the highest 15.1 g/100 ml. Intervals of width 1 g/100 ml were chosen, leading to eight groups in the frequency distribution. Labelling the groups 8–, 9–, . . . is clear. An acceptable alternative would have been 8.0–8.9, 9.0–9.9 and so on. Note that labelling them 8–9, 9–10 and so on would have been confusing, since it would not then be clear to which group a measurement of 9.0 g/100 ml, for example, belonged.

Once the format of the table is decided, the numbers of observations in each group are counted. If this is done by hand, mistakes are most easily avoided by going through the data in order. For each value, a mark is put against the appropriate group. To facilitate the counting, these marks are arranged in groups of five by putting each fifth mark horizontally through the previous four (~~1111~~); these groups are called **five-bar gates**. The process is called **tallying**.

As well as the number of women, it is useful to show the percentage of women in each of the groups.

Histograms

Frequency distributions are usually illustrated by **histograms**, as shown in Figure 3.3 for the haemoglobin data. Either the frequencies or the percentages may be used; the shape of the histogram will be the same.

The construction of a histogram is straightforward when the grouping intervals of the frequency distribution are all equal, as is the case in Figure 3.3. If the intervals are of different widths, it is important to take this into account when drawing the histogram, otherwise a distorted picture will be obtained. For example, suppose the two highest haemoglobin groups had been combined in compiling Table 3.2(b). The frequency for this combined group (14.0–15.9 g/100 ml) would be six, but clearly it would be misleading to draw a rectangle of height six from 14 to 16 g/100 ml. Since this interval would be twice the width of all the others, the correct height of the line would be three, half the total frequency for this group. This is illustrated by the dotted line in Figure 3.3. The general rule for drawing a histogram when the intervals are not all the same width is to make the heights of the rectangles proportional to the frequencies divided by the widths, that is to make the areas of the histogram bars proportional to the frequencies.

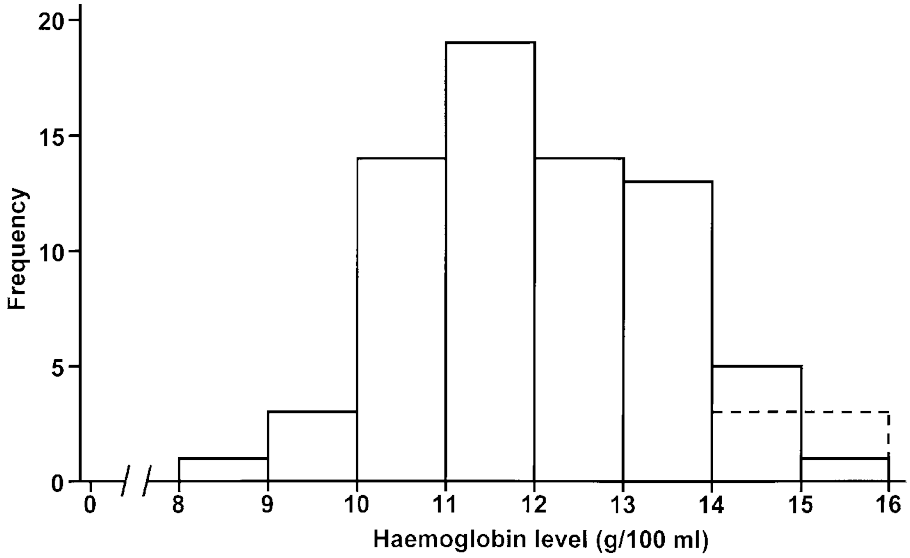


Fig. 3.3 Histogram of haemoglobin levels of 70 women.

Frequency polygon

An alternative but less common way of illustrating a frequency distribution is a **frequency polygon**, as shown in Figure 3.4. This is particularly useful when comparing two or more frequency distributions by drawing them on the same diagram. The polygon is drawn by imagining (or lightly pencilling) the histogram and joining

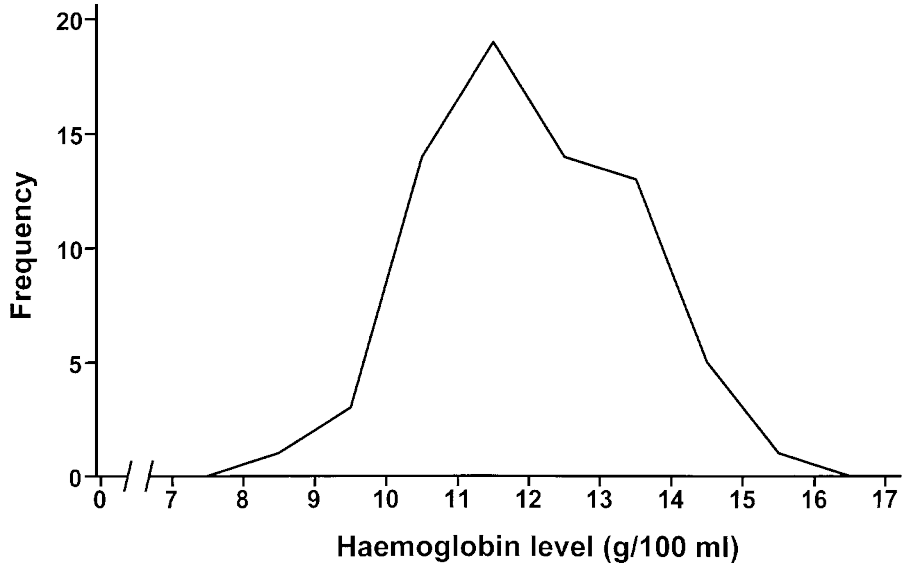


Fig. 3.4 Frequency polygon of haemoglobin levels of 70 women.

the midpoints of the tops of its rectangles. The endpoints of the resulting line are then joined to the horizontal axis at the midpoints of the groups immediately below and above the lowest and highest non-zero frequencies respectively. For the haemoglobin data, these are the groups 7.0–7.9 and 16.0–16.9 g/100 ml. The frequency polygon in Figure 3.4 is therefore joined to the axis at 7.5 and 16.5 g/100 ml.

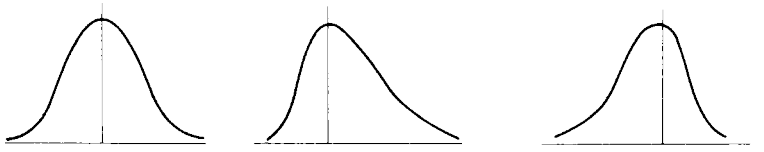
Frequency distribution of the population

Figures 3.3 and 3.4 illustrate the frequency distribution of the haemoglobin levels of a sample of 70 women. We use these data to give us information about the distribution of haemoglobin levels among women in general. For example, it seems uncommon for a woman to have a level below 9.0 g/100 ml or above 15.0 g/100 ml. Our confidence in drawing general conclusions from the data depends on how many individuals were measured. The larger the sample, the finer the grouping interval that can be chosen, so that the histogram (or frequency polygon) becomes smoother and more closely resembles the distribution of the total population. At the limit, if it were possible to ascertain the haemoglobin levels of the whole population of women, the resulting diagram would be a smooth curve.

Shapes of frequency distributions

Figure 3.5 shows three of the most common shapes of frequency distributions. They all have high frequencies in the centre of the distribution and low frequencies at the two extremes, which are called the **upper** and **lower tails** of the distribution. The distribution in Figure 3.5(a) is also **symmetrical** about the centre; this shape of curve is often described as 'bell-shaped'. The two other distributions are asymmetrical or **skewed**. The upper tail of the distribution in Figure 3.5(b) is longer than the lower tail; this is called **positively skewed** or skewed to the right. The distribution in Figure 3.5(c) is **negatively skewed** or skewed to the left.

All three distributions in Figure 3.5 are **unimodal**, that is they have just one peak. Figure 3.6(a) shows a **bimodal** frequency distribution, that is a distribution with two peaks. This is occasionally seen and usually indicates that the data are a mixture of



(a) Symmetrical and bell-shaped, e.g. height

(b) Positively skewed or skewed to the right, e.g. triceps skinfold measurement

(c) Negatively skewed or skewed to the left, e.g. period of gestation

Fig. 3.5 Three common shapes of frequency distributions with an example of each.

(a) Bimodal, e.g. hormone levels of males and females

(b) Reverse J-shaped, e.g. survival time after diagnosis of lung cancer

(c) Uniform, e.g. month of occurrence of disease with no seasonal pattern

Fig. 3.6 Three less-common shapes of frequency distributions with an example of each.

two separate distributions. Also shown in Figure 3.6 are two other distributions that are sometimes found, the **reverse J-shaped** and the **uniform** distributions.

3.3 CUMULATIVE FREQUENCY DISTRIBUTIONS, QUANTILES AND PERCENTILES

Cumulative frequency distributions

Frequency distributions (and histograms) indicate the way data are distributed over a range of values, by showing the number or percentage of individuals within each group of values. **Cumulative distributions** start from the lowest value and show how the number and percentage of individuals accumulate as the values increase. For example, the cumulative frequency distribution for the first five observations of haemoglobin levels is shown in Table 3.3. There were 70 observations, so each represents $100/70 = 1.43\%$ of the total distribution. Rounding to one decimal place, the first observation (8.8 g/100 ml) corresponds to 1.4% of the distribution, the first and second observations to 2.9% of the distribution, and so on. Table 3.3 shows the values of these **cumulative percentages**, for different observations in the range of observed haemoglobin levels in the 70 women. A total of four women (5.7%) had levels below 10 g/100 ml. Similarly, 18 women (25.7%) had haemoglobin levels below 11 g/100 ml.

The **cumulative frequency distribution** is illustrated in Figure 3.7. This is drawn as a **step function**: the vertical jumps correspond to the increases in the cumulative percentages at each observed haemoglobin level. (Another example of plots that use step functions is **Kaplan–Meier** plots of cumulative survival probabilities over time; see Section 26.3.) Cumulative frequency curves are steep where there is a concentration of values, and shallow where values are sparse. In this example, where the majority of haemoglobin values are concentrated in the centre of the distribution, the curve is steep in the centre, and shallow at low and high values. If the haemoglobin levels were evenly distributed across the range, then the cumulative frequency curve would increase at a constant rate; all the steps would be the same width as well as the same height. An advantage of cumulative frequency distributions is that they display the shape of the distribution without the need for grouping, as required in plotting histograms (see Section 3.2). However the shape of a distribution is usually more clearly seen in a histogram.

Table 3.3 Cumulative percentages for different ranges of haemoglobin levels of 70 women.

Observation	Cumulative percentage	Haemoglobin level (g/100 ml)		Quartile
1	1.4	8.8	Minimum = 8.8	1
2	2.9	9.3		1
3	4.3	9.4		1
4	5.7	9.7		1
5	7.1	10.2		
⋮	⋮	⋮		
15	21.4	10.8		1
16	22.9	10.9		1
17	24.3	10.9	Lower quartile = 10.9	1
18	25.7	10.9		1
19	27.1	11.0		2
20	28.6	11.0		2
⋮	⋮	⋮		
33	47.1	11.7		2
34	48.6	11.8		2
35	50.0	11.8	Median = 11.85	2
36	51.4	11.9		3
37	52.9	11.9		3
38	54.3	12.0		3
⋮	⋮	⋮		
50	71.4	12.9		3
51	72.9	12.9		3
52	74.3	13.0		3
53	75.7	13.1	Upper quartile = 13.1	4
54	77.1	13.1		4
55	78.6	13.2		4
⋮	⋮	⋮		
66	94.3	14.6		4
67	95.7	14.6		4
68	97.1	14.7		4
69	98.6	14.9		4
70	100	15.1	Maximum = 15.1	4

Median and quartiles

Cumulative frequency distributions are useful in recoding a numerical variable into a categorical variable. The **median** is the midway value; half of the distribution lies below the median and half above it.

$$\text{Median} = \frac{(n+1)\text{th}}{2} \text{ value of the ordered observations}$$
$$(n = \text{number of observations})$$

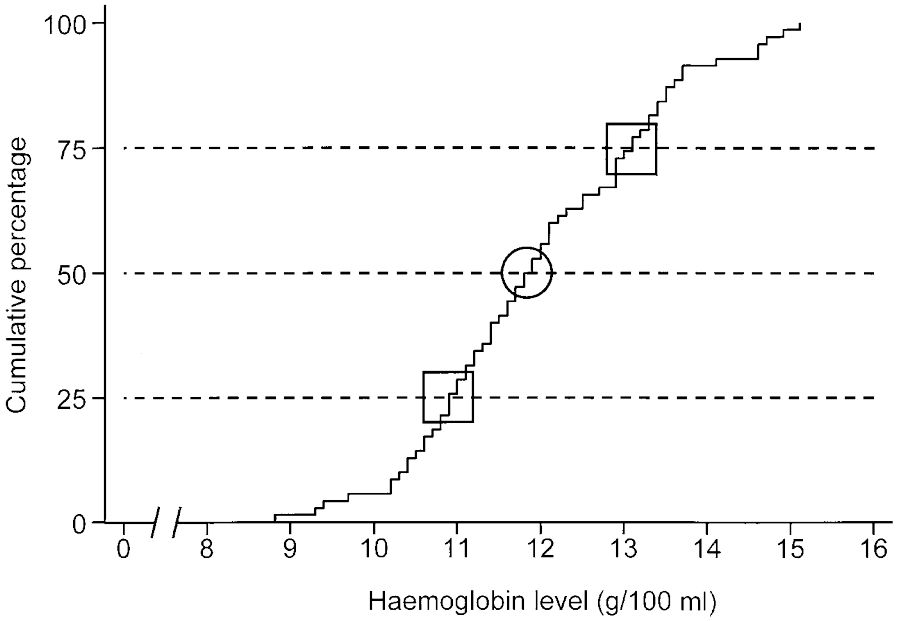


Fig. 3.7 Cumulative frequency distribution of haemoglobin levels of 70 women, with the median marked by a circle, and lower and upper quartiles marked by squares.

For the haemoglobin data, the median is the $71/2 = 35.5$th observation and so we take the average of the 35th and 36th observations. Thus the median is $(11.8+11.9)/2 = 11.85$, as shown in Table 3.3. Calculation of the median is also described in Section 4.2. When the sample size is reasonably large, the median can be estimated from the cumulative frequency distribution; it is the haemoglobin value corresponding to the point where the 50% line crosses the curve, as shown in Figure 3.7.

Also marked on Figure 3.7 are the two points where the 25% and 75% lines cross the curve. These are called the **lower** and **upper quartiles** of the distribution, respectively, and together with the median they divide the distribution into four equally-sized groups.

$$\text{Lower quartile} = \frac{(n+1)\text{th}}{4} \text{ value of the ordered observations}$$

$$\text{Upper quartile} = \frac{3 \times (n+1)\text{th}}{4} \text{ value of the ordered observations}$$

In the haemoglobin data, the lower quartile is the $71/4 = 17.75$th observation. This is calculated by taking three quarters of the difference between the 17th and 18th observations and adding it to the 17th observation. Since both the 17th and 18th observations equal 10.9 g/100 ml, so does the lower quartile, as shown

in Table 3.3. Similarly, $3 \times 71/4 = 53.25$, and since both the 53rd and 54th observations equal 13.1 g/100 ml, so does the upper quartile.

The **range** of the distribution is the difference between the minimum and maximum values. From Table 3.3, the minimum and maximum values for the haemoglobin data are 8.8 and 15.1 g/100 ml, so the range is $15.1 - 8.8 = 6.3$ g/100 ml. The difference between the lower and upper quartiles of the haemoglobin data is 2.2 g/100 ml. This is known as the **interquartile range**.

Range = highest value − lowest value

Interquartile range = upper quartile − lower quartile

A useful plot, based on these values, is a **box and whiskers plot**, as shown in Figure 3.8. The box is drawn from the lower quartile to the upper quartile; its length gives the interquartile range. The horizontal line in the middle of the box represents the median. Just as a cat's whiskers mark the full width of its body, the 'whiskers' in this plot mark the full extent of the data. They are drawn on either end of the box to the minimum and maximum values.

The right hand column of Table 3.3 shows how the median and lower and upper quartiles may be used to divide the data into equally sized groups called **quartiles**.

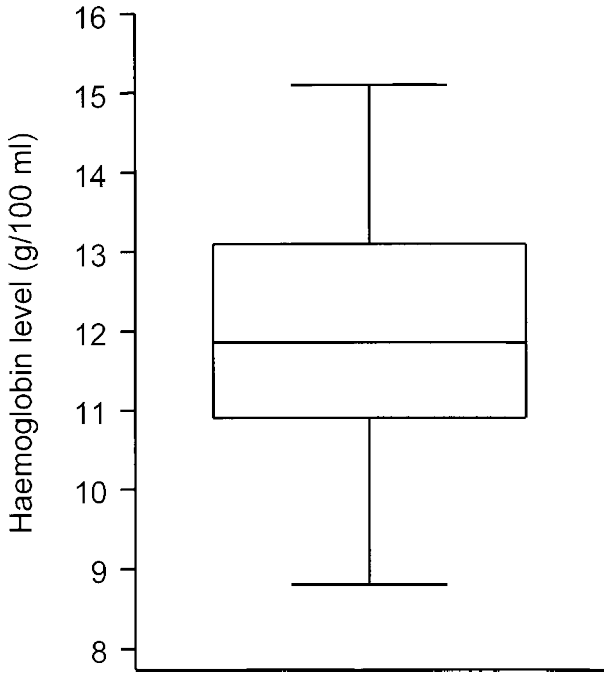


Fig. 3.8 Box and whiskers plot of the distribution of the haemoglobin levels of 70 women.

Values between 8.8 and 10.9 g/100 ml are in the first quartile, those between 11 and 11.8 g/100 ml are in the second quartile and so on. Note that equal values should always be placed in the same group, even if the groups are then of slightly different sizes.

Quantiles and percentiles

Equal-sized divisions of a distribution are called **quantiles**. For example, we may define **tertiles**, which divide the data into three equally-sized groups, and **quintiles**, which divide them into five. An example was described in Section 2.3, where the range of values observed for average monthly income was used to divide the sample into five equally-sized income groups, and a new variable 'income group' created with '1' corresponding to the least affluent group in the population and '5' to the most affluent group. Quintiles are estimated from the intersections with the cumulative frequency curve of lines at 20%, 40%, 60% and 80%. Divisions into ten equally sized groups are called **deciles**.

More generally, the kth **percentile** (or **centile** as it is also called) is the point below which k% of the values of the distribution lie. For a distribution with n observations, it is defined as:

$$k\text{th percentile} = \frac{k \times (n+1)\text{th}}{100} \text{ value of ordered observations}$$

It can also be estimated from the cumulative frequency curve; it is the x value corresponding to the point where a line drawn at k% intersects the curve. For example, the 5% point of the haemoglobin values is estimated to be 9.6 g/100 ml.

3.4 DISPLAYING THE ASSOCIATION BETWEEN TWO VARIABLES

Having examined the distribution of a single variable, we will often wish to display the way in which the distribution of one variable relates to the distribution of another. Appropriate methods to do this will depend on the type of the two variables.

Cross tabulations

When both variables are categorical, we can examine their relationship informally by cross-tabulating them in a **contingency table**. A useful convention is for the rows of the table to correspond to the exposure values and the columns to the outcomes. For example, Table 3.4 shows the results from a survey to compare the principal water sources in 150 households in three villages in West Africa. In this example, it would be natural to ask whether the household's village affects their likely water source, so that water source is the *outcome* and village is the *exposure*.

Table 3.4 Comparison of principal sources of water used by household in three villages in West Africa.

	Water source		
Village	River	Pond	Spring
A	20	18	12
B	32	20	8
C	18	12	10

The interpretability of contingency tables can be improved by including **marginal totals** and **percentages**:

- The marginal row totals show the total number of households in each village, and the marginal columns show the total numbers using each water source.
- Percentages (or proportions) can be calculated with respect to the row variable, the column variable, or the total number of individuals. A useful guide is that the percentages should correspond to the *exposure* variable. If the exposure is the row variable, as here, then row percentages should be presented, whereas if it is the column variable then column percentages should be presented.

In Table 3.4, the exposure variable, village, is the row variable, and Table 3.5 therefore shows row percentages together with marginal (row and column) totals. We can now see that, for example, the proportion of households mainly using a river was highest in Village B, while village A had the highest proportion of households mainly using a pond. By examining the column totals we can see that overall, rivers were the principal water source for 70 (47%) of the 150 households.

Table 3.5 Comparison of principal sources of water used by households in three villages in West Africa, including marginal totals and row percentages.

	Water source			
Village	River	Pond	Spring	Total
A	20 (40%)	18 (36%)	12 (24%)	50 (100%)
B	32 (53%)	20 (33%)	8 (13%)	60 (100%)
C	18 (45%)	12 (30%)	10 (25%)	40 (100%)
Total	70 (47%)	50 (33%)	30 (20%)	150 (100%)

Scatter plots

When we wish to examine the relationship between two numerical variables, we should start by drawing a scatter plot. This is a simple graph where each pair of values is represented by a symbol whose horizontal position is determined by the value of the first variable and vertical position is determined by the value of the second variable. By convention, the outcome variable determines vertical position and the exposure variable determines horizontal position.

For example, Figure 3.9 shows data from a study of lung function among 636 children aged 7 to 10 years living in a deprived suburb of Lima, Peru. The maximum volume of air which the children could breath out in 1 second (Forced Expiratory Volume in 1 second, denoted as FEV_1) was measured using a spirometer. We are interested in how FEV_1 changes with age, so that age is the exposure variable (horizontal axis) and FEV_1 is the outcome variable (vertical axis). The plot gives the clear impression that FEV_1 increases in an approximately linear manner with age.

Scatter plots may also be used to display the relationship between a categorical variable and a continuous variable. For example, in the study of lung function we are also interested in the relationship between FEV_1 and respiratory symptoms experienced by the child over the previous 12 months. Figure 3.10 shows a scatter plot that displays this relationship.

This figure is difficult to interpret, because many of the points overlap, particularly in the group of children who did not report respiratory symptoms. One solution to this is to scatter the points randomly along the horizontal axis, a process known as '**jittering**'. This produces a clearer picture, as shown in Figure 3.11. We can now see that FEV_1 tended to be higher in children who did not report respiratory symptoms in the previous 12 months than in those who did.

An alternative way to display the relationship between a numerical variable and a discrete variable is to draw **box and whiskers plots**, as described in Section 3.3. Table 3.6 shows the data needed to do this for the two groups of children: those who did and those who did not report respiratory symptoms. All the statistics displayed are

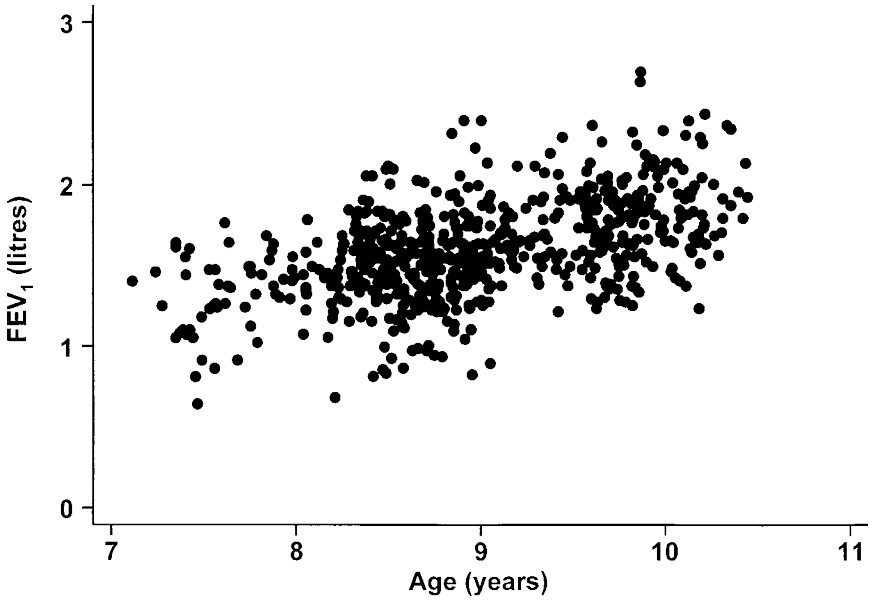


Fig. 3.9 Scatter plot showing the relationship between FEV_1 and age in 636 children living in a deprived suburb of Lima, Peru.

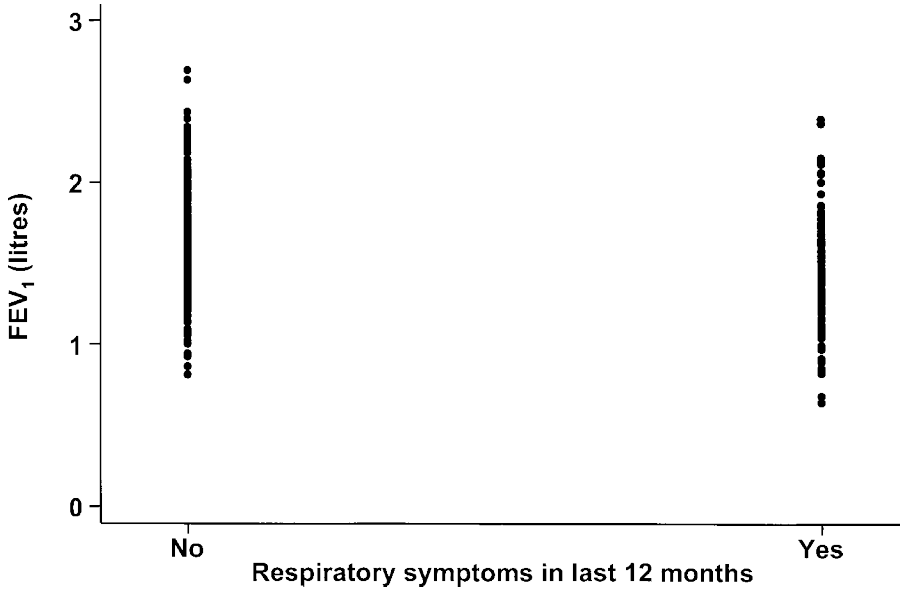


Fig. 3.10 Scatter plot showing the relationship between FEV_1 and respiratory symptoms in 636 children living in a deprived suburb of Lima, Peru.

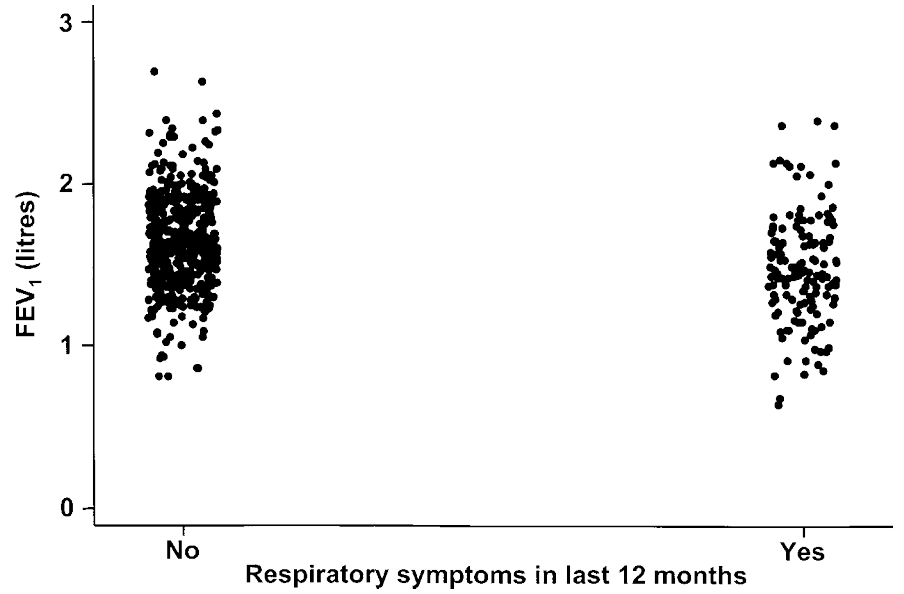


Fig. 3.11 Scatter plot showing the relationship between FEV_1 and respiratory symptoms in 636 children living in a deprived suburb of Lima, Peru. The position of the points on the horizontal axis was moved randomly ('jittered') in order to separate them.

Table 3.6 Median, interquartile range, and range of FEV_1 measurements on 636 children living in a deprived suburb of Lima, Peru, according to whether the child reported respiratory symptoms in the previous 12 months.

Respiratory symptoms in the previous 12 months	*n*	Lowest FEV_1 value	Lower quartile (25th centile)	Median	Upper quartile (75th centile)	Highest FEV_1 value
No	491	0.81	1.44	1.61	1.82	2.69
Yes	145	0.64	1.28	1.46	1.65	2.39
Totals	636	0.64	1.40	1.58	1.79	2.69

lower in children who reported symptoms. This is reflected in Figure 3.12, where all the points in the box and whiskers plot of FEV_1 values for children who reported respiratory symptoms are lower than the corresponding points in the box and whiskers plot for children who did not report symptoms.

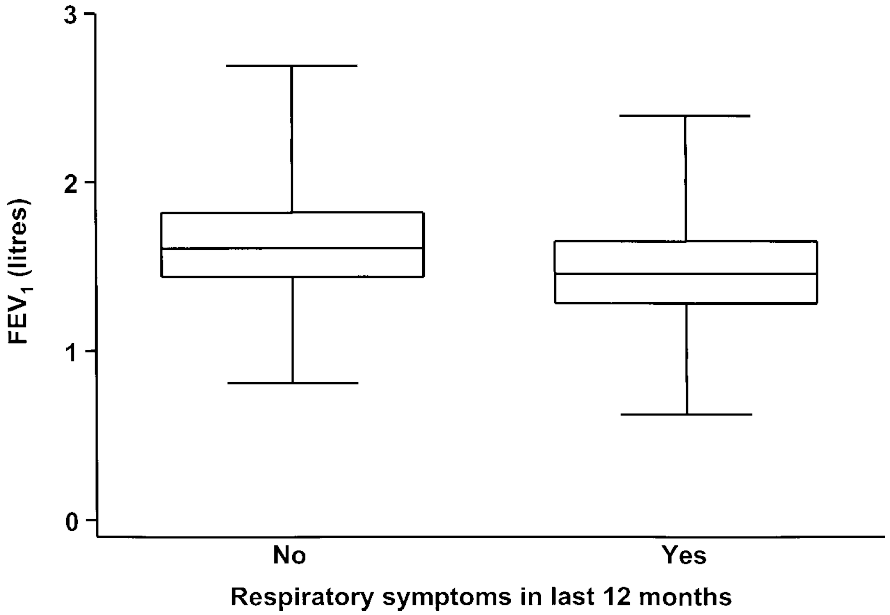


Fig. 3.12 Box and whiskers plots of the distribution of FEV_1 in 636 children living in a deprived suburb of Lima, Peru, according to whether they reported respiratory symptoms in the previous 12 months.

3.5 DISPLAYING TIME TRENDS

Graphs are also useful for displaying trends over time, such as the declines in child mortality rates that have taken place in all regions of the world in the latter half of the twentieth century, as shown in Figure 3.13. The graph also indicates the enormous differentials between regions that still remain. Note that the graph shows absolute changes in mortality rates over time. An alternative would be to

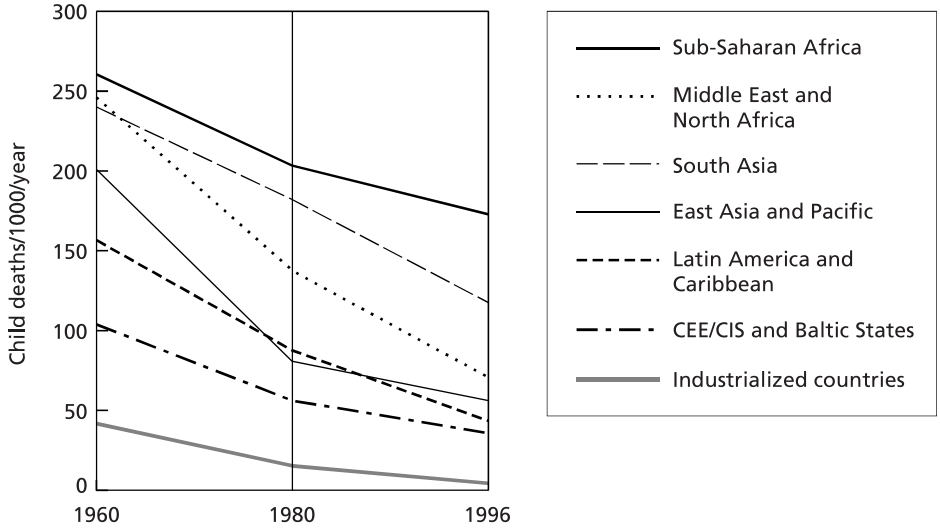


Fig. 3.13 Trends in under-five mortality rates by region of the world.

plot the logarithms of the death rates (see Chapter 13). The slopes of the lines would then show proportional declines, enabling rates of progress between regions to be readily compared.

Breaks and discontinuities in the scale(s) should be clearly marked, and avoided whenever possible. Figure 3.14(a) shows a common form of misrepresentation due to an inappropriate use of scale. The decline in infant mortality rate (IMR) has been made to look dramatic by expanding the vertical scale, while in reality the decrease over the 10 years displayed is only slight (from 22.7 to 22.1 deaths/1000 live births/year). A more realistic representation is shown in Figure 3.14(b), with the vertical scale starting at zero.

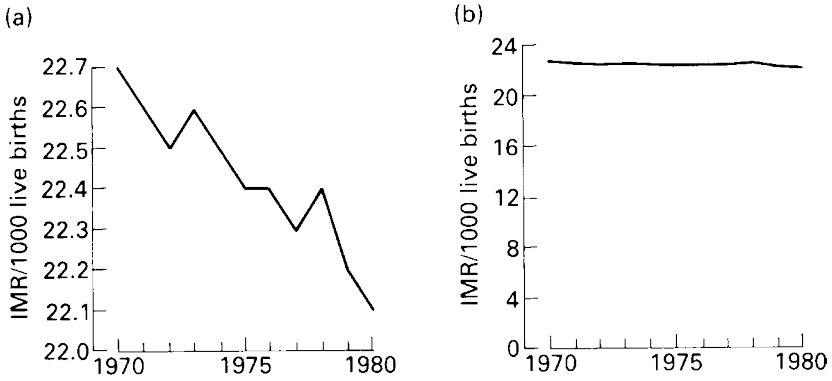


Fig. 3.14 Decline in infant mortality rate (IMR) between 1970 and 1980. (a) Inappropriate choice of scale has misleadingly exaggerated the decline. (b) Correct use of scale.

PART B

ANALYSIS OF NUMERICAL OUTCOMES

In this part of the book we describe methods for the analysis of studies where the outcome variable is **numerical**. Examples of such variables include blood pressure, antibody levels, birth weight and so on. We begin, in Chapter 4, by describing how to summarize characteristics of the distribution of a numerical variable; having defined the **mean** and **standard deviation** of a distribution, we introduce the important concept of **sampling error**. Chapter 5 describes the **normal distribution**, which occupies a central role in statistical analysis. We explain that the normal distribution is important not only because it is a good empirical description of the distribution of many variables, but also because the **sampling distribution** of a mean is normal, even when the individual observations are not normally distributed. We build on this in the next three chapters, introducing the two fundamental ways of reporting the results of a statistical analysis, **confidence intervals** (Chapters 6 and 7) and ***P*-values** (Chapters 7 and 8).

Chapter 6 deals with the analysis of a single variable. The remainder of this part of the book deals with ways of analysing the relationship between a numerical outcome (response) variable and one or more exposure (explanatory) variables. We describe how to compare means between two exposure groups (Chapters 7 and 8), and extend these methods to comparison of means in several groups using **analysis of variance** (Chapter 9) and the use of **linear regression** to examine the association between numerical outcome and exposure variables (Chapter 10). All these methods are shown to be special cases of **multiple regression**, which is described in Chapter 11.

We conclude by describing how we can examine the assumptions underlying these methods (Chapter 12), and the use of **transformations** of continuous variables to facilitate data analysis when these assumptions are violated (Chapter 13).

This page intentionally left blank

CHAPTER 4

Means, standard deviations and standard errors

4.1 **Introduction**
4.2 **Mean, median and mode**
4.3 **Measures of variation**
Range and interquartile range
Variance
Degrees of freedom
Standard deviation
Interpretation of the standard deviation
Change of units
Coefficient of variation
4.4 **Calculating the mean and standard deviation from a frequency distribution**
4.5 **Sampling variation and standard error**
Understanding standard deviations and standard errors

4.1 INTRODUCTION

A frequency distribution (see Section 3.2) gives a general picture of the distribution of a variable. It is often convenient, however, to summarize a numerical variable still further by giving just two measurements, one indicating the average value and the other the spread of the values.

4.2 MEAN, MEDIAN AND MODE

The average value is usually represented by the arithmetic mean, customarily just called the **mean**. This is simply the sum of the values divided by the number of values.

$$\text{Mean, } \bar{x}x = \frac{\Sigma x}{n}$$

where x denotes the values of the variable, Σ (the Greek capital letter sigma) means 'the sum of' and n is the number of observations. The mean is denoted by $\bar{x}x$ (spoken 'x bar').

Other measures of the average value are the **median** and the **mode**. The median was defined in Section 3.3 as the value that divides the distribution in half. If the observations are arranged in increasing order, the median is the middle observation.

$$\text{Median} = \frac{(n+1)}{2}\text{th value of ordered observations}$$

If there is an even number of observations, there is no middle one and the average of the two 'middle' ones is taken. The **mode** is the value which occurs most often.

Example 4.1

The following are the plasma volumes of eight healthy adult males:

$$2.75, 2.86, 3.37, 2.76, 2.62, 3.49, 3.05, 3.12 \text{ litres}$$

(a) $n = 8$

$\Sigma x = 2.75 + 2.86 + 3.37 + 2.76 + 2.62 + 3.49 + 3.05 + 3.12 = 24.02$ litres

Mean, $\bar{x} = \Sigma x/n = 24.02/8 = 3.00$ litres

(b) Rearranging the measurements in increasing order gives:

$$2.62, 2.75, 2.76, 2.86, 3.05, 3.12, 3.37, 3.49 \text{ litres}$$

Median $= (n + 1)/2 = 9/2 = 4.5$th value

$=$ average of 4th and 5th values

$= (2.86 + 3.05)/2 = 2.96$ litres

(c) There is no estimate of the mode, since all the values are different.

The mean is usually the preferred measure since it takes into account each individual observation and is most amenable to statistical analysis. The median is a useful descriptive measure if there are one or two extremely high or low values, which would make the mean unrepresentative of the majority of the data. The mode is seldom used. If the sample is small, either it may not be possible to estimate the mode (as in Example 4.1c), or the estimate obtained may be misleading. The mean, median and mode are, *on average*, equal when the distribution is symmetrical and unimodal. When the distribution is positively skewed, a **geometric mean** may be more appropriate than the arithmetic mean. This is discussed in Chapter 13.

4.3 MEASURES OF VARIATION

Range and interquartile range

Two measures of the amount of variation in a data set, the range and the interquartile range, were introduced in Section 3.3. The **range** is the simplest measure, and is the difference between the largest and smallest values. Its disadvantage is that it is based on only two of the observations and gives no idea of how the other observations are arranged between these two. Also, it tends to be larger, the larger the size of the sample. The **interquartile range** indicates the spread of the middle 50% of the distribution, and together with the median is a useful adjunct to the range. It is less sensitive to the size of the sample, providing that this is not too

small; the lower and upper quartiles tend to be more stable than the extreme values that determine the range. These two ranges form the basis of the **box and whiskers plot**, described in Sections 3.3 and 3.4.

Range = highest value − lowest value

Interquartile range = upper quartile − lower quartile

Variance

For most statistical analyses the preferred measure of variation is the **variance** (or the **standard deviation**, which is derived from the variance, see below). This uses all the observations, and is defined in terms of the *deviations* $(x-\bar{x})$ of the observations from the mean, since the variation is small if the observations are bunched closely about their mean, and large if they are scattered over considerable distances. It is not possible simply to average the deviations, as this average will always be zero; the positive deviations corresponding to values above the mean will balance out the negative deviations from values below the mean. An obvious way of overcoming this difficulty would be simply to average the sizes of the deviations, ignoring their sign. However, this measure is not mathematically very tractable, and so instead we average the *squares* of the deviations, since the square of a number is always positive.

$$\text{Variance, } s^2 = \frac{\Sigma(x-\bar{x})^2}{(n-1)}$$

Degrees of freedom

Note that the sum of squared deviations is divided by $(n-1)$ rather than n, because it can be shown mathematically that this gives a better estimate of the variance of the underlying population. The denominator $(n-1)$ is called the number of **degrees of freedom** of the variance. This number is $(n-1)$ rather than n, since only $(n-1)$ of the deviations $(x-\bar{x})$ are independent from each other. The last one can always be calculated from the others because all n of them must add up to zero.

Standard deviation

A disadvantage of the variance is that it is measured in the square of the units used for the observations. For example, if the observations are weights in grams, the

variance is in grams squared. For many purposes it is more convenient to express the variation in the original units by taking the *square root* of the variance. This is called the **standard deviation** (s.d.).

$$\text{s.d., } s = \sqrt{\frac{\Sigma(x-\bar{x})^2}{(n-1)}}$$

or equivalently

$$s = \sqrt{\frac{\Sigma x^2 - (\Sigma x)^2/n}{(n-1)}}$$

When using a calculator, the second formula is more convenient for calculation, since the mean does not have to be calculated first and then subtracted from each of the observations. The equivalence of the two formulae is demonstrated in Example 4.2. (Note: Many calculators have built-in functions for the mean and standard deviation. The keys are commonly labelled $\bar{x}$ and σ_{n-1}, respectively, where σ is the lower case Greek letter sigma.)

Example 4.2

Table 4.1 shows the steps for the calculation of the standard deviation of the eight plasma volume measurements of Example 4.1.

$$\Sigma x^2 - (\Sigma x)^2/n = 72.7980 - (24.02)^2/8 = 0.6780$$

gives the same answer as $\Sigma(x-\bar{x})^2$, and

$$s = \sqrt{(0.6780/7)} = 0.31 \text{ litres}$$

Table 4.1 Calculation of the standard deviation of the plasma volumes (in litres) of eight healthy adult males (same data as in Example 4.1). Mean, $\bar{x} = 3.00$ litres.

	Plasma volume x	Deviation from the mean $x-\bar{x}$	Squared deviation $(x-\bar{x})^2$	Squared observation x^2
	2.75	−0.25	0.0625	7.5625
	2.86	−0.14	0.0196	8.1796
	3.37	0.37	0.1369	11.3569
	2.76	−0.24	0.0576	7.6176
	2.62	−0.38	0.1444	6.8644
	3.49	0.49	0.2401	12.1801
	3.05	0.05	0.0025	9.3025
	3.12	0.12	0.0144	9.7344
Totals	24.02	0.00	0.6780	72.7980

Interpretation of the standard deviation

Usually about 70% of the observations lie within one standard deviation of their mean, and about 95% lie within two standard deviations. These figures are based on a theoretical frequency distribution, called the normal distribution, which is described in Chapter 5. They may be used to derive reference ranges for the distribution of values in the population (see Chapter 5).

Change of units

Adding or subtracting a constant from the observations alters the mean by the same amount but leaves the standard deviation unaffected. Multiplying or dividing by a constant changes both the mean and the standard deviation in the same way.

For example, suppose a set of temperatures is converted from Fahrenheit to centigrade. This is done by subtracting 32, multiplying by 5, and dividing by 9. The new mean may be calculated from the old one in exactly the same way, that is by subtracting 32, multiplying by 5, and dividing by 9. The new standard deviation, however, is simply the old one multiplied by 5 and divided by 9, since the subtraction does not affect it.

Coefficient of variation

$$\text{cv} = \frac{s}{\bar{x}x} \times 100\%$$

The **coefficient of variation** expresses the standard deviation as a percentage of the sample mean. This is useful when interest is in the size of the variation relative to the size of the observation, and it has the advantage that the coefficient of variation is independent of the units of observation. For example, the value of the standard deviation of a set of weights will be different depending on whether they are measured in kilograms or pounds. The coefficient of variation, however, will be the same in both cases as it does not depend on the unit of measurement.

4.4 CALCULATING THE MEAN AND STANDARD DEVIATION FROM A FREQUENCY DISTRIBUTION

Table 4.2 shows the distribution of the number of previous pregnancies of a group of women attending an antenatal clinic. Eighteen of the 100 women had no previous pregnancies, 27 had one, 31 had two, 19 had three, and five had four previous pregnancies. As, for example, adding 2 thirty-one times is

Table 4.2 Distribution of the number of previous pregnancies of a group of women aged 30–34 attending an antenatal clinic.

	No. of previous pregnancies					
	0	1	2	3	4	Total
No. of women	18	27	31	19	5	100

equivalent to adding the product (2×31), the total number of previous pregnancies is calculated by:

$$\begin{aligned}\Sigma x &= (0 \times 18) + (1 \times 27) + (2 \times 31) + (3 \times 19) + (4 \times 5) \\ &= 0 + 27 + 62 + 57 + 20 = 166\end{aligned}$$

The average number of previous pregnancies is, therefore:

$$\bar{x}x = 166/100 = 1.66$$

In the same way:

$$\begin{aligned}\Sigma x^2 &= (0^2 \times 18) + (1^2 \times 27) + (2^2 \times 31) + (3^2 \times 19) + (4^2 \times 5) \\ &= 0 + 27 + 124 + 171 + 80 = 402\end{aligned}$$

The standard deviation is, therefore:

$$s = \sqrt{\frac{(402 - 166^2/100)}{99}} = \sqrt{\frac{126.44}{99}} = 1.13$$

If a variable has been grouped when constructing a frequency distribution, its mean and standard deviation should be calculated using the original values, not the frequency distribution. There are occasions, however, when only the frequency distribution is available. In such a case, approximate values for the mean and standard deviation can be calculated by using the values of the mid-points of the groups and proceeding as above.

4.5 SAMPLING VARIATION AND STANDARD ERROR

As discussed in Chapter 2, the sample is of interest not in its own right, but for what it tells the investigator about the population which it represents. The sample mean, $\bar{x}x$and standard deviation, s, are used to estimate the mean and standard deviation of the population, denoted by the Greek letters μ (mu) and σ (sigma) respectively.

The sample mean is unlikely to be exactly equal to the population mean. A different sample would give a different estimate, the difference being due to

sampling variation. Imagine collecting many independent samples of the same size from the same population, and calculating the sample mean of each of them. A frequency distribution of these means (called the **sampling distribution**) could then be formed. It can be shown that:

1 the mean of this frequency distribution would be the population mean, and
2 the standard deviation would equal $\sigma/\sqrt{n}$. This is called the **standard error of the sample mean**, and it measures how precisely the population mean is estimated by the sample mean. The size of the standard error depends both on how much variation there is in the population and on the size of the sample. The larger the sample size n, the smaller is the standard error.

We seldom know the population standard deviation, σ, however, and so we use the sample standard deviation, s, in its place to estimate the standard error.

$$\text{s.e.} = \frac{s}{\sqrt{n}}$$

Example 4.3

The mean of the eight plasma volumes shown in Table 4.1 is 3.00 litres (Example 4.1) and the standard deviation is 0.31 litres (Example 4.2). The standard error of the mean is therefore estimated as:

$$s/\sqrt{n} = 0.31/\sqrt{8} = 0.11 \text{ litres}$$

Understanding standard deviations and standard errors

Example 4.4

Figure 4.1 shows the results of a game played with a class of 30 students to illustrate the concepts of sampling variation, the sampling distribution, and standard error. Blood pressure measurements for 250 airline pilots were used, and served as the population in the game. The distribution of these measurements is shown in Figure 4.1(a). The population mean, μ, was 78.2 mmHg, and the population standard deviation, σ, was 9.4 mmHg. Each value was written on a small disc and the 250 discs put into a bag.

Each student was asked to shake the bag, select ten discs, write down the ten diastolic blood pressures, work out their mean, $\bar{x}$and return the discs to the bag. In this way 30 different samples were obtained, with 30 different sample means, each estimating the same population mean. The mean of these sample means was 78.23 mmHg, close to the population mean. Their distribution is shown in Figure 4.1(b). The standard deviation of the sample means was 3.01 mmHg, which agreed well with the theoretical value, $\sigma/\sqrt{n} = 9.4/\sqrt{10} = 2.97$ mmHg, for the standard error of the mean of a sample of size ten.

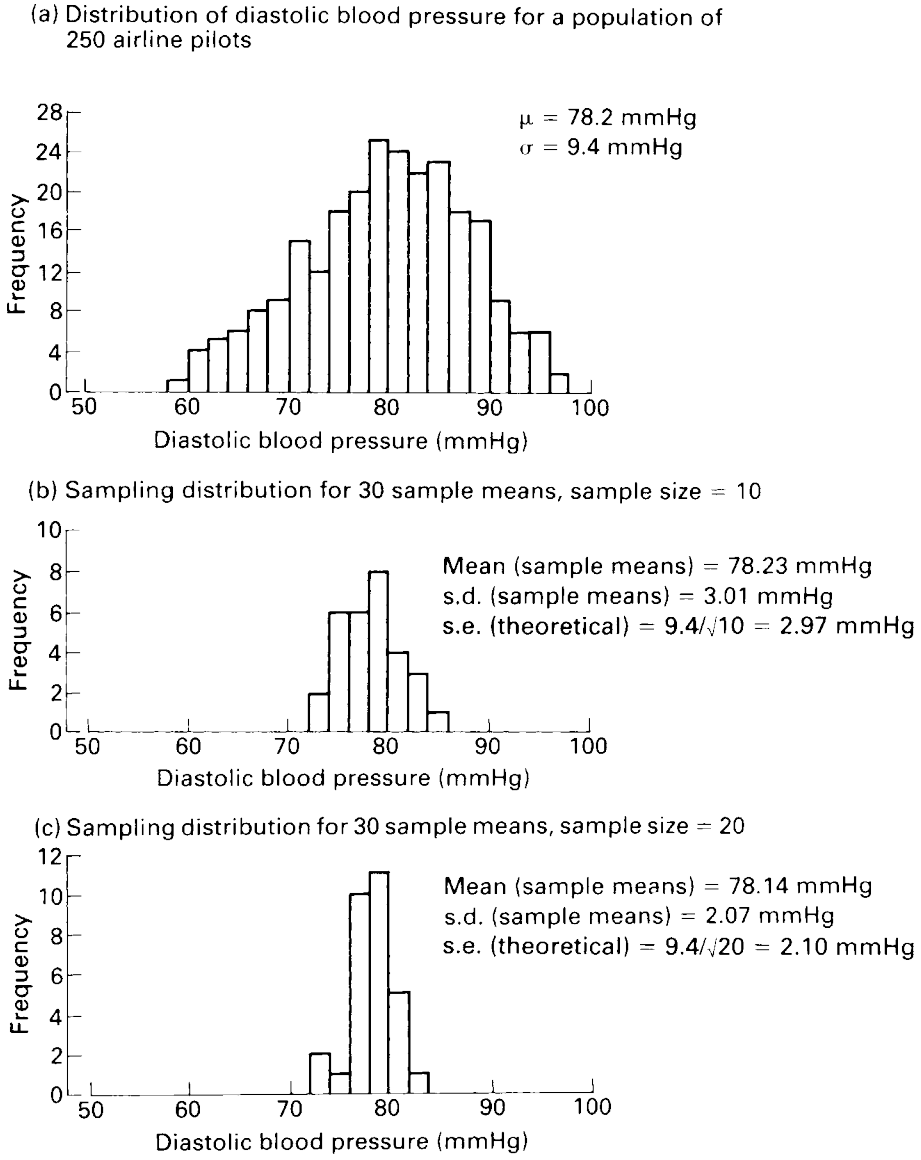


Fig. 4.1 Results of a game played to illustrate the concepts of sampling variation, the sampling distribution, and the standard error.

The exercise was repeated taking samples of size 20. The results are shown in Figure 4.1(c). The reduced variation in the sample means resulting from increasing the sample size from 10 to 20 can be clearly seen. The mean of the sample means was 78.14 mmHg, again close to the population mean. The standard deviation was 2.07 mmHg, again in good agreement with the theoretical value, $9.4/\sqrt{20} = 2.10$ mmHg, for the standard error of the mean of a sample of size 20.

In this game, we had the luxury of results from several different samples, and could draw the sampling distribution. Usually we are not in this position: we have just one sample that we wish to use to estimate the mean of a larger population, which it represents. We can draw the frequency distribution of the values in our sample (see, for example, Figure 3.3 of the histogram of haemoglobin levels of 70 women). Providing the sample size is not too small, this frequency distribution will be similar in appearance to the frequency distribution of the underlying population, with a similar spread of values. In particular, the sample standard deviation will be a fairly accurate estimate of the population standard deviation. As stated in Section 4.2, approximately, 95% of the sample values will lie within two standard deviations of the sample mean. Similarly, approximately 95% of all the values in the population will lie within this same amount of the population mean.

The sample mean will not be exactly equal to the population mean. The theoretical distribution called the **sampling distribution** gives us the spread of values we would get if we took a large number of additional samples; this spread depends on the amount of variation in the underlying population and on our sample size. The standard deviation of the sampling distribution is called the **standard error** and is equal to the standard deviation of the population, divided by the square root of n. This means that approximately 95% of the values in this theoretical sampling distribution of sample means lie within two standard errors of the population mean. This fact can be used to construct a range of likely values for the (unknown) population mean, based on the observed sample mean and its standard error. Such a range is called a **confidence interval**. Its method of construction is not described until Chapter 6 since it depends on using the normal distribution, described in Chapter 5. In summary:

- The standard deviation measures the amount of variability in the population.
- The standard error (= standard deviation $/\sqrt{n}$) measures the amount of variability in the sample mean; it indicates how closely the population mean is likely to be estimated by the sample mean.
- Because standard deviations and standard errors are often confused it is very important that they are clearly labelled when presented in tables of results.