4 Summarizing data

4.1 Types of data

In Chapters 2 and 3 we looked at ways in which data are collected. In this chapter we shall see how data can be summarized to help to reveal information they contain. We do this by calculating numbers from the data which extract the important material. These numbers are called **statistics**. A statistic is anything calculated from the data alone.

It is often useful to distinguish between three types of data: qualitative, discrete quantitative, and continuous quantitative. **Qualitative** data arise when individuals may fall into separate classes. These classes may have no numerical relationship with one another at all, e.g. sex: male, female; types of dwelling: house, maisonette, flat, lodgings; eye colour: brown, grey, blue, green, etc. **Quantitative** data are numerical, arising from counts or measurements. If the measurements can have only certain specific values, like the number of people in a household, or number of teeth which have been filled, those data are said to be **discrete**. Discrete variables usually have integer or whole number values. If the values of the measurements can take any number in a range, such as height or weight, the data are said to be **continuous**. In practice there is overlap between these categories. Most continuous data are limited by the accuracy with which measurements can be made. Human height, for example, is difficult to measure more accurately than to the nearest millimetre and is more usually measured to the nearest centimetre. So only a finite set of possible measurements is actually available, although the quantity 'height' can take an infinite number of possible values, and the measured height is really discrete. However, the methods described below for continuous data will be seen to be those appropriate for its analysis.

We shall refer to qualities or quantities such as sex, height, age, etc., as **variables**, because they vary from one member of a sample or population to another. A qualitative variable is also termed a **categorical variable**, **nominal variable**, or an **attribute**. We shall use these terms interchangeably.

4.2 Frequency distributions

When data are purely qualitative, the simplest way to deal with them is to count the number of cases in each category. For example, in the analysis of the census of a psychiatric hospital population (Section 3.2), one of the variables of interest was the patient's principal diagnosis (Bewley *et al.* 1975). To summarize these data, we count the number of patients having each diagnosis. The results are shown in Table 4.1. The count of individuals having a particular quality is called the **frequency** of that quality. For example, the frequency of schizophrenia is 474. The proportion of individuals having the quality is called the

Table 4.1 Principal diagnosis of patients in Tooting Bec Hospital (data from Bewley *et al.* 1975)

Diagnosis	Number of patients
Schizophrenia	474
Affective disorders	277
Organic brain syndrome	405
Subnormality	58
Alcoholism	57
Other and not known	196
Total	1 467

relative frequency or **proportional frequency**. The relative frequency of schizophrenia is 474/1467 = 0.32 or 32%. The set of frequencies of all the possible categories is called the **frequency distribution** of the variable.

In this census we assessed whether patients were 'unlikely to be discharged', 'possibly to be discharged', or 'likely to be discharged'. The frequencies of these categories are shown in Table 4.2. Likelihood of discharge is a qualitative variable, like diagnosis, but the categories are ordered. This enables us to use another set of summary statistics, the cumulative frequencies. The **cumulative frequency** for a value of a variable is the number of individuals with values less than or equal to that value. Thus, if we order likelihood of discharge from 'unlikely', through 'possibly' to 'likely', the cumulative frequencies are 871, 1 210 (= 871 + 339), and 1 467. The **relative cumulative frequency** for a value is the proportion of individuals in the sample with values less than or equal to that value. For the example they are 0.59 (= 871/1 467), 0.82, and 1.00. Thus we can see that the proportion of patients for whom discharge was not thought likely was 0.82 or 82%.

As we have noted, likelihood of discharge is a qualitative variable, with ordered categories. Sometimes this ordering is taken into account in analysis, sometimes not. Although the categories are ordered these are not quantitative data. There is no sense in which the difference between 'likely' and 'possibly' is the same as the difference between 'possibly' and 'unlikely'.

Table 4.3 shows the frequency distribution of a quantitative variable, parity. This shows the number of previous pregnancies for a sample of women booking for delivery at St George's Hospital. Only certain values are possible, as the number of pregnancies must be an integer, so this variable is discrete. The frequency of each separate value is given.

Table 4.2 Likelihood of discharge of patients in Tooting Bec Hospital (data from Bewley *et al.* 1975)

Discharge:	Frequency	Relative frequency	Cumulative frequency	Relative cumulative frequency
Unlikely	871	0.59	871	0.59
Possible	339	0.23	1 210	0.82
Likely	257	0.18	1 467	1.00
Total	1 467	1.00	1 467	1.00

Table 4.3 Parity of 125 women attending antenatal clinics at St George's Hospital (data supplied by Rebecca McNair, personal communication)

Parity	Frequency	Relative frequency (per cent)	Cumulative frequency	Relative cumulative frequency (per cent)
0	59	47.2	59	47.2
1	44	35.2	103	82.4
2	14	11.2	117	93.6
3	3	2.4	120	96.0
4	4	3.2	124	99.2
5	1	0.8	125	100.0
Total	125	100.0	125	100.0

Table 4.4 FEV1 (litres) of 57 male medical students (data from Physiology practical class, St George's Hospital Medical School)

2.85	3.19	3.50	3.69	3.90	4.14	4.32	4.50	4.80	5.20
2.85	3.20	3.54	3.70	3.96	4.16	4.44	4.56	4.80	5.30
2.98	3.30	3.54	3.70	4.05	4.20	4.47	4.68	4.90	5.43
3.04	3.39	3.57	3.75	4.08	4.20	4.47	4.70	5.00	
3.10	3.42	3.60	3.78	4.10	4.30	4.47	4.71	5.10	
3.10	3.48	3.60	3.83	4.14	4.30	4.50	4.78	5.10	

Table 4.4 shows a continuous variable, forced expiratory volume in one second (FEV1) in a sample of male medical students. As most of the values occur only once, to get a useful frequency distribution we need to divide the FEV1 scale into class intervals, e.g. from 3.0 to 3.5, from 3.5 to 4.0, and so on, and count the number of individuals with FEV1s in each class interval. The class intervals should not overlap, so we must decide which interval contains the boundary point to avoid it being counted twice. It is usual to put the lower boundary of an interval into that interval and the higher boundary into the next interval. Thus the interval starting at 3.0 and ending at 3.5 contains 3.0 but not 3.5. We can write this as '3.0 – ' or '3.0 – 3.5⁻' or '3.0 – 3.499'. Including the lower boundary in the class interval has this advantage: most distributions of measurements have a zero point below which we cannot go, whereas few have an exact upper limit. If we were to include the upper boundary in the interval instead of the lower, we would have two possible ways of dealing with zero. It could be left as an isolated point, not in an interval. Alternatively, it could be included in the lowest interval, which would then not be exactly comparable with the others as it would include both boundaries while all the other intervals only included the upper.

If we take a starting point of 2.5 and an interval of 0.5, we get the frequency distribution shown in Table 4.5. Note that this is not unique. If we take a starting point of 2.4 and an interval of 0.2, we get a different set of frequencies.

The frequency distribution can be calculated easily and accurately using a computer. Manual calculation is not so easy and must be done carefully and systematically.

Table 4.5 Frequency distribution of FEV1 in 57 male medical students (data from Physiology practical class, St George's Hospital Medical School)

FEV1	Frequency	Relative frequency (per cent)
2.0 –	0	0.0
2.5 –	3	5.3
3.0 –	9	15.8
3.5 –	14	24.6
4.0 –	15	26.3
4.5 –	10	17.5
5.0 –	6	10.5
5.5 –	0	0.0
Total	57	100.0

One way recommended by many older texts (e.g. Hill 1977) is to set up a tally system, as in Table 4.6. We go through the data and for each individual make a tally mark by the appropriate interval. We then count up the number in each interval. In practice this is very difficult to do accurately, and it needs to be checked and double-checked. Hill (1977) recommends writing each number on a card and dealing the cards into piles corresponding to the intervals. It is then easy to check that each pile contains only those cases in that interval and count them. This is undoubtedly superior to the tally system. Another method is to order the observations from lowest to highest before marking the interval

Table 4.6 Tally system for finding the frequency distribution of FEV1 (data from Physiology practical class, St George's Hospital Medical School)

FEV1		Frequency
2.0 –		0
2.5 –	///	3
3.0 –	///// ////	9
3.5 –	///// ///// ////	14
4.0 –	///// ///// /////	15
4.5 –	///// /////	10
5.0 –	///// /	6
5.5 –		0
Total		57

boundaries and counting, or to use the stem and leaf plot described below. Personally, I always use a computer.

4.3 Histograms and other frequency graphs

Graphical methods are very useful for examining frequency distributions. Figure 4.1 shows a graph of the cumulative frequency distribution for the FEV1 data. This is called a **step function**, because the frequency increases in abrupt steps. We can smooth this by joining successive points where the cumulative frequency changes by straight lines, to give a **cumulative frequency polygon**. Figure 4.2 shows this for the cumulative relative frequency distribution of FEV1.

The most common way of depicting a frequency distribution is by a **histogram**. This is a diagram where the class intervals are on an axis and rectangles with heights or areas proportional to the frequencies erected on them. Figure 4.3 shows the histogram for the FEV1 distribution in Table 4.5. The vertical scale shows frequency, the number of observations in each interval.

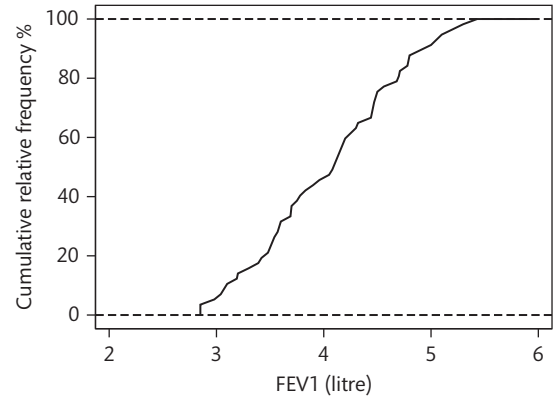

Figure 4.2 Cumulative frequency polygon of FEV1 (data from Physiology practical class, St George's Hospital Medical School).

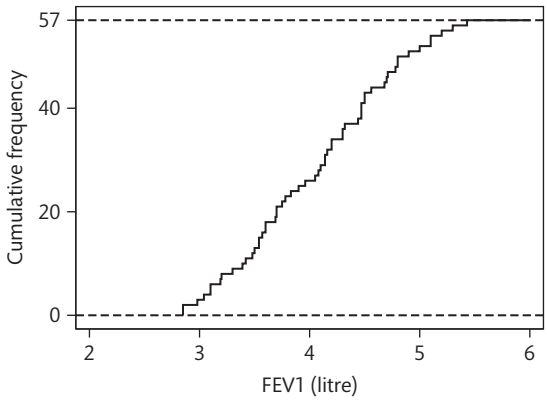

Figure 4.1 Cumulative frequency distribution of FEV1 in a sample of male medical students (data from Physiology practical class, St George's Hospital Medical School).

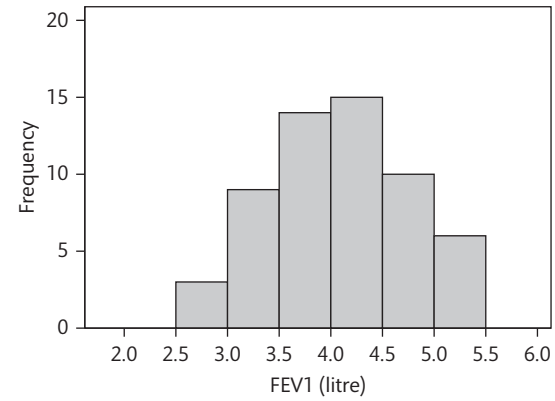

Figure 4.3 Histogram of FEV1: frequency scale (data from Physiology practical class, St George's Hospital Medical School).

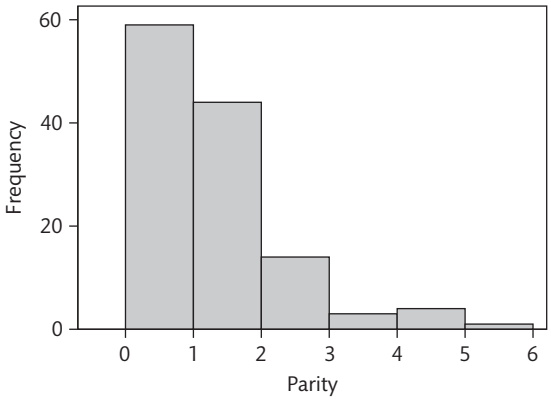

Figure 4.4 Histogram of parity (Table 4.3) using integer cut-off points for the intervals (data supplied by Rebecca McNair, personal communication).

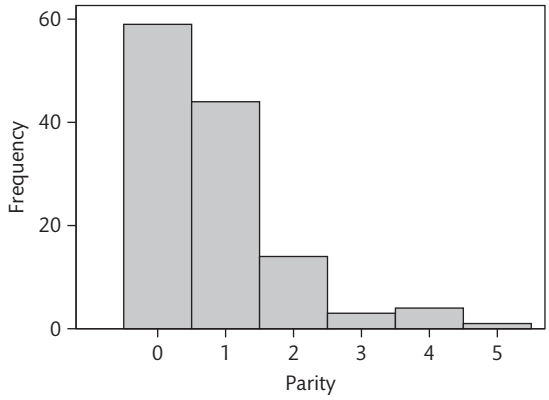

Figure 4.5 Histogram of parity (Table 4.3) using fractional cut-off points for the intervals (data supplied by Rebecca McNair, personal communication).

Sometimes we want to show the distribution of a discrete variable (e.g. Table 4.3) as a histogram. If our intervals are 0 — 1⁻, 1 — 2⁻, etc., the actual observations will all be at one end of the interval (Figure 4.4). Making the starting point of the interval a fraction rather than an integer gives a slightly better picture (Figure 4.5). This can also be helpful for continuous data when there is a lot of digit preference (Section 20.1). For example, where most observations are recorded as integers or as something point five, starting the interval at something .75 can give a more accurate picture. We can also emphasize the discrete nature of the variable by using a narrow interval (Figure 4.6). We could even use simple vertical lines.

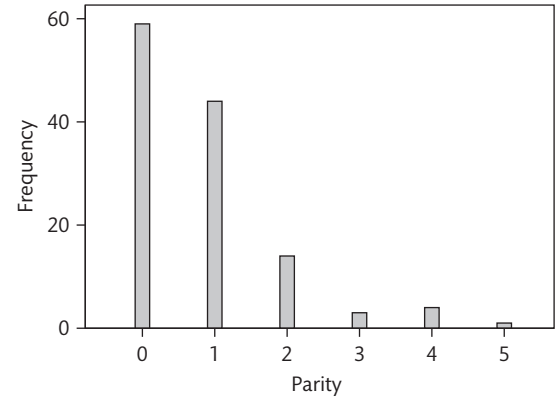

Figure 4.6 Histogram of parity (Table 4.3) using fractional cut-off points and narrow intervals (data supplied by Rebecca McNair, personal communication).

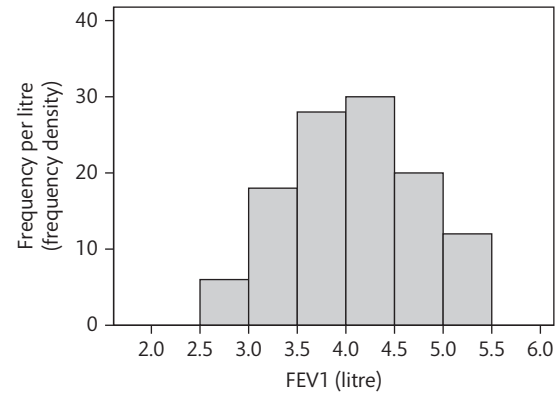

Figure 4.7 Histogram of FEV1: frequency per unit FEV1 or frequency density scale (data from Physiology practical class, St George's Hospital Medical School).

Figure 4.7 shows a histogram for the same distribution as Figure 4.3, with frequency per unit FEV1 (or frequency density) shown on the vertical axis. The distributions appear identical and we may well wonder whether it matters which method we choose. We see that it does matter when we consider a frequency distribution with unequal intervals, as in Table 4.7. If we plot the histogram using the heights of the rectangles to represent relative frequency in the interval we get the histogram in Figure 4.8, whereas if we use the relative frequency per year we get the histogram in Figure 4.9. These histograms tell different stories. Figure 4.8 suggests that the most common age for accident victims is between 15 and 44 years,

Table 4.7 Distribution of age in people suffering accidents in the home (data from Whittington 1977)

Age group	Relative frequency (per cent)	Relative frequency per year (per cent)
0–4	25.3	5.06
5–14	18.9	1.89
15–44	30.3	1.01
45–64	13.6	0.68
65+	11.7	0.33

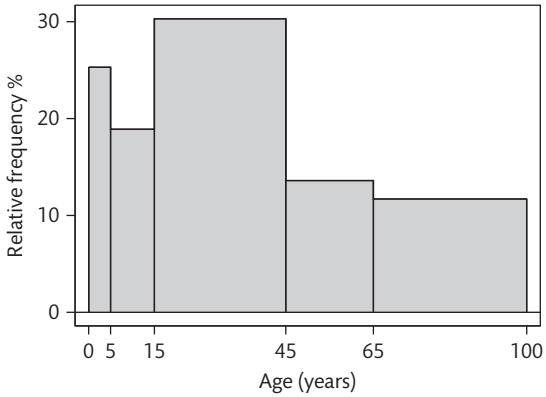

Figure 4.8 Histograms of age distribution of home accident victims using the relative frequency scale (data from Whittington 1977).

whereas Figure 4.9 suggests it is between 0 and 4. Figure 4.9 is correct, Figure 4.8 being distorted by the unequal class intervals. It is therefore preferable in general to use the frequency per unit (frequency density) rather than per class interval when plotting a histogram with unequal class intervals. The frequency for a particular interval is then represented by the area of the rectangle on that interval. Only when the class intervals are all equal can the frequency for the class interval be represented by the height of the rectangle. The computer programmer finds equal intervals much easier, however, and histograms with unequal intervals are now uncommon. I have used equal intervals and the frequency scale in most of this book.

Rather than a histogram consisting of vertical rectangles, we can plot a **frequency polygon** instead. To do

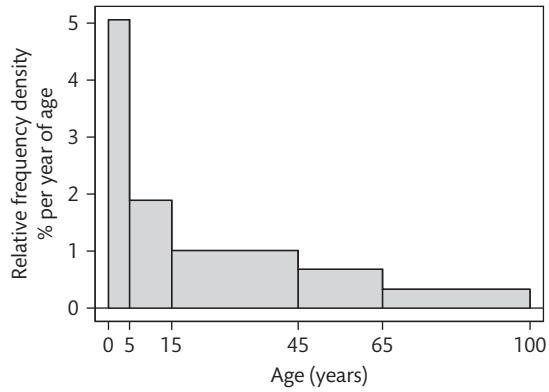

Figure 4.9 Histogram of age distribution of home accident victims using the relative frequency density scale (data from Whittington 1977).

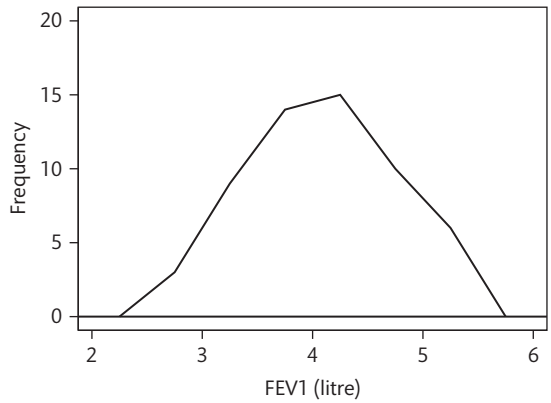

Figure 4.10 Frequency polygon of FEV1 in medical students (data from Physiology practical class, St George's Hospital Medical School).

this we join the centre points of the tops of the rectangles, then omit the rectangles (Figure 4.10). Where a cell of the histogram is empty, we join the line to the centre of the cell at the horizontal axis (Figure 4.11, males). This can be useful if we want to show two or more frequency distributions on the same graph, as in Figure 4.11. When we do this, the comparison is easier if we use relative frequency or relative frequency density rather than frequency. This makes it easier to compare distributions with different numbers of subjects.

A different version of the histogram has been developed by Tukey (1977), the **stem and leaf plot** (Figure 4.12). The rectangles are replaced by the numbers themselves. The 'stem' is the first digit or digits of

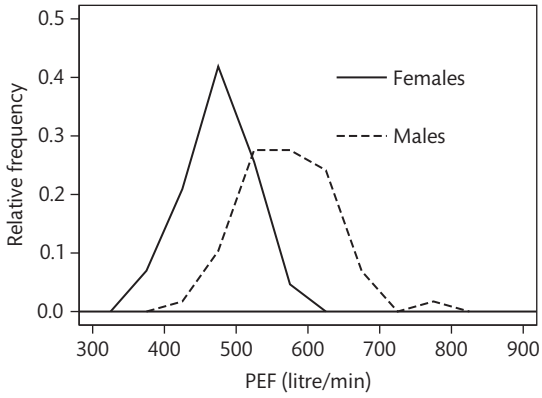

Figure 4.11 Frequency polygons of PEF in medical students (data from Physiology practical class, St George's Hospital Medical School).

2	8 8 9
3	0 1 1 1 2 3 3 4 4 5 5 5 5 6 6 6 7 7 7 7 8 9 9
4	0 0 1 1 1 1 2 2 3 3 3 4 4 4 4 5 5 5 6 7 7 7 8 8 9
5	0 1 1 2 3 4

Figure 4.12 Stem and leaf plot for the FEV1 data, rounded down to one decimal place (data from Physiology practical class, St George's Hospital Medical School).

the number and the 'leaf' the trailing digit. The first row of Figure 4.12 represents the numbers 2.8, 2.8, and 2.9, which in the data are 2.85, 2.85, and 2.98. The plot provides a good summary of data structure while at the same time we can see other characteristics such as a tendency to prefer some trailing digits to others, called digit preference (Section 20.1). It is also easy to construct and much less prone to error than the tally method of finding a frequency distribution.

4.4 Shapes of frequency distribution

Figure 4.3 shows a frequency distribution of a shape often seen in medical data. The distribution is roughly symmetrical about its central value and has frequency concentrated about one central point. The most frequent value is called the **mode** of the distribution and the interval with the greatest frequency is called the **modal interval** or **modal class**. Figure 4.3 has one such point. It is **unimodal**. Figure 4.13 shows a very different shape. Here there are two distinct modes, one near 100 mm Hg and the other near 170 mm Hg. There is pronounced dip in the region between 120 and 160 mm Hg, where we might expect the systolic pressures of many members of the general population to be found. This distribution is **bimodal**. We must be careful to distinguish between the unevenness in the histogram which results from using a small sample to represent a large population and that resulting from genuine bimodality in the data. The trough between 120 and 160 in Figure 4.13 is very marked and might represent a genuine bimodality. In this case we

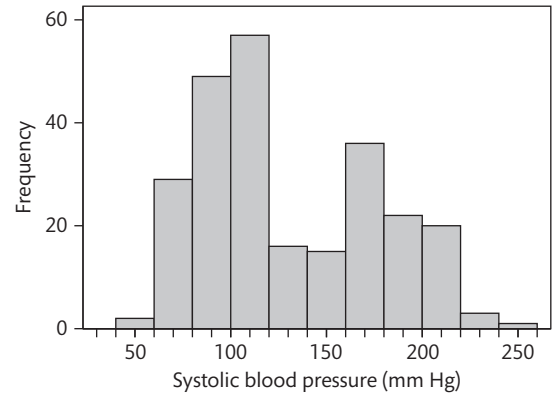

Figure 4.13 Systolic blood pressure in a sample of patients in an intensive therapy unit (data from Friedland *et al.* 1996).

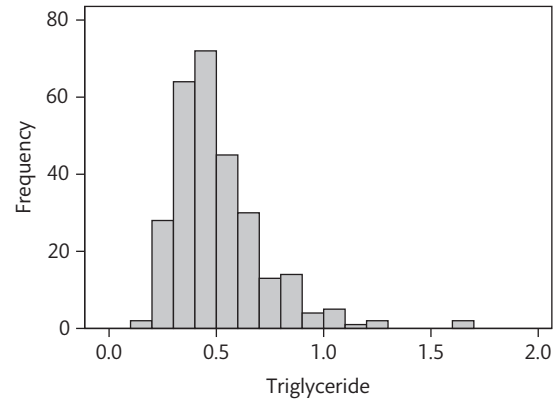

Figure 4.14 Serum triglyceride in cord blood from 282 babies (Table 4.8) (data supplied by Tessi Hanid, personal communication).

Table 4.8 Serum triglyceride measurements in cord blood from 282 babies (data supplied by Tessi Hanid, personal communication)

0.15	0.29	0.32	0.36	0.40	0.42	0.46	0.50	0.56	0.60	0.70	0.86
0.16	0.29	0.33	0.36	0.40	0.42	0.46	0.50	0.56	0.60	0.72	0.87
0.20	0.29	0.33	0.36	0.40	0.42	0.47	0.52	0.56	0.60	0.72	0.88
0.20	0.29	0.33	0.36	0.40	0.44	0.47	0.52	0.56	0.61	0.74	0.88
0.20	0.29	0.33	0.36	0.40	0.44	0.47	0.52	0.56	0.62	0.75	0.95
0.20	0.29	0.33	0.36	0.40	0.44	0.47	0.52	0.56	0.62	0.75	0.96
0.21	0.30	0.33	0.36	0.40	0.44	0.47	0.52	0.56	0.63	0.76	0.96
0.22	0.30	0.33	0.36	0.40	0.44	0.48	0.52	0.56	0.64	0.76	0.99
0.24	0.30	0.33	0.37	0.40	0.44	0.48	0.52	0.56	0.64	0.78	1.01
0.25	0.30	0.34	0.37	0.40	0.44	0.48	0.53	0.57	0.64	0.78	1.02
0.26	0.30	0.34	0.37	0.40	0.44	0.48	0.54	0.57	0.64	0.78	1.02
0.26	0.30	0.34	0.37	0.40	0.44	0.48	0.54	0.58	0.64	0.78	1.04
0.26	0.30	0.34	0.38	0.40	0.45	0.48	0.54	0.58	0.65	0.78	1.08
0.27	0.30	0.34	0.38	0.40	0.45	0.48	0.54	0.58	0.66	0.78	1.11
0.27	0.30	0.34	0.38	0.41	0.45	0.48	0.54	0.58	0.66	0.80	1.20
0.27	0.31	0.34	0.38	0.41	0.45	0.48	0.54	0.59	0.66	0.80	1.28
0.28	0.31	0.34	0.38	0.41	0.45	0.48	0.55	0.59	0.66	0.82	1.64
0.28	0.32	0.35	0.39	0.41	0.45	0.48	0.55	0.59	0.66	0.82	1.66
0.28	0.32	0.35	0.39	0.41	0.46	0.48	0.55	0.59	0.67	0.82	
0.28	0.32	0.35	0.39	0.41	0.46	0.49	0.55	0.60	0.67	0.82	
0.28	0.32	0.35	0.39	0.41	0.46	0.49	0.55	0.60	0.68	0.83	
0.28	0.32	0.35	0.39	0.42	0.46	0.49	0.55	0.60	0.70	0.84	
0.28	0.32	0.35	0.40	0.42	0.46	0.50	0.55	0.60	0.70	0.84	
0.28	0.32	0.36	0.40	0.42	0.46	0.50	0.55	0.60	0.70	0.84	

have people in intensive care, who are very sick. Some have a condition which results in dangerously high pressure, others a condition which results in dangerously low pressure. We actually have multiple populations represented with some overlap between them. However, almost all distributions encountered in medical statistics are unimodal.

Figure 4.14 differs from Figure 4.3 in a different way (Table 4.8). The distribution of serum triglyceride is **skewed**, that is, the distance from the central value to the extreme is much greater on one side than it is on the other. The parts of the histogram near the extremes are called the **tails** of the distribution. If the tails are similar in length the distribution is **symmetrical**, as in

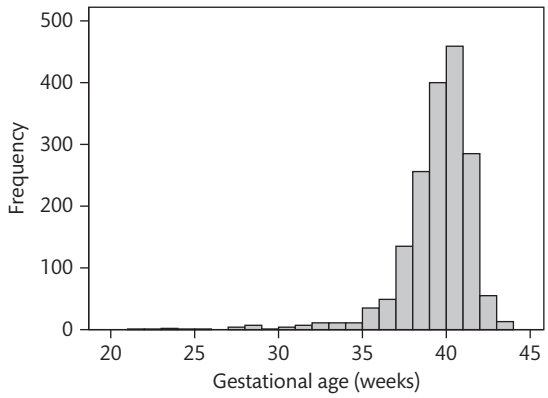

Figure 4.15 Gestational age at birth for 1 749 deliveries at St George's Hospital (data supplied by Rebecca McNair, personal communication).

Figure 4.3. If the tail on the right is longer than the tail on the left as in Figure 4.14, the distribution is **skewed to the right** or **positively skewed**. If the tail on the left is longer, the distribution is **skewed to the left** or **negatively skewed**. This is unusual, but Figure 4.15 shows an example. The negative skewness comes about because babies can be born alive at any gestational age from about 20 weeks, but soon after 40 weeks the baby will have to be born. Pregnancies will not be allowed to go on for more than 44 weeks; the birth would be induced artificially. Most distributions encountered in medical work are symmetrical or skewed to the right, for reasons we shall discuss later (Section 7.4).

4.5 Medians and quantiles

We often want to summarize a frequency distribution in a few numbers, for ease of reporting or comparison. The most direct method is to use quantiles. The **quantiles** are values which divide the distribution such that there is a given proportion of observations below the quantile. For example, the median is a quantile. The **median** is the central value of the distribution, such that half the observations are less than or equal to it and half are greater than or equal to it. We can estimate any quantiles easily from the cumulative frequency distribution or a stem and leaf plot. For the FEV1 data the median is 4.1, the 29th value in Table 4.4. If we have an even number of points, we choose a value midway between the two central values.

In general, we estimate the q quantile, the value such that a proportion q will be below it, as follows. We have n ordered observations which divide the scale into $n + 1$ parts: below the lowest observation, above the highest, and between each adjacent pair. The proportion of the distribution which lies below the ith observation is estimated by $i/(n + 1)$. We set this equal to q and get $i = q(n + 1)$. If i is an integer, the ith observation is the required quantile estimate. If not, let j be the integer part of i, the part before the decimal point. The quantile will lie between the jth and $j + 1$th observations. We estimate it by

$$x_j + (x_{j+1} - x_j) \times (i - j)$$

For the median, for example, the 0.5 quantile, $i = q(n + 1) = 0.5 \times (57 + 1) = 29$, the 29th observation as before.

Other quantiles which are particularly useful are the **quartiles** of the distribution. The quartiles divide the distribution into four equal parts, called **fourths** or **quarters**. The second quartile is the median. For the FEV1 data the first and third quartiles are 3.54 and 4.53. For the first quartile, $i = 0.25 \times 58 = 14.5$. The quartile is between the 14th and 15th observations, which are both 3.54. For the third quartile, $i = 0.75 \times 58 = 43.5$, so the quartile lies between the 43rd and 44th observations, which are 4.50 and 4.56. The quantile is given by $4.50 + (4.56 - 4.50) \times (43.5 - 43) = 4.53$. We often divide the distribution at 99 **centiles** or **percentiles**. The median is thus the 50th centile. For the 20th centile of FEV1, $i = 0.2 \times 58 = 11.6$, so the quantile is between the 11th and 12th observations, 3.42 and 3.48, and can be estimated by $3.42 + (3.48 - 3.42) \times (11.6 - 11) = 3.46$. We can estimate these easily from Figure 4.2 by finding the position of the quantile on the vertical axis, e.g. 0.2 for the 20th centile or 0.5 for the median, drawing a horizontal line to intersect the cumulative frequency polygon, and reading the quantile off the horizontal axis. The term 'quartile' is often used incorrectly to mean the fourth or quarter of the observations which fall between two quartiles. The related words 'quintile' and 'tertile' often suffer in the same way.

Tukey (1977) used the median, quartiles, maximum and minimum as a convenient five figure summary of

a distribution. He also suggested a neat graph, the **box and whisker plot**, which represents this (Figure 4.16). The box shows the distance between the quartiles, with the median marked as a line, and the 'whiskers' show the extremes. The different shapes of the FEV1 and serum triglyceride distributions are clear from the graph. For display purposes, an observation whose distance from the edge of the box (i.e. the quartile) is more than 1.5 times the length of the box (i.e. the interquartile range, Section 4.7) may be called an **outlier**. Outliers may be shown as separate points, as for the serum triglyceride measurements in Figure 4.16. The plot can be useful for showing the comparison of several groups (Figure 4.17).

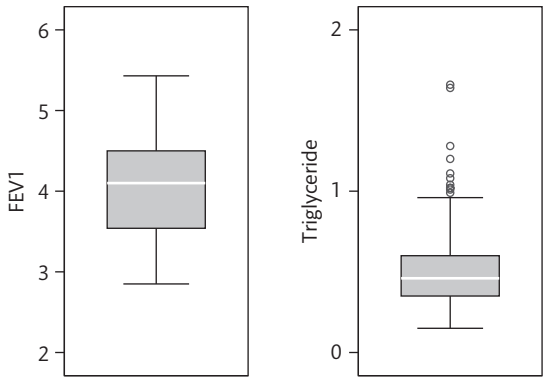

Figure 4.16 Box and whisker plots for FEV1 and for serum triglyceride (data from Physiology practical class, St George's Hospital Medical School/Tessi Hanid).

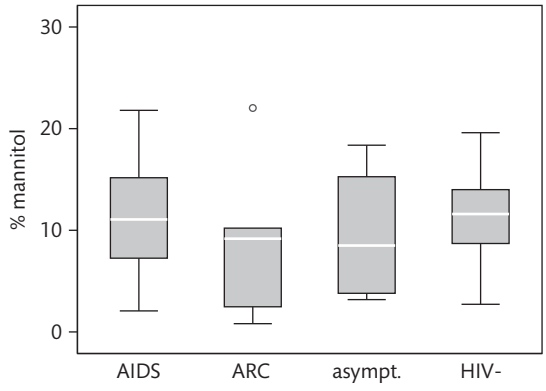

Figure 4.17 Box plots showing a roughly symmetrical variable in four groups, with an outlying point (data in Table 10.7) (data supplied by Moses Kapembwa, personal communication).

4.6 The mean

The median is not the only measure of central value for a distribution. Another is the **arithmetic mean** or **average**, usually referred to simply as the **mean**. This is found by taking the sum of the observations and dividing by their number. For example, consider the following hypothetical data:

2 3 9 5 4 0 6 3 4

The sum is 36 and there are nine observations, so the mean is 36/9 = 4.0. At this point we need to introduce some algebraic notation, widely used in statistics. We denote the observations by

$$x_1, x_2, \ldots, x_i, \ldots, x_n$$

There are n observations and the ith of these is x_i. For the example, $x_4 = 5$ and $n = 9$. The sum of all the x_i is

$$\sum_{i=1}^{n} x_i$$

The summation sign is an upper case Greek letter, sigma, the Greek S. When it is obvious that we are adding the values of x_i for all values of i, which runs from 1 to n, we may abbreviate this to $\sum x_i$ or simply to $\sum x$. The mean of the x_i is denoted by $\bar{x}$ ('x bar'), and

$$\bar{x} = \frac{1}{n}\sum x_i = \frac{\sum x_i}{n}$$

The sum of the 57 FEV1s is 231.51 and hence the mean is 231.51/57 = 4.06. This is very close to the median, 4.1, so the median is within 1% of the mean. This is not so for the triglyceride data. The median triglyceride (Table 4.8) is 0.46 but the mean is 0.51, which is higher. The median is 10% away from the mean. If the distribution is symmetrical, the sample mean and median will be about the same, but in a skewed distribution they will usually not. If the distribution is skewed to the right, as for serum triglyceride, the mean will usually be greater, if it is skewed to the left the median will usually be greater. This is because the values in the tails affect the mean but not the median.

The sample mean has much nicer mathematical properties than the median and is thus more useful for the comparison methods described later. The median is a very useful descriptive statistic, but not much used for other purposes.

4.7 Variance, range, and interquartile range

The mean and median are measures of the position of the middle of the distribution, which we call the **central tendency**. We also need a measure of the spread or variability of the distribution, called the **dispersion**.

One obvious measure is the **range**, the difference between the highest and lowest values. For the data of Table 4.4, the range is 5.43 – 2.85 = 2.58 litres. The range is often presented as the two extremes, 2.85–5.43 litres, rather than their difference. The range is a useful descriptive measure, but has two disadvantages. Firstly, it depends only on the extreme values and so can vary a lot from sample to sample. Secondly, it depends on the sample size. The larger the sample is, the further apart the extremes are likely to be. We can see this if we consider a sample of size 2. If we add a third member to the sample, the range will only remain the same if the new observation falls between the other two, otherwise the range will increase. We can get round the second of these problems by using the **interquartile range**, the difference between the first and third quartiles. For the data of Table 4.4, the interquartile range is 4.53 – 3.54 = 0.99 litres. The interquartile range, too, is often presented as the two extremes, 3.54–4.53 litres. However, the interquartile range is quite variable from sample to sample and is also mathematically intractable. Although a useful descriptive measure, it is not the one preferred for purposes of comparison.

The most frequently used measures of dispersion are the variance and standard deviation. We start by calculating the difference between each observation and the sample mean, called the **deviations from the mean** (Table 4.9). If the data are widely scattered, many of the observations x_i will be far from the mean $\bar{x}$ and so many deviations $x_i - \bar{x}$ will be large. If the data are narrowly scattered, very few observations will be far from the mean and so few deviations $x_i - \bar{x}$ will be large. We need some kind of average deviation to measure the scatter. If we add all the deviations together, we get zero, because $\sum(x_i - \bar{x}) = \sum x_i - \sum \bar{x} = \sum x_i - n\bar{x}$ and $n\bar{x} = \sum x_i$. Instead we square the deviations and then add them, as shown in Table 4.9. This removes the effect of sign; we are only measuring the size of the deviation, not the direction. This gives us $\sum(x_i - \bar{x})^2$, in the example equal to 52, called the **sum of squares about the mean**, usually abbreviated to **sum of squares**.

Table 4.9 Deviations from the mean of nine observations

Observations x_i	Deviations from the mean $x_i - \bar{x}$	Squared deviations $(x_i - \bar{x})^2$
2	–2	4
3	–1	1
9	5	25
5	1	1
4	0	0
0	–4	16
6	2	4
3	–1	1
4	0	0
36	0	52

Clearly, the sum of squares will depend on the number of observations as well as the scatter. We want to find some kind of average squared deviation. This leads to a difficulty. Although we want an average squared deviation, we divide the sum of squares by $n-1$, not n. This is not the obvious thing to do and puzzles many students of statistical methods. The reason is that we are interested in estimating the scatter of the population, rather than the sample, and the sum of squares about the sample mean is proportional to $n - 1$ (Appendix 4A, Appendix 6B). Dividing by n would lead to small samples producing lower estimates of variability than large samples. The minimum number of observations from which the variability can be estimated is two, a single observation cannot tell us how variable the data are. If we used n as our divisor, for $n = 1$ the sum of squares would be zero, giving a variance of zero. With the correct divisor of $n - 1$, $n = 1$ gives the meaningless ratio 0/0, reflecting the impossibility of estimating variability from a single observation. The estimate

of variability is called the **variance**, defined by

$$\text{variance} = \frac{1}{n-1}\sum(x_i - \bar{x})^2$$

We have already said that $\sum(x_i - \bar{x})^2$ is called the sum of squares. The quantity $n - 1$ is called the **degrees of freedom** of the variance estimate (Appendix 7A). We have:

$$\text{variance} = \frac{\text{sum of squares}}{\text{degrees of freedom}}$$

We shall usually denote the variance by s^2. In the example, the sum of squares is 52 and there are nine observations, giving 8 degrees of freedom. Hence $s^2 = 52/8 = 6.5$.

The formula $\sum(x_i - \bar{x})^2$ gives us a rather tedious calculation. There is another formula for the sum of squares, which makes the calculation easier to carry out. This is simply an algebraic manipulation of the first form and gives exactly the same answers. We thus have two formulae for variance:

$$s^2 = \frac{1}{n-1}\sum\left(x_i - \bar{x}^2\right)$$

$$s^2 = \frac{1}{n-1}\left(\sum x_i^2 - \frac{(\sum x_i)^2}{n}\right)$$

The algebra is quite simple and is given in Appendix 4B. For example, using the second formula for the nine observations, we have:

$$\begin{aligned}\sum x_i^2 &= 2^2 + 3^2 + 9^2 + 5^2 + 4^2 + 0^2 + 6^2 + 3^2 + 4^2\\ &= 4 + 9 + 81 + 25 + 16 + 0 + 36 + 9 + 16\\ &= 196\\ \sum x_i &= 36\\ s^2 &= \frac{1}{n-1}\left(\sum x_i^2 - \frac{(\sum x_i)^2}{n}\right)\\ &= \frac{1}{9-1}\left(196 - \frac{36^2}{9}\right)\\ &= \frac{1}{8}(196 - 144)\\ &= 52/8\\ &= 6.5\end{aligned}$$

as before. On a calculator this is a much easier formula than the first, as the numbers need only be put in once. It can be inaccurate, because we may subtract one large number from another to get a small one. For this reason the first formula would be used in a computer program.

4.8 Standard deviation

The variance is calculated from the squares of the observations. This means that it is not in the same units as the observations, which limits its use as a descriptive statistic. The obvious answer to this is to take the square root, which will then have the same units as the observations and the mean. The square root of the variance is called the **standard deviation**, usually denoted by s. Thus,

$$\begin{aligned}s &= \sqrt{\frac{1}{n-1}\sum(x_i - \bar{x})^2}\\ &= \sqrt{\frac{1}{n-1}\left(\sum x_i^2 - \frac{(\sum x_i)^2}{n}\right)}\end{aligned}$$

Returning to the FEV data, we calculate the variance and standard deviation as follows. We have $n = 57$, $\sum x_i = 231.51$, $\sum x_i^2 = 965.45$.

$$\begin{aligned}\text{Sum of squares} &= \sum x_i^2 - \frac{(\sum x_i)^2}{n}\\ &= 965.45 - \frac{231.51^2}{57}\\ &= 965.45 - 940.296\\ &= 25.154\\ s^2 &= \frac{\text{sum of squares}}{n-1}\\ &= \frac{25.154}{57-1}\\ &= 0.449\end{aligned}$$

The standard deviation is $s = \sqrt{s^2} = \sqrt{0.449} = 0.67$ litres.

Figures 4.18 and 4.19 show the relationship between mean, standard deviation, and frequency distribution. For FEV1, we see that the majority of observations are within one standard deviation of the mean, and nearly all within two standard deviations of the mean (Figure 4.18). There is a small part of the histogram outside the $\bar{x} - 2s$ to $\bar{x} + 2s$ interval, on either side of this symmetrical histogram. Figure 4.19 shows the same thing for the highly skewed triglyceride data. In this case, however, the outlying observations are all in one tail of the distribution. In general, we expect roughly two-thirds of observations to

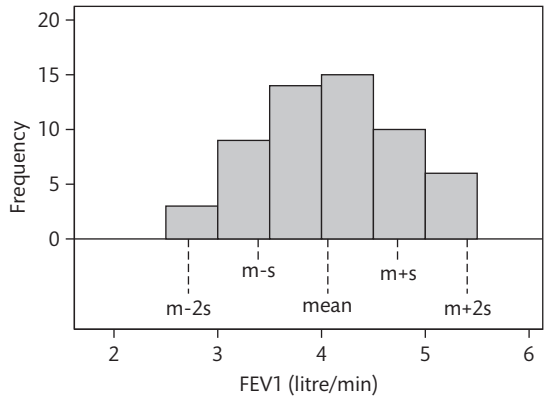

Figure 4.18 Histogram of FEV1 with mean and standard deviation (data from Physiology practical class, St George's Hospital Medical School).

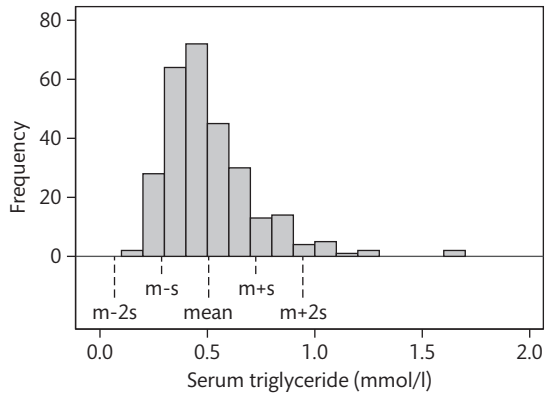

Figure 4.19 Histogram of triglyceride with mean and standard deviation (data supplied by Tessi Hanid, personal communication).

lie within one standard deviation of the mean and 95% to lie within two standard deviations of the mean, but where the outlying observations are will depend on symmetry or skewness.

4.9 Multiple choice questions: Summarizing data

(Each branch is either true or false.)

4.1 Which of the following are qualitative variables:
(a) sex;
(b) parity;
(c) diastolic blood pressure;
(d) diagnosis;
(e) height.

4.2 Which of the following are continuous variables:
(a) blood glucose;
(b) peak expiratory flow rate;
(c) age last birthday;
(d) exact age;
(e) family size.

4.3 When a distribution is skewed to the right:
(a) the median is greater than the mean;
(b) the distribution is unimodal;
(c) the tail on the left is shorter than the tail on the right;
(d) the standard deviation is less than the variance;
(e) the majority of observations are less than the mean.

4.4 The shape of a frequency distribution can be described using:
(a) a box and whisker plot;
(b) a histogram;
(c) a stem and leaf plot;
(d) mean and variance;
(e) a table of frequencies.

4.5 For the sample 3, 1, 7, 2, 2:
(a) the mean is 3;
(b) the median is 7;
(c) the mode is 2;
(d) the range is 1;
(e) the standard deviation is 6.0.

4.6 Diastolic blood pressure has a distribution which is slightly skewed to the right. If the mean and standard deviation were calculated for the diastolic pressures of a random sample of men:
(a) there would be fewer observations below the mean than above it;
(b) the standard deviation would be approximately equal to the mean;
(c) the majority of observations would be more than one standard deviation from the mean;
(d) the standard deviation would estimate the accuracy of blood pressure measurement;
(e) about 95% of observations would be expected to be within two standard deviations of the mean.

4.10 Exercise: Student measurements and a graph of study numbers

There are two sets of data in this exercise. The first, observations made during a student anatomy practical, provide practice in reading histograms and standard deviations. The second, from a publication on the numbers of participants in studies, presents a challenging graphical interpretation.

Figure 4.20 shows the distribution of the mid upper arm circumferences of 120 female biomedical sciences, medical, nursing, physiotherapy, and radiography students.

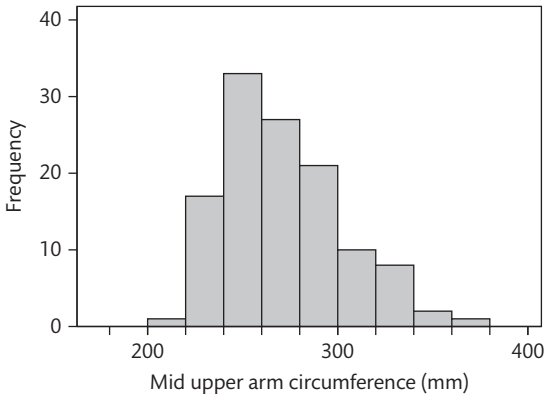

Figure 4.20 Distribution of the mid upper arm circumferences of 120 female students (data from Anatomy practical class, St George's Hospital Medical School).

4.1 What kind of variable is arm circumference?
4.2 What kind of graph is Figure 4.20?
4.3 On the graph, where are the mode, the lower tail, and the upper tail of this distribution?
4.4 From the graph, how would you describe the shape of the distribution of arm circumference and why?
4.5 From the graph, approximately what would you estimate the median and the first and third quartiles to be? Where would they appear along the horizontal axis?
4.6 From the graph, approximately what would you estimate the mean and the standard deviation to be? Where would they appear along the horizontal axis?

Table 4.10 shows eye colour, as recorded by another student, for male and female students.

Table 4.10 Recorded eye colour by sex for 183 students

Eye colour	Sex		
	Female	**Male**	**Total**
Black	6	4	10
Brown	47	32	79
Blue	27	16	43
Grey	10	1	11
Hazel	9	5	14
Green	16	4	20
Other	4	1	5
Missing	1	0	1
Total	120	63	183

4.7 What kind of variable is eye colour? What kind of variable is sex?

My friend and colleague Doug Altman sent me this interesting graph. Shiraishi *et al.* (2009) analysed studies of diagnostic tests reported in the journal *Radiology* between 1997 and 2006. They looked at all the studies of diagnostic methods which used a numerical variable with a particular cut-off value to decide the diagnosis. Such studies are analysed and presented using a ROC curve, or Receiver Operating Characteristic curve (Section 20.6). Shiraishi *et al.* looked at the total number of people, both with and without the condition under investigation, who were included in each of these studies. They gave a graph similar to Figure 4.21 for the distribution of the number of participants included.

4.8 How would you describe the shape of this distribution?
4.9 What feature of the intervals makes it difficult to draw a histogram for these data?
4.10 If we were to present this distribution as a histogram, what would the horizontal scale show?
4.11 If we were to present this distribution as a histogram, what would the vertical scale show?
4.12 Think about how this might appear as a valid histogram. How would you describe the shape of this distribution?

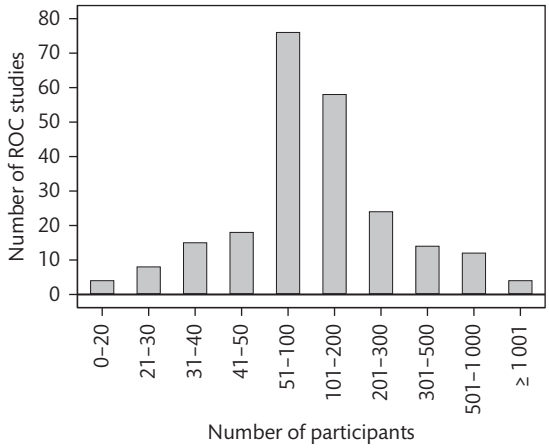

Figure 4.21 Numbers of participants for each study in a sample of 233 diagnostic studies (reproduced from Shiraishi J *et al*. Experimental design and data analysis in receiver operating characteristic studies: lessons learned from reports in *Radiology* from 1997 to 2006. *Radiology* 2009 253:(3) 822–830, with permission from the Radiological Society of North America).

Appendix 4A: The divisor for the variance

The variance is found by dividing the sum of squares about the sample mean by $n - 1$, not by n. This is because we want the scatter about the population mean, and the scatter about the sample mean is always less. The sample mean is 'closer' to the data points than is the population mean. We shall try a little sampling experiment to show this. Table 4.11 shows a set of 100 random digits which we shall take as the population to be sampled. They have mean 4.74 and the sum of squares about the mean is 811.24. Hence the average squared difference from the mean is 8.1124. We can take samples of size two at random from this population using a pair of decimal dice, which will enable us to choose any digit numbered from 00 to 99. The first pair chosen was 5 and 6 which has mean 5.5. The sum of squares about the population mean 4.74 is $(5 - 4.74)^2 + (6 - 4.74)^2 = 1.655$. The sum of squares about the sample mean is $(5 - 5.5)^2 + (6 - 5.5)^2 = 0.5$.

The sum of squares about the population mean is greater than the sum of squares about the sample mean, and this will always be so. Table 4.12 shows this for 20 such samples of size two. The average sum of squares about the population mean is 13.6, and about the sample mean it is 5.7. Hence dividing by the sample size ($n = 2$), we have mean square differences of 6.8 about the population mean and 2.9 about the sample mean. Compare this with 8.1 for the population as a whole. We see that the sum of squares about the population mean is quite close to 8.1, while the sum of squares about the sample mean is much less. However, if we divide the sum of squares about the sample mean by $n - 1$, i.e. 1, instead of n we have 5.7, which is not much different from the 6.8 from the sum of squares about the population mean.

Table 4.13 shows the results of a similar experiment with more samples being taken. The table shows the two average variance estimates using n and $n - 1$ as the divisor of the sum of squares, for sample sizes 2, 3, 4, 5, and 10. We see that the sum of squares about the sample mean divided by n increases steadily with sample size, but if we divide it by $n - 1$ instead of n, the estimate does not change as the sample size increases. The sum of squares about the sample mean is proportional to $n - 1$.

Table 4.11 Population of 100 random digits for a sampling experiment

9	1	0	7	5	6	9	5	8	8	1	0	5	7	6	5	0	2	2	2
1	8	8	8	5	2	4	8	3	1	6	5	5	7	4	1	7	3	3	3
2	8	1	8	5	8	4	0	1	9	2	1	6	9	4	4	7	6	1	7
1	9	7	9	7	2	7	7	0	8	1	6	3	8	0	5	7	4	8	6
7	0	2	8	8	7	2	5	4	1	8	6	8	3	5	8	2	7	2	4

Table 4.12 Sampling pairs from Table 4.11

Sample		$\sum(x_i - \mu)^2$	$\sum(x_i - \bar{x})^2$	Sample		$\sum(x_i - \mu)^2$	$\sum(x_i - \bar{x})^2$
5	6	1.655	0.5	8	3	13.655	12.5
8	8	21.255	0.0	5	7	5.175	2.0
6	1	15.575	12.5	5	2	5.575	4.5
9	3	21.175	18.0	5	7	5.175	2.0
5	5	0.135	0.0	8	8	21.255	0.0
7	7	10.215	0.0	3	2	10.535	0.5
1	7	19.095	18.0	0	4	23.015	8.0
9	8	28.775	0.5	9	3	21.175	18.0
3	3	6.055	0.0	5	2	7.575	4.5
5	1	14.055	8.0	6	9	19.735	4.5
Mean						13.6432	5.7

Table 4.13 Mean sums of squares about the sample mean for sets of 100 random samples from Table 4.12

Number in sample, n	Mean variance estimates	
	$\frac{1}{n}\sum(x_i - \bar{x})^2$	$\frac{1}{n-1}\sum(x_i - \bar{x})^2$
2	4.5	9.1
3	5.4	8.1
4	5.9	7.9
5	6.2	7.7
10	7.2	8.0

Appendix 4B: Formulae for the sum of squares

The different formulae for sums of squares are derived as follows:

$$\begin{aligned}\text{sum of squares} &= \sum(x_i - \bar{x})^2 \\ &= \sum(x_i^2 - 2x_i\bar{x} + \bar{x}^2) \\ &= \sum x_i^2 - \sum 2x_i\bar{x} + \sum \bar{x}^2 \\ &= \sum x_i^2 - 2\bar{x}\sum x_i + n\bar{x}^2\end{aligned}$$

because $\bar{x}$ has the same value for each of the n observations. Now, $\sum x_i = n\bar{x}$, so

$$\begin{aligned}\text{sum of squares} &= \sum x_i^2 - 2\bar{x}n\bar{x} + n\bar{x}^2 \\ &= \sum x_i^2 - 2n\bar{x}^2 + n\bar{x}^2 \\ &= \sum x_i^2 - n\bar{x}^2\end{aligned}$$

and putting $\bar{x} = \frac{1}{n}\sum x_i$

$$\begin{aligned}\text{sum of squares} &= \sum x_i^2 - n\left(\frac{1}{n}\sum x_i\right)^2 \\ &= \sum x_i^2 - \frac{(\sum x_i)^2}{n}\end{aligned}$$

We thus have three formulae for variance:

$$\begin{aligned}s^2 &= \frac{1}{n-1}\sum(x_i - x)^2 \\ &= \frac{1}{n-1}\left(\sum x_i^2 - n\bar{x}^2\right) \\ &= \frac{1}{n-1}\left(\sum x_i^2 - \frac{(\sum x_i)^2}{n}\right)\end{aligned}$$