CHAPTER 2

# Descriptive statistics

This chapter presents an introduction to the key elements of **descriptive statistics**: the ways in which quantitative information can be presented and described. The first step in the analysis of quantitative data is its organization and presentation in tables and graphs. The basic features of the data – such as its central or most common values and the way the observations are distributed around these central values – can then be summarized in various ways.

## 2.1 Presentation of numerical data

### 2.1.1 Frequency distributions

Quantitative data in their raw form consist of a series of numbers or categories. For example, the Poor Law data on *per capita* relief expenditure in 1831 by the 24 parishes in Kent could be set out in its original sequence as in table 2.1, with the data rounded to one decimal place.

Even with 24 observations it is difficult to get much sense of the data in this form. It would help a little if the values were arranged in ascending or descending order of magnitude (known as an **array**). This would immediately show the highest and lowest values and give some indication of the most common level of payment, but with a large number of observations it would still be hard to interpret the information in this form.

A much clearer picture can be obtained if the data are re-arranged as a **frequency distribution**. This is the most common form in which data are summarized and presented. With a discrete variable (see §1.3.3) it makes sense to measure the frequency of occurrence of *each* of the values. For example, if a lecturer had set an examination with nine questions, she might find it informative to compile a frequency distribution showing how many candidates had answered each question.

However, this is less appropriate for a continuous variable such as the

**Table 2.1** *Per capita* expenditure on relief in 24 Kent parishes in 1831 (shillings)

| | | | | | |
|---|---|---|---|---|---|
| 20.4 | 29.1 | 14.9 | 24.1 | 18.2 | 20.7 |
| 8.1 | 14.0 | 18.4 | 34.5 | 16.1 | 24.6 |
| 25.4 | 12.6 | 13.3 | 27.3 | 29.6 | 13.6 |
| 11.4 | 21.5 | 20.9 | 11.6 | 18.2 | 37.9 |

payments of relief, since we would not normally be interested in minute differences, e.g. in finding out how many parishes paid 18.1 shillings, and how many paid 18.2 shillings, etc. Instead we choose appropriate **class intervals** and note the frequency of the occurrence of payments in each class.

For table 2.2 we have chosen seven class intervals, each with a width of 5 shillings; so that all the payments in the first interval are equal to or greater than 5 shillings (abbreviated to ≥5) but less than 10 shillings (< 10); those in the next are equal to or greater than 10 shillings but less than 15 shillings, and so on.[1] The **frequency** is shown in column (3) and the **relative frequency** in column (4). In order to show how the frequency is derived the underlying tally is shown in column (2), but this would not normally be included in a frequency table. The relative frequency shows the share of the

**Table 2.2** Frequency, relative frequency, and cumulative frequency of *per capita* relief payments in Kent in 1831

| (1) **Class intervals (shillings)** | (2) **Tally** | (3) **Frequency (*f*)** | (4) **Relative frequency (%) (*f*/n × 100)** | (5) **Cumulative frequency** | (6) **Cumulative relative frequency (%)** |
|---|---|---|---|---|---|
| ≥5 but <10 | 1 | 1 | 4.2 | 1 | 4.2 |
| ≥10 but <15 | ~~1 1 1 1~~ 1 1 | 7 | 29.2 | 8 | 33.4 |
| ≥15 but <20 | 1 1 1 1 | 4 | 16.6 | 12 | 50.0 |
| ≥20 but <25 | ~~1 1 1 1~~ 1 | 6 | 25.0 | 18 | 75.0 |
| ≥25 but <30 | 1 1 1 1 | 4 | 16.6 | 22 | 91.6 |
| ≥30 but <35 | 1 | 1 | 4.2 | 23 | 95.8 |
| ≥35 but <40 | 1 | 1 | 4.2 | 24 | 100.0 |
| | | 24 | 100.0 | | |

*Note:*
In column (1) ≥ stands for 'equal to or greater than', and < for 'less than'.

observations in each class interval. In table 2.2 this is expressed as a percentage of the total (adding to 100), but it could equally well be a proportion (adding to 1).

It is sometimes also useful to know the **cumulative frequency** of a variable. This gives the number of observations which are less than or equal to any selected value or class interval. The **cumulative relative frequency** gives the corresponding proportions or percentages. For example, we can see from column (5) of table 2.2 that *per capita* relief payments were less than 20 shillings in 12 Kent parishes, and from column (6) that this represented 50 per cent of the parishes.

By summarizing the data in the form of a frequency distribution some information is lost, but it is now much easier to get a clear idea of the pattern of relief payments in the Kent parishes.

### *Number and width of the class intervals*

The precise number and width of the class (or cell) intervals in a frequency table is a matter of judgement. Too many narrow intervals and the pattern is again obscured; too few and too much information is lost. For example, in relation to table 2.2, covering a sample of 24 parishes, having (say) 2 classes would tell us very little, whereas 20 classes would be little improvement on the presentation of all the data in table 2.1. A rough rule of thumb is that it is usually best to have between five and 20 classes, depending on the number of observations.

It is preferable if the width can be made uniform throughout. Classes of unequal width (e.g. 10 – 14.9, 15 – 24.9, 25 – 29.9) are confusing to the reader and a nuisance in computations. However, this is not always possible. For example, when there are a few extreme values (as in data on wealth holdings) the intervals may have to be changed to accommodate the very uneven distribution.

It is sometimes necessary to work with open-ended classes, e.g. 'under 10 shillings' or '50 shillings and over'. However, if the data were given in this form in the original source this can cause problems in calculations that require a figure for the mid-point of the class intervals (as in §2.4). At the lower end it may not be plausible to assume that the values are (roughly) evenly distributed all the way down to a lower bound of zero. At the upper end, in the absence of information about the values in excess of the lower limit given for the open-ended class, the best that can be done is to make an informed guess about the likely values within this class.

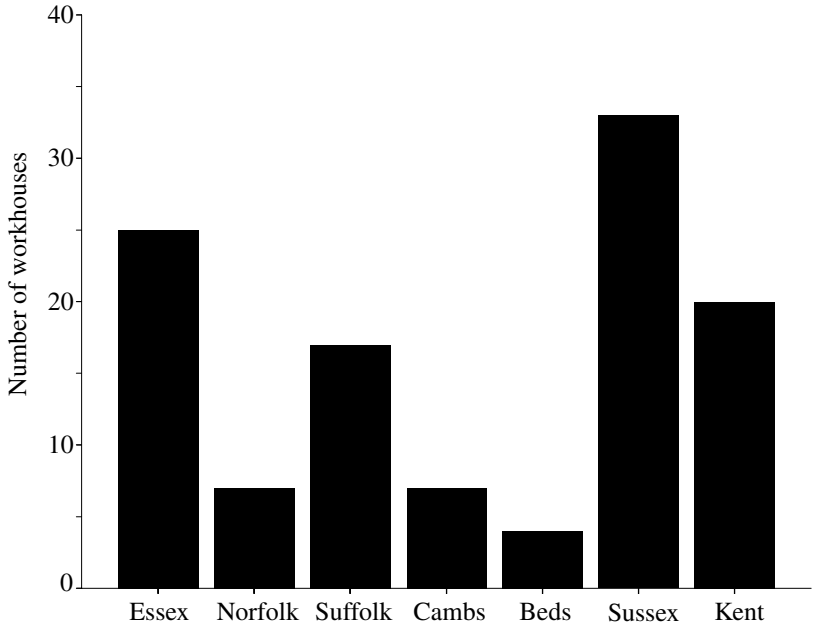



**Figure 2.1** Bar chart of number of workhouses in selected counties in 1831

### 2.1.2 Bar charts and histograms

It is often useful for both statistician and reader to present results in a form that can be interpreted visually. Two common methods of doing this with frequency distributions are bar charts and histograms.

**Bar charts** are commonly used to present data on the frequency distribution of qualitative data, or data which falls into discrete categories. Thus the number of workhouses in the sample parishes in seven counties in the south and south-east of England could be represented in a bar chart such as figure 2.1, in which there is a bar for each of the selected counties, and the *height* of each bar indicates the number (or frequency) of workhouses reported in the Poor Law data set. There is no statistical convention relating to the width of each bar, though it is visually more appealing to present bars of common width for each category equally spaced.

**Histograms** are used where the frequencies to be charted are grouped in class intervals. Each bar (or rectangle) represents either the absolute number or the proportion of cases with values in the class interval. To illustrate the construction of histograms we will use the Poor Law data on *per capita* relief payments by all 311 parishes. The information is set out in table 2.3 using eight class intervals, each with a width of 6 shillings. The number of parishes in each class is given in column (2), the corresponding

**Table 2.3** Frequency of *per capita* relief payments in 311 parishes in 1831

| (1) Class intervals (shillings) | (2) Frequency | (3) Relative frequency | (4) Cumulative relative frequency |
|---|---|---|---|
| ≥0 but <6 | 6 | 1.93 | 1.93 |
| ≥6 but <12 | 78 | 25.08 | 27.01 |
| ≥12 but <18 | 93 | 29.90 | 56.91 |
| ≥18 but <24 | 63 | 20.26 | 77.17 |
| ≥24 but <30 | 46 | 14.79 | 91.96 |
| ≥30 but <36 | 17 | 5.47 | 97.43 |
| ≥36 but <42 | 5 | 1.61 | 99.04 |
| ≥42 but <48 | 3 | 0.96 | 100.00 |
| | 311 | 100.00 | |

*Note:*
In column (1) ≥ stands for 'equal to or greater than', and < for 'less than'.

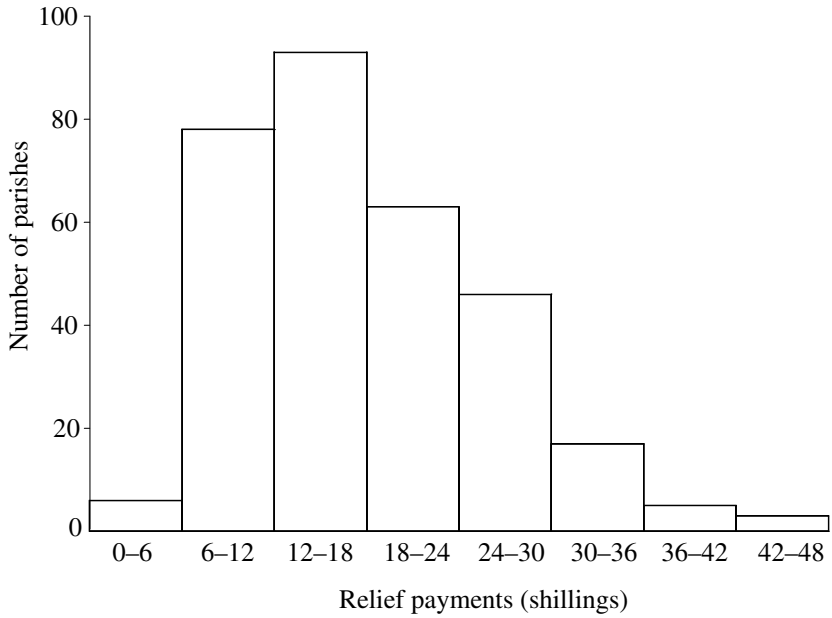
relative frequency in column (3), and the cumulative relative frequency in column (4). The relative frequencies in column (3) are shown with a total of 100. It would make no difference if they were all divided by 100 and summed to 1.

Since there is by construction a strictly proportional relationship between the absolute frequency in column (2) and the relative frequency in column (3) – the latter is always exactly a fraction of the former (in this case, equal to $1 \div 311$) – it will make no difference whatsoever to the appearance of the histogram whether we plot it using column (2) or column (3).

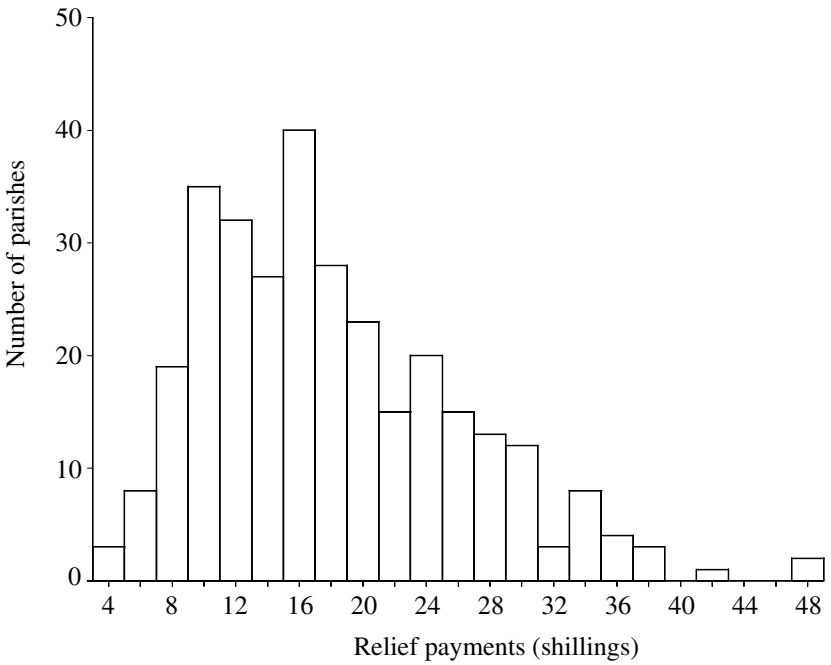
If it is possible to keep all class intervals *the same width*, then the visual impact of the histogram is effectively determined by the height of each of the rectangles. Strictly speaking, however, it is the *area* of each bar that is proportional to the absolute or relative frequency of the cases in the corresponding class interval.

The first two versions of the histogram for the *per capita* relief payments by all 311 parishes in figure 2.2 are of this type. In (a) there are eight class intervals as in table 2.3, each with a width of 6 shillings; in (b) the number of intervals is increased to 23, each with a width of only 2 shillings. For (b) the horizontal axis is labelled by the mid-point of each range.

**Figure 2.2**
Histograms with different class intervals for *per capita* relief payments in 311 parishes in 1831

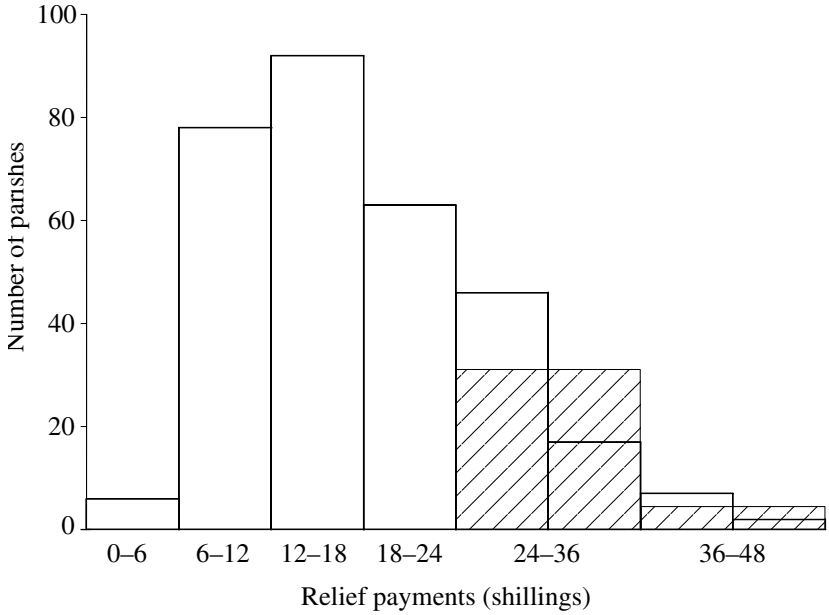



(a) With 8 equal class intervals, each 6 shillings wide




(b) With 23 equal class intervals, each 2 shillings wide

**Figure 2.2** (*cont.*)




(c) With 6 class intervals of unequal width

In this case we could select these (or other) class intervals with equal widths because we have the data for every individual parish in the data set. But what would we do if the data had already been grouped before it was made available to us, and the class intervals were not of equal width? Assume, for example, that the information on relief payments in the 311 parishes had been published with only six class intervals, the first four with a width of 6 shillings as in table 2.3, and the last two with a width of 12 shillings (i.e. with absolute frequencies of 63 and 8, and relative frequencies of 20.26 and 2.57).

Since the *width* of these last two intervals has been *doubled,* they must be shown on the histogram with their *height* set at only *half* that of the 6 shilling intervals. The *areas* (i.e. width $\times$ height) of the two sets of intervals will then be proportional. In figure 2.2 (c) the original eight-class equal-width histogram of figure 2.2 (a) is reproduced, with the alternative rectangles for the two wider class intervals superimposed on the last four intervals.

The height of the first of these 12-shilling intervals is set at 31.5, equal to half the 63 parishes in this class; the height of the second at 4, equal to half of the eight parishes in this final class. It is easy to see that when done on this basis the area of each of the new rectangles is exactly equal to that of the two corresponding rectangles in the previous version.

The advantage of the histogram in such situations is thus that despite the change in width of the intervals it is still possible to represent either the absolute or the relative frequency of the cases in each class interval by comparison of the *area* of their rectangles. Because it is the area rather than the height that is proportional to the frequency, it is always wise when looking at a histogram to refer first to the horizontal axis to ascertain the widths, and not to be misled by the visual impact of the height of each of the columns.

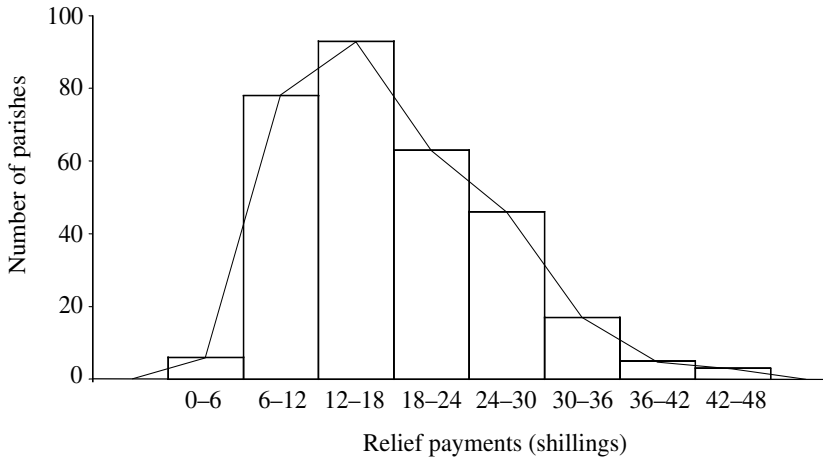
### 2.1.3 Frequency curves

Imagine a line that is drawn so that it passes through the mid-point of each of the class intervals of a frequency distribution for a *continuous* variable such as *per capita* relief payments. This graph is called a **frequency polygon.** If the histogram is like the one in figure 2.2 (a), with relatively few cases and wide class intervals, the polygon will look very different to the histogram. There will be numerous triangles that fall within the polygon but lie outside the histogram, and others that are outside the polygon but inside the histogram. However, if we look closely at figure 2.3 (a), in which the two graphs are shown on the same diagram, we see that all these triangles form exactly matching pairs, so that the area under the two graphs is *exactly* the same.

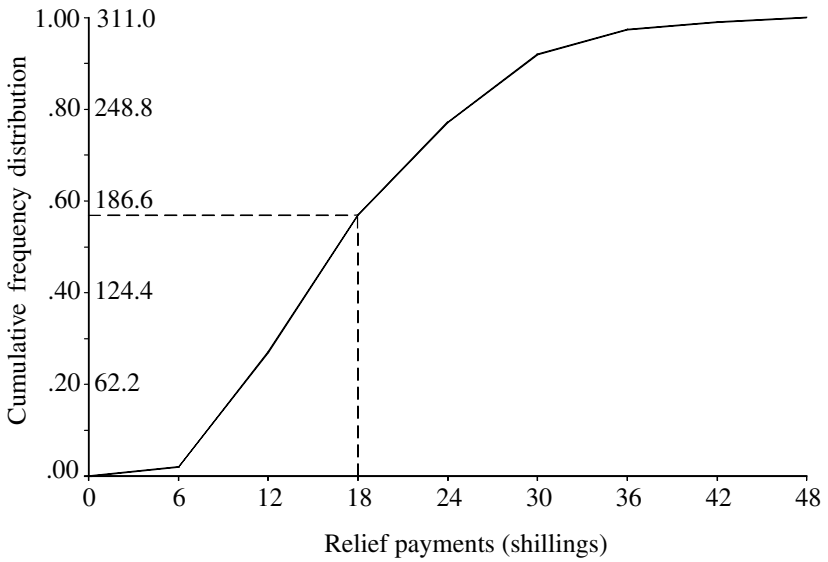
Now assume that we can gradually reduce the width of the class intervals to take full advantage of the whole scale over which a continuous variable is measured.[a] As we increased the number of rectangles in the histogram and narrowed their width, we would find that the shape of the polygons drawn through the mid-points of the rectangles approximated more and more closely to a smooth curve. We would ultimately get a perfectly smooth **frequency curve.**

It was noted in §2.1.2 that the relative frequencies in column (3) of table 2.3 could be treated as if they summed to 1, and also that the areas of the rectangles making up a histogram are proportional to the relative frequencies. Since together the rectangles cover all the class intervals, the *total area under the histogram* can also be thought of as equal to 1. We have now seen how we might in principle obtain a perfectly smooth frequency curve, and that the area under such a curve is identical to the area under a histogram. Thus the total *area under the frequency curve* can also be thought of as equal

[a] We also need to assume that we can increase the number of cases we are dealing with (i.e. that we can work with very much larger numbers than the 311 parishes in our Poor Law data set). By increasing the number we avoid the irregularities in the distribution that might occur with small numbers of cases in narrow class intervals.




(a) Histogram and frequency polygon




(b) Cumulative frequency distribution

**Figure 2.3** *Per capita* relief payments in 311 parishes in 1831

to 1. This simple but important idea of the area under a smooth frequency curve will be extremely useful in relation to the normal distribution, a concept which plays a major part in many of the statistical techniques we employ later in this book.

A different form of frequency polygon can also be used to display the **cumulative frequency distribution** defined in §2.1.1. This diagram (also known as an **ogive**) is illustrated in figure 2.3 (b) with the data from column (4) of table 2.3 for the *per capita* relief payments by the 311 parishes. The scale shown on the horizontal axis indicates the *upper* limit of the successive class intervals, and that on the vertical axis indicates both the cumulative relative and cumulative absolute frequency. (The values for the latter are obtained by cumulating column (2) of table 2.3.)

The curve of the cumulative frequency distribution traces either the proportion or the number of cases that are *less than* the corresponding value of the variable shown on the horizontal axis. Since there is an exact correspondence between the absolute and relative frequencies, the two graphs are identical. For example, the position marked off by the broken lines in figure 2.3 (b) indicates that relief payments of less than 18 shillings per person were made by 177 parishes, corresponding to a proportion of 0.569 (56.9 per cent). The s-shape is characteristic of such cumulative frequency diagrams.

## 2.2 Measures of central tendency

If we refer again to the data set in table 2.1, there are 24 observations, each giving the *per capita* expenditure on relief in one of the Kent parishes in 1831. Grouping the data in class intervals in a frequency table (such as table 2.2), and plotting this as a histogram, provides a more helpful indication of some features of the data, but this is neither a convenient nor a precise way in which to *summarize* the information. In order to achieve this we need to determine three aspects of the data:

(a) Which are the *central* (i.e. the most common or typical) values within the distribution?
(b) How is the distribution *spread* (dispersed) around those central values?
(c) What is the *shape* of the distribution?

Each of these features can be described by one or more simple statistics. These are the basic elements of descriptive statistics, and together they provide a precise and comprehensive summary of a data set. We shall now look at each in turn.

There are three principal measures that are used to locate the most common or most typical cases. They are referred to collectively as **measures of central tendency**.

### 2.2.1 The mean, the median, and the mode

(a) The **arithmetic mean**, also called the arithmetic average, is obtained by adding up all the values and dividing this total by the number of observations.[b] It is usually denoted by $\bar{X}$ (X bar) when referring to a sample, and by $\mu$ (lower case Greek mu) for the corresponding population. Thus

$$\bar{X} = \frac{\Sigma X}{n} \tag{2.1}$$

where n is the number of observations in the data set for which there are values for the variable, X.[c]

(b) The **median** is the value that has one-half of the number of observations respectively above and below it, when the series is set out in an ascending or descending array. Its value thus depends entirely on whatever happens to be the value of the observation in the middle of the array; it is not influenced by the values of any of the other observations in the series.

For example, if there are five observations, the median is the value of the third observation, and there are two cases above this and two below. When there is an even number of observations, the average of the two middle observations is taken as the median.

(c) The **mode** is the value that occurs most frequently (i.e. is most fashionable). When data are grouped in class intervals the mode is taken as the mid-point of the category with the highest frequency. In a frequency distribution the mode is represented by the highest point on the curve. A distribution may have more than one mode; one with two modes is described as **bimodal.**

[b] This measure is frequently referred to simply as the mean. When no further qualification is given the term 'mean' can be understood to refer to the arithmetic mean. However, it is also possible to calculate other means, such as a **geometric mean**, and when there is any possibility of confusion the full titles should be used. The geometric mean is calculated by multiplying all the values and taking the appropriate root: thus if there are five terms in a series ($X_1$, $X_2$, ... , $X_5$) the geometric mean would be the fifth root of their product:

$$\sqrt[5]{X_1 \times X_2 \times X_3 \times X_4 \times X_5}$$

[c] We will follow the convention of using lower case n to refer to the number of observations in a sample. The symbol for the corresponding number in the population is upper case N.

The three primary measures can be illustrated with the following simple data set containing seven observations set out in ascending order:

3, 5, 5, 7, 9, 10 and 45
The arithmetic mean = $\Sigma X/n = 84/7 = 12$[d]
The median = 7
The mode is 5

The most useful of these measures for most purposes are the mean and the median. The mean is generally to be preferred because it is based on all the observations in the series, but this can also be its weakness if there are some extreme values. For example, in the series just given the mean is raised by the final very large value, whereas this observation has no effect on the median. It is a matter of judgement in any particular study whether or not to give weight to the extreme values in reporting a measure of central tendency. For the data on RELIEF in table 2.1 the mean is 20.3 and the median 19.4.

The mode is the only one of these three measures that can be used with either nominal or ordinal level measurements.

### 2.2.2 Weighted averages and standardized rates

The simple mean defined in §2.2.1 effectively treats all n observations as having equal importance, and is the measure of central tendency most often referred to in statistical work. However, for certain purposes it may not be appropriate to give the same weight to each observation when some observations represent a larger share of the relevant total than others. For example, if the aim were to measure the average change in prices of various farm products between 1920 and 1938, it would not be appropriate to give the same weight to minor products (such as pumpkins) as to major products (such as wheat).

To illustrate a very useful alternative procedure we will take the data for CNTYMIG in the Irish emigration data set (for simplicity we restrict the example to the first four counties and round the data to 1 decimal place). To begin with, consider only the data in columns (1) and (2) of table 2.4 for 1881.

It is immediately evident that there are considerable differences in the rate of migration per 1,000 of population from these four counties, and that the one with the lowest rate – Dublin – also has the largest population.

[d] The geometric mean is 8.02, indicating that the extreme value (45) has less effect on this measure than on the arithmetic mean.

**Table 2.4** Population and migration for four Irish counties in 1881 and 1911

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| | 1881 | | 1911 | |
| County | Population (000) | CNTYMIG (per 1,000) | Population (000) | CNTYMIG (per 1,000) |
| Carlow | 46.6 | 29.0 | 36.3 | 5.9 |
| Dublin | 418.9 | 6.8 | 477.2 | 2.3 |
| Kildare | 75.8 | 19.8 | 66.6 | 4.4 |
| Kilkenny | 99.5 | 11.9 | 75.0 | 4.9 |

$$\text{Arithmetic mean of CNTYMIG in 1881} = \frac{29.0 + 6.8 + 19.8 + 11.9}{4} = \frac{67.5}{4} = 16.9$$

$$\text{Weighted arithmetic mean of CNTYMIG in 1881} = \frac{(29.0 \times 46.6) + (6.8 \times 418.9) + (19.8 \times 75.8) + (11.9 \times 99.5)}{46.6 + 418.9 + 75.8 + 99.5} = \frac{6884.7}{640.8} = 10.7$$

Conversely Carlow has the highest rate of migration but the smallest population. If we ignore these differences in the migration rates and size of the counties and calculate the simple mean we get an average of 16.9 migrants per 1,000 (as shown at the foot of table 2.4).

If we wish to take these differences into account, we must instead use a **weighted arithmetic mean**. This requires that each county is given an appropriate weight, in this case its population size in 1881. The weighted mean is then calculated by multiplying each rate of migration by its weight, and dividing the total of these products by the sum of the weights. The calculation is set out at the foot of table 2.4, and the result is 10.7 per 1,000. This is considerably lower than the unweighted mean, and gives a much more accurate measure of the overall rate of migration from these counties.

More generally, the formula for a weighted mean, $\bar{X}'$, is

$$\bar{X}' = \frac{\Sigma(w_i X_i)}{\Sigma w_i} \tag{2.1a}$$

where $w_i$ is the appropriate weight for each observation, $X_i$, in the series.

We will also introduce one particular form of weighted average that is very widely used in demographic and medical history, the standardized rate.

If we look at the 1911 data in columns (3) and (4) of table 2.4 we see that the emigration rates have fallen sharply in all four counties, and that the population in Dublin has increased since 1881 while that in the other three counties has fallen. The weighted average rate of CNTYMIG for 1911 is only 3.0.

However, this decline as compared to 10.7 in 1881 is partly a result of changes in migration from the individual counties, and partly a result of changes in the relative importance of the different counties. If we wish to get a summary measure of the rate of emigration which is *unaffected by the population changes* we can calculate a **standardized rate**. In this case we would ask, in effect: what would the 1911 rate have been if the distribution of the population by county were fixed at the proportions of 1881? The answer is obtained by weighting the four 1911 county rates by their respective 1881 populations, and is 3.2.

This procedure has very wide application, with the possibility of standardizing a variety of rates for one or more relevant factors. For example, crude birth rates per 1,000 women are frequently standardized by age in order to exclude the effects of changes in the age composition of the female population. Crude death rates per 1,000 of the population might be standardized by both age and occupation; suicide rates by both geographical region and gender; marriage rates by social class; and so on. In every case the standardized rate is effectively a weighted average of a particular set of specific rates such as the age-specific birth rates or the age- and occupation-specific death rates.

### 2.2.3 Percentiles, deciles, and quartiles

It was noted in §2.2.1 that the median is calculated by taking the value in an array above and below which there are one-half of the observations. Exactly the same principle can be applied for other measures, but taking other fractions.

(a) **Quartiles:** Instead of dividing the observations into two equal halves we can divide them into four equal quarters. The first quartile is then equal to the value that has one-quarter of the values below it, and three-quarters above it. Conversely the third quartile has three-quarters of the values below it, and one-quarter above. The second quartile has two quarters above and two quarters below, and is thus identical to the median.

(b) **Percentiles and deciles**: Other common divisions are percentiles, which divide the distribution into 100 portions of equal size, and deciles, which divide it into 10 portions of equal size. So if you are told at the end of your course on quantitative methods that your mark is at the ninth decile, you will know that nine-tenths of the students had a lower mark than you, and only one-tenth had a better mark.

The fifth decile and the fiftieth percentile are the same as the median; the 25th and 75th percentiles are the same as the first and third quartiles.

## 2.3 Measures of dispersion

Knowledge of the mean or some other measure of central tendency tells us something important about a data set, but two sets of numbers can have the same mean and still be markedly different. Consider, for example, the following (fictitious) numbers relating to the profits (in dollars) earned by two groups of five farmers in counties A and B:

County A: 14, 16, 18, 20, and 22
County B: 2, 8, 18, 29, and 33

Both these series sum to 90 and have the same mean of 18 dollars, but it is immediately obvious that the spread of numbers around this shared central value is very much greater in the second county. Our view of the risks and rewards of farming in the two regions would be influenced accordingly, and we would want to include some indication of this factor in any summary statement about farm incomes.

To do this we need some measure of the **dispersion** or spread of the data around its central value, and there are a variety of different ways such a measure can be constructed.

### 2.3.1 The range and the quartile deviation

(a) The **range** is a very crude measure of dispersion and is simply the difference between the highest and lowest value in the series. Its obvious weakness is that it is based entirely on these two extreme values, and takes no account whatsoever of all the other observations.
(b) The **quartile deviation** (or inter-quartile range) is a slightly better measure. Instead of representing the range as the difference between the maximum and minimum values it measures the distance between the first and third quartiles (see §2.2.3). It is sometimes divided by 2 and called the **semi-interquartile range**. In either form it is thus a measure of the distance covered by the middle half of all the observations, and is less sensitive to extreme values of the distribution than the range. However, it still ignores a considerable part of the data, and it does not provide any measure of the variability among the cases in the middle of the series.

The remaining measures of dispersion avoid both these weaknesses.

### 2.3.2 The mean deviation, the variance, and the standard deviation

If the objective is to get a good measure of the spread of *all* the observations in a series, one method which suggests itself is to find out how much each

value differs from the mean (or some other central or typical value). If we then simply added up all these differences, the result would obviously be affected by the number of cases involved. So the final step must be to take some kind of average of all the differences.

The three following measures all do this in varying ways.

(a) The **mean deviation** is obtained by calculating the difference between each observation ($X_i$) and the mean of the series ($\overline{X}$) and then finding the average of those deviations.

In making this calculation the *sign* of the deviation (i.e. whether the value of the observation is higher or lower than the mean) has to be ignored. If not, the sum of the positive deviations would always be exactly equal to the sum of the negative deviations. Since these two sums would automatically cancel out, the result would *always* be zero.

The mean deviation does take account of all the observations in the series and is easy to interpret. However, it is less satisfactory from a theoretical statistical point of view than the next two measures.

(b) The **variance** uses an alternative way of getting rid of the negative signs: it does so by calculating the *square* of the deviations.[e] It is then calculated by finding the mean of the squared deviations. This is equal to the sum of the squared deviations from the mean, divided by the size of the sample.[f]

**THE VARIANCE**

The variance, $s^2$, is equal to
the *arithmetic mean of the squared deviations from the mean.*

The formula is thus

$$s^2 = \frac{\Sigma(X_i - \overline{X})^2}{n} \qquad (2.2)$$

[e] Taking squares solves the problem because one of the rules of elementary arithmetic is that when a negative number is multiplied by another negative number the result is a positive number; e.g. $-4 \times -4 = +16$.

[f] When we get to the deeper waters of chapter 5 we will find that the statisticians recommend that n should be replaced by $n - 1$ when using sample data in the calculation of the variance and certain other measures. We will ignore this distinction until we reach that point. Note, however, that if you use a computer program or calculator to check some of the calculations done 'by hand' in chapters 2–4 you will get different results if your program uses $n - 1$ rather than n. The distinction between the two definitions matters only when the sample size is small. If $n = 30$ or more there would be practically no difference between the two estimates.

**Table 2.5** Calculation of the sample variance for two sets of farm incomes

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| | | County A | | | County B | |
| | $X_i$ | $(X_i - \bar{X})$ | $(X_i - \bar{X})^2$ | $X_i$ | $(X_i - \bar{X})$ | $(X_i - \bar{X})^2$ |
| | 14 | −4 | 16 | 2 | −16 | 256 |
| | 16 | −2 | 4 | 8 | −10 | 100 |
| | 18 | 0 | 0 | 18 | 0 | 0 |
| | 20 | 2 | 4 | 29 | 11 | 121 |
| | 22 | 4 | 16 | 33 | 15 | 225 |
| Sum | 90 | 0 | 40 | 90 | 0 | 702 |
| n = 5 | | | | | | |
| Mean ($\bar{X}$) | 18 | | | 18 | | |
| Variance ($s^2$) | | | 8 | | | 140.4 |

The calculation of the variance for the simple data set given above for the hypothetical farming incomes in the two counties is set out in table 2.5.

The variance of County A = 40/5 = 8, while that of County B = 702/5 = 140.4, and is very much greater. Note that the result of squaring the deviations from the mean is that the more extreme values (such as incomes of 2 dollars or 33 dollars) have a very large impact on the variance.

This example was deliberately kept very simple to illustrate how the variance is calculated and to show how it captures the much greater spread of incomes in County B. In a subsequent exercise we will take more realistic (and larger) samples, and use a computer program to calculate the variance.

The variance has many valuable theoretical properties and is widely used in statistical work. However, it has the disadvantage that it is expressed in square units. In the farm income example we would have to say that the variance in County A was 8 'squared dollars', which is not very meaningful. The obvious way to get rid of these awkward squared units is to take the *square root of the variance*. This leads to our final measure of dispersion.

(c) The **standard deviation** is the square root of the variance.

Thus:

The standard deviation in County A equals $\sqrt{8} = 2.8$ dollars
The standard deviation in County B equals $\sqrt{140.4} = 11.8$ dollars

**THE STANDARD DEVIATION**

The standard deviation, s,

is equal to
the *square root of the*
*arithmetic mean of the squared deviations from the mean.*

The formula is thus

$$s = \sqrt{\frac{\Sigma(X_i - \overline{X})^2}{n}} \qquad (2.3)$$

The standard deviation is the most useful and widely used measure of dispersion. It is measured in the same units as the data series to which it refers. Thus we can say that the mean income in County A is 18 dollars, with a standard deviation (s.d.) of 2.8 dollars. Such results will often be reported in the form: $\overline{X} = 18$, s.d. $= 2.8$, or in even more summary form as: $\overline{X} = 18 \pm 2.8$.

The standard deviation can be thought of as the average or typical (hence standard) deviation from the mean. Thus it will be seen that in County A the deviations from the mean (see column (2) of table 2.5) vary between 0 and 4, and the standard deviation is 2.8. Similarly in County B the spread of the deviations in column (5) is from 0 to 16, and the standard deviation is 11.8. The standard deviation thus lies somewhere between the smallest and largest deviation from the mean.

The variance and the standard deviation have several mathematical properties that make them more useful than the mean deviation. Both play an extremely important role in many aspects of statistics.

### 2.3.3 The coefficient of variation

The standard deviation is calculated in the same units (e.g. of weight or length or currency or time) as the series to which it refers. This makes it specific to that series, and means that it is difficult or misleading to compare the *absolute* standard deviations of two series measured in different underlying units. This would also be true even with the same unit, if that is used for values where there is a substantial change in level.

For example, we might discover that in 1914, when the mean weekly wage of adult male workers in the United Kingdom was £1.60, the standard deviation was £0.30; and that in 1975 when the mean was £75, the standard

deviation was £17.50 (these are very rough approximations to the actual data). Because growth and inflation has completely altered the level of wage payments, it is impossible to tell from this whether the dispersion of wages was larger or smaller in the later period.

To do this we need a measure of *relative rather than absolute variation.* This can be obtained by dividing the standard deviation by the mean. The result is known as the **coefficient of variation,** abbreviated to CV (or cv)

$$CV = s/\bar{X} \tag{2.4}$$

The two estimates for CV are, therefore, $0.3/1.6 = 0.19$ and $17.5/75 = 0.23$ and it is evident that the variation of wages has increased slightly.[g]

## 2.4 Measures of central tendency and dispersion with grouped data

In the preceding description of the procedures for computation of the measures of central tendency and dispersion it was taken for granted that the full set of values was available for all the cases. It may be, however, that for some projects the relevant information is available only in the form of **grouped** data. Imagine, for example, that instead of the individual data for relief payments in each of the 311 parishes, the only information that survived in the historical record was the grouped data of table 2.3. Certain additional *assumptions* would then be required in order to calculate the measures of central tendency and dispersion.

The results obtained for grouped data on these assumptions will not be exactly the same as those that we would get if we had all the individual data. However the inaccuracies will generally be minor unless the class intervals are very wide or there are serious errors in the values attributed to the open-ended classes.[h]

### 2.4.1 The median and mean with grouped data

In order to calculate the median, mean, and other measures with such grouped data it is necessary to make some assumptions about the distribution of the unknown individual cases *within* the class intervals. For the median, quartiles, and similar measures the conventional assumption is that the cases are all spread *at equal distances* within their respective intervals.

[g] CV is sometimes expressed as a percentage. In the above example multiplying by 100 would give results of 19 per cent and 23 per cent, respectively.

[h] The scale of the possible differences is illustrated by question 1 in the exercises at the end of this chapter.

To illustrate this procedure, consider the data on *per capita* relief payments in table 2.3. Since there are 311 parishes, the median parish will be the 156th. There are 84 parishes in the first two class intervals, so we need another 72 to reach the median parish, and it will thus fall somewhere in the third class interval, which has a lower limit of 12 shillings. There are a total of 93 parishes in that class interval and, by assumption, they are spread at equal distances along the interval of 6 shillings. The median parish will therefore occur at the value equal to 12 shillings plus 72/93 of 6 shillings, or 16.6 shillings.[i]

For measures involving the mean, the corresponding assumption is that all the cases within the group have a value equal to the *mid-point* of the group. Thus we take a value of 15 shillings for all 93 parishes in table 2.3 with relief payments between 12 and 18 shillings, and similarly for the other intervals.[j]

Let us denote each of these mid-points by $X_i$, and the frequency with which the cases occur (as shown in column (2) of table 2.3) by $f$. These two values can then be multiplied to get the product for each class interval, $fX_i$. The sum of those products is thus $\Sigma fX_i$, and the formula for the mean with grouped data is then this sum divided by the sum of the frequencies, $\Sigma f$

$$\bar{X} = \frac{\Sigma f X_i}{\Sigma f} \tag{2.5}$$

There are two points to note with regard to this formula. First, for the data in table 2.3, the denominator in this formula, $\Sigma f$, is 311, which is precisely what n would be when calculating the mean with ungrouped data using the formula in (2.1). Secondly, the procedure is exactly equivalent to the weighted arithmetic mean introduced in §2.2.1, as can be readily seen by comparing this formula with the one in (2.1a). The mid-points of the class intervals correspond to the observations in the series, and the frequencies become the weights.

### 2.4.2 The variance and standard deviation with grouped data

The calculation of the variance with grouped data then involves the following steps. Find the deviations of each of the mid-points from this mean value $(X_i - \bar{X})$; square each of these terms $(X_i - \bar{X})^2$; and multiply this by the corresponding frequency, $f$, to obtain the product $f(X_i - \bar{X})$. The variance is then equal to the sum of these products divided by the sum of the frequencies

[i] This result is rounded to 1 decimal place. Note that a result can never be more precise than the data on which it is based.

[j] If there are any open-ended classes (e.g. if the last class in table 2.3 had been reported only in the form of a lower limit of 'equal to or greater than 42 shillings') it would be necessary to make an intelligent guess as to a sensible value for the mid-point.

$$s^2 = \frac{\Sigma f(X_i - \overline{X})^2}{\Sigma f} \tag{2.6}$$

The standard deviation can similarly be calculated as the square root of the variance

$$s = \sqrt{\frac{\Sigma f(X_i - \overline{X})^2}{\Sigma f}} \tag{2.7}$$

## 2.5 The shape of the distribution

### 2.5.1 Symmetrical and skewed distributions

The final aspect of the distribution that is of interest is its shape, i.e. the way in which the data are distributed around the central point. The distribution around the mean may be symmetrical or it may be skewed either to the left or to the right.

In §2.1.3 we noted that if we have a sufficiently large number of cases for a variable measured on a continuous scale, and if these cases are then grouped in sufficiently narrow class intervals, a line drawn through the mid-point of each of the class intervals ultimately becomes a perfectly smooth frequency curve.

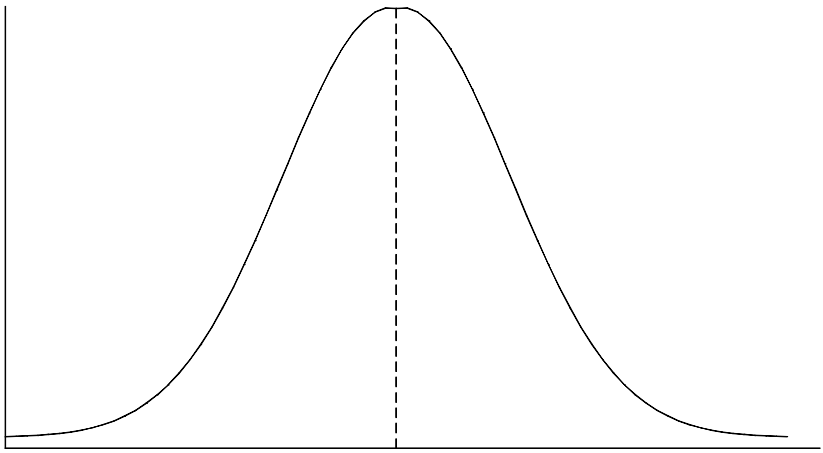
One subset of such smooth frequency curves is perfectly symmetric. This applies to the **rectangular distribution**, in which case the proportion of frequencies in each of a succession of equally spaced intervals is identical, and also to the **pyramidical distribution.**[2] The most important member of this family of *perfectly smooth and symmetrical* frequency curves is also *bell-shaped* and is known as the **normal curve,** illustrated in figure 2.4 (a). Because the curve is symmetrical and unimodal, the mean, median, and mode are all identical. The normal curve plays a very large part in this book and will be discussed more fully in §2.6.1.

Other families of distribution are smooth but *not symmetrical.* For example, a distribution may be *skewed to the right*, as in figure 2.4 (b), where there are large numbers of small or medium values and a tail stretching to the right with a small number of very large values (**positive skewness**). Examples of such positive skewness are often found with data on income distribution or holdings of land and other property, and on the size of firms. Data on the age of marriage or on the number of births per married women would also be positively skewed.[k]

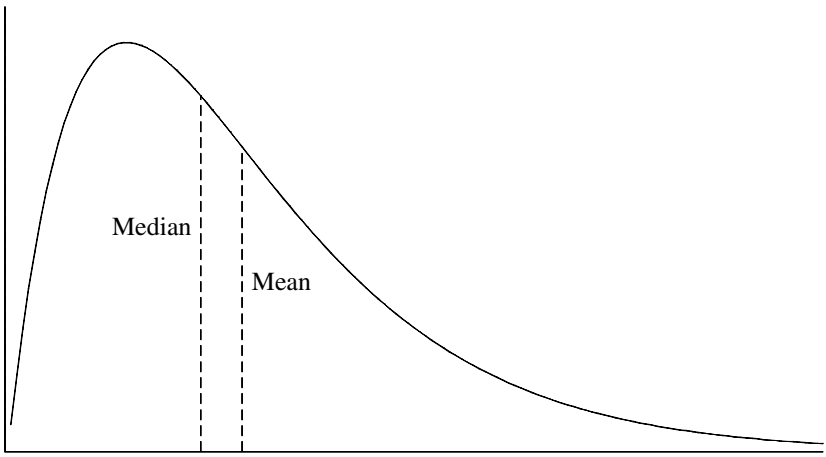
Alternatively, a distribution may be *skewed to the left*, as in figure 2.4 (c),

[k] Such distributions are sometimes loosely referred to as **log-normal.** A true log-normal distribution is one that is strongly positively skewed when the data are entered in ordinary numbers and normally distributed when the data are converted to logs.

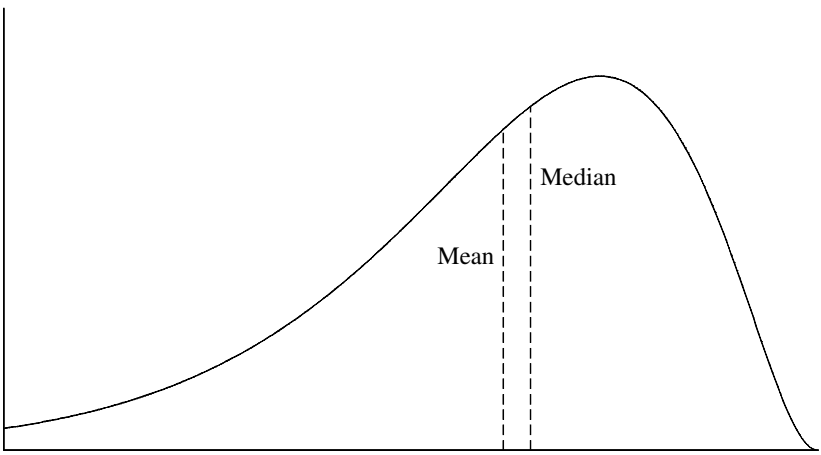
**Figure 2.4**
Symmetrical and skewed frequency curves



(a) The normal curve




(b) Smooth curve skewed to the right (positive skewness)




(c) Smooth curve skewed to the left (negative skewness)

with a tail made up of a small number of very low values (**negative skewness**). This would be characteristic of data on the age of death in Britain or the United States in any normal peacetime year. There would be a small number of deaths at low ages as a result of accidents and illness, but the great majority would occur at ages 60–90 and the right hand tail would effectively end abruptly a little above age 100.

### 2.5.2 Measures of skewness

Measures of skewness should be independent of the units in which the variable is measured, and should be zero when the distribution is perfectly symmetrical so that there is no skewness. One simple measure of the extent to which a distribution is asymmetrical (skewed) is given by the relationship between the mean and the median: the greater the skewness the larger the difference between the mean and the median. If the distribution is positively skewed, the mean will be raised by the large values in the right-hand tail of the distribution, whereas the median is unaffected by these extreme values. The mean will thus be larger than the median. Conversely, if the distribution is negatively skewed, the mean will be smaller than the median. A very rough indication of the presence of either positive or negative skewness in any particular distribution can thus be obtained by a comparison of the mean and median of a distribution.

A simple measure based on these features is a variant of the Pearson coefficient of skewness

$$\text{Skewness} = \frac{3(\text{Mean} - \text{Median})}{\text{Standard deviation}} \tag{2.8}$$

This satisfies the two criteria mentioned above and would lie between –3 and +3, though these limits are seldom encountered in practice.

Statistical packages tend to use more precise measures based on the deviations of the individual observations from the mean. In one such measure, these deviations are cubed, and expressed in terms of the number of standard deviations, also cubed. The arithmetic mean of this term is then calculated. The formula is thus

$$\frac{\Sigma(X_i - \overline{X})^3/s^3}{n} \tag{2.9}$$

## 2.6 The normal distribution

### 2.6.1 Properties of the normal distribution

Since the *perfect* normal curve is a theoretical distribution based on an infinitely large number of observations it can be only approximated by *actual* frequency distributions, but we will find many examples of random variables that show a very good approximation to the theoretical normal distribution. We will also find that the normal distribution has a number of remarkable properties.

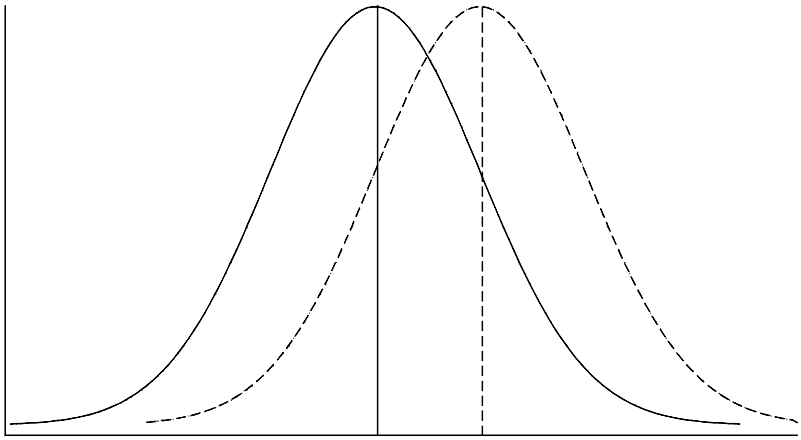
A variable is defined as a **random variable** if its numerical value is *unknown until it is observed*; it is not perfectly predictable. The observation can be the outcome of an experiment such as tossing a coin or planting seeds, or it can be the measurement of a variable such as height, the price of wheat, or the age of marriage. If it can take only a finite number of values (for example, the number of births per married women is limited to positive whole numbers such as 0, 1, 2, or 3) it is called a **discrete** random variable. If it can take any value (within a relevant interval) it is called a **continuous** random variable.

The normal curve is defined by an equation involving two constants and the mean and standard deviation of the distribution of a continuous random variable, X. The equation gives the value of Y (the height of the curve, shown on the vertical axis) for any value of X (measured along the horizontal axis). Y thus measures the frequency with which each value of X occurs. Although the form of this equation is considerably more complicated than the one used in §1.5 for a straight line ($Y = a + bX$), the principle is exactly the same. We select the values for the constants, in this case they are $\pi$ (lower case Greek pi, approximately 3.1416) and e (approximately 2.7183), plug in the changing values for the variable(s) in the equation (in this case the values for X), and calculate the corresponding values for Y.

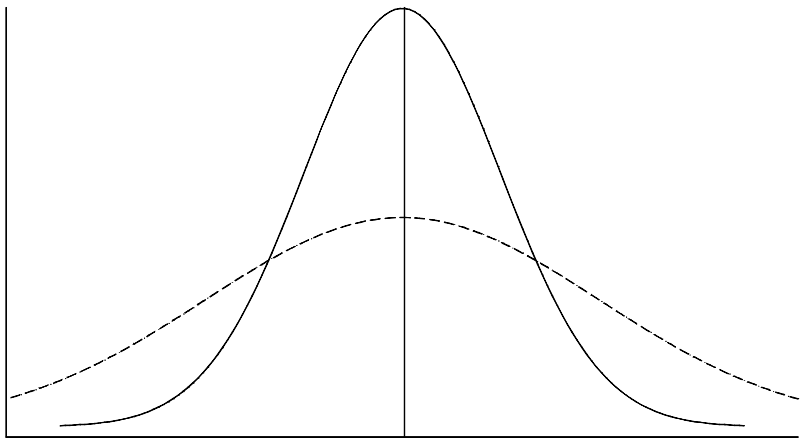
The normal curve is not a single, unique, curve. All normal curves will have the same smooth and symmetrical bell-like shape, but there are very many different normal curves, each defined by its particular mean and standard deviation. Two normal curves may have the same spread (standard deviation) but different means, in which case the one with the larger mean will be further to the right along the horizontal (X) axis, as in figure 2.5 (a). Or they may have the same mean but different standard deviations, in which case the one with the larger standard deviation will be spread out more widely on either side of the mean, as in figure 2.5 (b). Or both means and standard deviations may be different, as in figure 2.5 (c).

When a statistical program is used to plot a histogram for any particular data set it can also be requested to add the particular normal curve

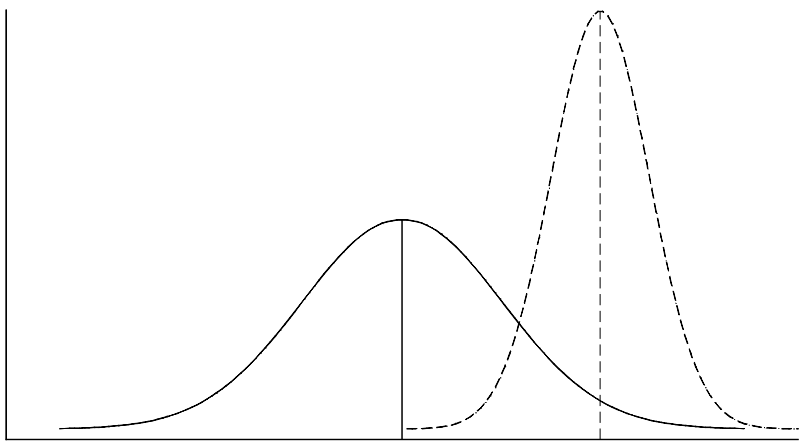
**Figure 2.5** A family of normal curves



(a) Different means but the same standard deviation



(b) The same mean but different standard deviations



(c) Different means and standard deviations

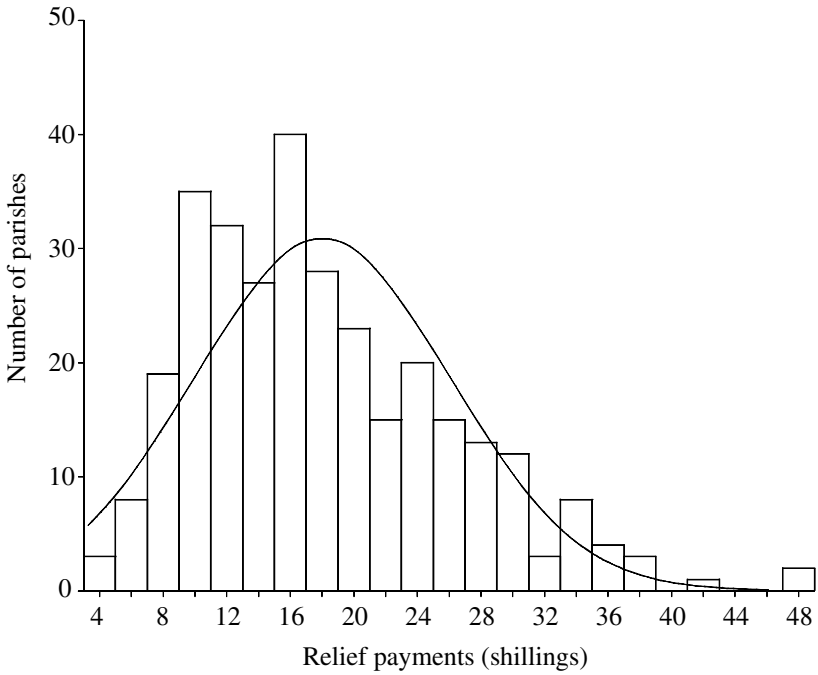



**Figure 2.6**
Histogram of *per capita* relief payments in 311 parishes in 1831 with normal curve superimposed

given by the mean and standard deviation of that data. It is thus possible to see how far the data conform to the normal distribution. This is done, for example, in figure 2.6 with the histogram of parish relief payments previously shown in figure 2.2 (b), and it can be seen that it is a reasonably good approximation.

The fact that many actual distributions approximate the theoretical normal distribution enables statisticians to make extensive use of the properties of the theoretical normal curve. One obvious property that we have already noted implicitly is that the mean, median, and mode are all equal; they coincide at the highest point of the curve and there is only one mode.

A second property of considerable importance relates to the area under the normal curve. Irrespective of the particular mean or standard deviation of the curve it will always be the case that *a constant proportion of all the cases will lie a given distance from the mean measured in terms of the standard deviation*. It is thus possible to calculate what the proportion is for any particular *distance from the mean expressed in terms of standard deviations (std devs)*.

Since the distribution is perfectly symmetrical we also know that exactly one-half of the above proportions are to the right of (greater than) the mean and one-half are to the left of (smaller than) the mean.

Three particular results, specified in the following box, are of special interest.

**SPREAD OF THE NORMAL CURVE**

**90 per cent** of all cases are within **1.645 std devs** either side of the mean, leaving 5 per cent in each of the two tails

**95 per cent** of all cases are within **1.96 std devs** either side of the mean, leaving 2.5 per cent in each of the two tails

**99 per cent** of all cases are within **2.58 std devs** either side of the mean, leaving 0.5 per cent in each of the two tails

These proportions are illustrated in the three panels of figure 2.7.

In the above description, particular areas were selected given in convenient whole numbers. It is equally possible to select convenient whole numbers for the number of standard deviations to be taken either side of the mean, in particular:

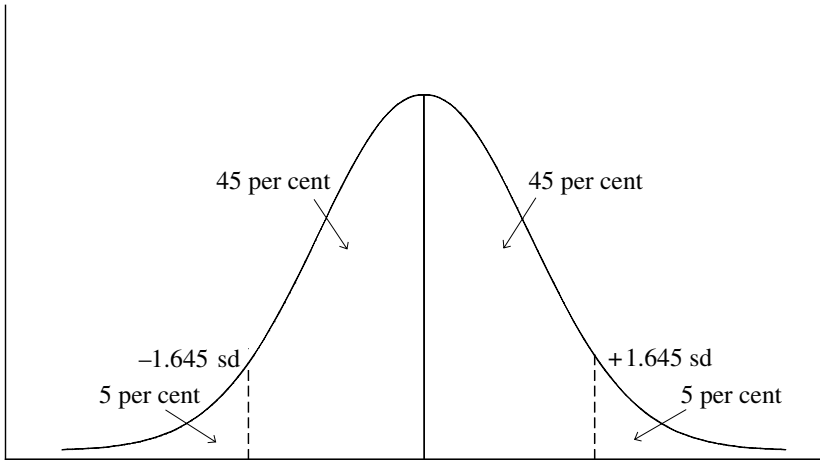
- The distance of $\pm$**1 std dev** from the mean covers **68.26 per cent** of all cases
- The distance of $\pm 2$ **std devs** from the mean covers **95.46 per cent** of all cases
- The distance of $\pm 3$ **std devs** from the mean covers **99.73 per cent** of all cases

To fix this extremely important property of the normal distribution in your mind, consider an example of a roughly normal distribution that should already be intuitively familiar. If you are told that the mean height of a large sample of male students at the University of Virginia is 6 feet (′) and the standard deviation (std dev) is 5 inches (″), you could apply the characteristics stated above and would find that:

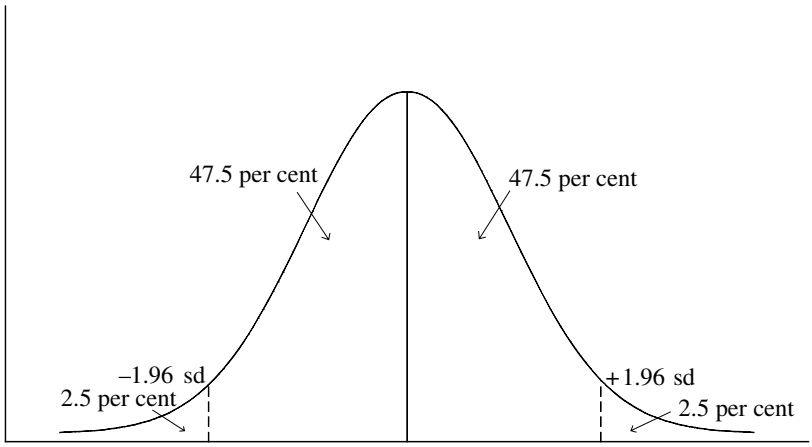
- *Roughly two-thirds* are between 5′ 7″ and 6′ 5″ tall (the mean $\pm 1$ std dev)
- That only a small minority, *less than 5 per cent,* are shorter than 5′ 2″ or taller than 6′ 10″ (the mean $\pm 2$ std devs)

**Figure 2.7** Areas under the normal curve

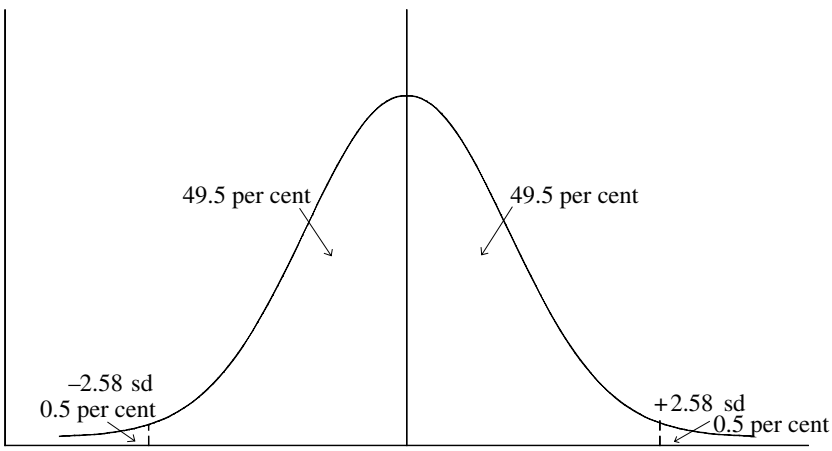
(a) 1.645 standard deviations either side of the mean covers 90 per cent of the area. This leaves 5 per cent in each tail.




(b) 1.96 standard deviations either side of the mean covers 95 per cent of the area. This leaves 2.5 per cent in each tail.




(c) 2.58 standard deviations either side of the mean covers 99 per cent of the area. This leaves 0.5 per cent in each tail.



- There are effectively *none* below 4′ 9″ or above 7′ 3″ (the mean ±3 std devs).[1]

These results should seem plausible, and should help you to understand how to interpret information about the standard deviation for (approximately normal) distributions that are less familiar than heights.

This leads directly to the idea that areas under the normal curve can be thought of in terms of the number of standard deviations.

### 2.6.2 The standard normal distribution

We have seen (§2.6.1) that there is a large family of normal curves, each defined by its particular mean and standard deviation. Because there are so many different normal curves it is helpful to *standardize* them.

To do this, note that one way to measure the distance from the mean of the distribution to any point above or below the mean is to ask how many standard units that distance represents. Thus with the height example used in the previous subsection, where the mean was 6′ and the standard deviation was 5″, we could say that if a student measured 6′ 11″ his height was 2.2 standard deviations above the mean.

This idea of the *distance from the mean* $(X - \overline{X})$ *expressed in units of standard deviations* can be generalized to standardize any normal distribution. The result creates a new variable, designated Z, and the distribution of Z is referred to as the **standard normal distribution**

$$Z = \frac{(X - \overline{X})}{s} \tag{2.10}$$

A very important feature of this *standardized* distribution is that the mean must, by definition, always be zero, and the standard deviation must always be equal to 1. This is illustrated in table 2.6, where a hypothetical (and approximately normally distributed) series for the heights of a sample of 40 male students is given in column (1), and the corresponding standardized distribution (Z) in column (4). (The distances exactly 1 or 2 standard deviations above and below the mean are highlighted.) Columns (2) and (3) give the information required to calculate Z and the standard

[1] For those more accustomed to think in metric units, the equivalent units are a mean of 180 cm and a standard deviation of 13 cm. The height of roughly two-thirds of the students is thus between 167 and 193 cm, and fewer than 5 per cent are either shorter than 154 cm or taller than 206 cm.

**Table 2.6** Heights of a sample of 40 male students and the standardized distribution

| (1) Xi Height in inches | (2) $(X_i - \bar{X})$ | (3) $(X_i - \bar{X})^2$ | (4) $(X_i - \bar{X})/s = Z$ | (5) $(Z_i - \bar{Z})^2 = Z^2$ |
|---|---|---|---|---|
| 61.1 | −10.9 | 118.81 | −2.18 | 4.75 |
| **62.0** | **−10.0** | **100.00** | **−2.00** | **4.00** |
| 63.5 | −8.5 | 72.25 | −1.70 | 2.89 |
| 64.7 | −7.3 | 53.29 | −1.46 | 2.13 |
| 65.3 | −6.7 | 44.89 | −1.34 | 1.80 |
| 66.3 | −5.7 | 32.49 | −1.14. | 1.30 |
| **67.0** | **−5.0** | **25.00** | **−1.00** | **1.00** |
| 68.0 | −4.0 | 16.00 | −0.80 | 0.64 |
| 68.2 | −3.8 | 14.44 | −0.76 | 0.58 |
| 68.3 | −3.7 | 13.69 | −0.74 | 0.55 |
| 68.5 | −3.5 | 12.25 | −0.70 | 0.49 |
| 69.2 | −2.8 | 7.84 | −0.56 | 0.31 |
| 69.3 | −2.7 | 7.29 | −0.54 | 0.29 |
| 70.0 | −2.0 | 4.00 | −0.40 | 0.16 |
| 70.3 | −1.7 | 2.89 | −0.34 | 0.12 |
| 70.7 | −1.3 | 1.69 | −0.26 | 0.07 |
| 71.4 | −0.6 | 0.36 | −0.12 | 0.01 |
| 71.7 | −0.3 | 0.09 | −0.06 | 0.00 |
| **72.0** | **0.0** | **0.00** | **0.00** | **0.00** |
| 72.4 | 0.4 | 0.16 | 0.08 | 0.01 |
| 72.6 | 0.6 | 0.36 | 0.12 | 0.01 |
| 72.8 | 0.8 | 0.64 | 0.16 | 0.03 |
| 72.9 | 0.9 | 0.81 | 0.18 | 0.03 |
| 73.4 | 1.4 | 1.96 | 0.28 | 0.08 |
| 73.7 | 1.7 | 2.89 | 0.34 | 0.12 |
| 73.9 | 1.9 | 3.61 | 0.38 | 0.14 |
| 74.1 | 2.1 | 4.41 | 0.42 | 0.18 |
| 74.3 | 2.3 | 5.29 | 0.46 | 0.21 |
| 74.5 | 2.5 | 6.25 | 0.50 | 0.25 |
| 75.4 | 3.4 | 11.56 | 0.68 | 0.46 |
| 75.6 | 3.6 | 12.96 | 0.72 | 0.52 |
| 76.3 | 4.3 | 18.49 | 0.86 | 0.74 |
| 76.5 | 4.5 | 20.25 | 0.90 | 0.81 |
| **77.0** | **5.0** | **25.00** | **1.00** | **1.00** |
| 77.3 | 5.3 | 28.09 | 1.06 | 1.12 |
| 77.5 | 5.5 | 30.25 | 1.10 | 1.21 |
| 78.8 | 6.8 | 46.24 | 1.36 | 1.85 |
| 80.3 | 8.3 | 68.89 | 1.66 | 2.76 |
| 81.2 | 9.2 | 84.64 | 1.84 | 3.39 |
| **82.0** | **10.0** | **100.00** | **2.00** | **4.00** |
| $\Sigma$ 2,880.0 | 0.00 | 1,000.00 | 0.00 | 40.00 |

n = 40

deviation of Z is calculated in column (5). The mean of Z is seen to be 0 and the standard deviation is 1.00.[m]

It will also be seen how the standardized distribution maps the initial distribution, so that the mean value of the original series (= 72 inches) corresponds to the mean value of the standardized distribution (= 0). Similarly, a distance of 1 standard deviation above or below the mean in the original distribution ($\bar{X} \pm 1$ std dev $= 72 \pm 5 = 67$ or 77 inches) corresponds to values of $Z = \bar{Z} \pm 1$ std dev $= 0 \pm 1 = -1$ or $+1$. In the same way a distance of 2 standard deviations above or below the mean (62 or 82 inches) corresponds to values of $Z = -2$ or $+2$.

Statisticians have compiled tables that show the proportion under the standardized curve for all values of Z. The proportions in these tables apply to the perfect *theoretical* distribution, but for any actual distribution that is approximately normal the proportions will be roughly the same. An extract from such a table of the standard normal distribution is reproduced in table 2.7.[n]

In this particular version the area under the curve is subdivided in five different ways, so that five different proportions of the total area are given for each *positive* value of Z. Since the curve is perfectly symmetrical the proportions for the corresponding *negative* values of Z would be exactly the same.

The five different proportions in columns (2)–(6) are:

(a) The cumulative proportion up to Z.
(b) The proportion from the mean to the specific positive value of Z. (Since the total proportion to the left of the mean = 0.5, this is always *0.5 less than the value in (a).*)
(c) The proportion in the right-hand tail beyond Z. (Since the total proportion under the curve = 1, this is always *1.0 minus the value in (a).*)
(d) The proportion in both tails given by the positive and negative value of Z. (This is always *twice the proportion in (c).*)
(e) The central proportion excluding the two tails defined by the positive and negative values of Z. (Since the total proportion under the curve = 1, this is always *1.0 minus the value in (d).*)

[m] The calculations are as follows: The mean height for these students $= \bar{X} = \Sigma X/n = 2{,}880/40 = 72.0$. The variance of the heights $= s^2 = \Sigma(X_i - \bar{X})^2/n = 1{,}000/40 = 25.00$, and so the standard deviation of these heights $= s = \sqrt{25} = 5.00$. This standard deviation is then used in column (4) to calculate Z. For Z, $\Sigma Z/n = 0.00/40 = 0.0$, and this is used in column (5) to calculate the variance of $Z = \Sigma(Z_i - \bar{Z})^2/n = 40/40 = 1.00$. The standard deviation is therefore $\sqrt{1.00} = 1.00$.

[n] A more complete table covering all values of Z can be consulted in D. V. Lindley and W. F. Scott, *New Cambridge Statistical Tables*, Cambridge University Press, 2nd edn., 1995, p.34, and in most general statistics books.

**Table 2.7** Areas under the standard normal curve for selected values of Z

| (1) Selected values of Z | (2) Cumulative area up to Z | (3) Area from mean up to Z | (4) Area in one tail beyond Z | (5) Area in both tails beyond Z | (6) Central area |
|---|---|---|---|---|---|
| 0.00 | 0.5000 | 0.0000 | 0.5000 | 1.0000 | 0.0000 |
| 0.20 | 0.5793 | 0.0793 | 0.4207 | 0.8414 | 0.1586 |
| 0.40 | 0.6554 | 0.1554 | 0.3446 | 0.6892 | 0.3108 |
| 0.60 | 0.7257 | 0.2257 | 0.2743 | 0.5486 | 0.4514 |
| 0.80 | 0.7881 | 0.2881 | 0.2119 | 0.4238 | 0.5762 |
| 1.00 | 0.8413 | 0.3413 | 0.1587 | 0.3174 | 0.6826 |
| 1.20 | 0.8849 | 0.3849 | 0.1151 | 0.2302 | 0.7698 |
| **1.28** | **0.8997** | **0.3997** | **0.1003** | **0.2006** | **0.7994** |
| 1.40 | 0.9192 | 0.4192 | 0.0808 | 0.1616 | 0.8384 |
| 1.60 | 0.9452 | 0.4452 | 0.0548 | 0.1096 | 0.8904 |
| **1.645** | **0.9500** | **0.4500** | **0.0500** | **0.1000** | **0.9000** |
| 1.80 | 0.9641 | 0.4641 | 0.0359 | 0.0718 | 0.9282 |
| **1.96** | **0.9750** | **0.4750** | **0.0250** | **0.0500** | **0.9500** |
| 2.00 | 0.9772 | 0.4772 | 0.0228 | 0.0456 | 0.9544 |
| 2.20 | 0.9861 | 0.4861 | 0.0139 | 0.0278 | 0.9722 |
| 2.40 | 0.9918 | 0.4918 | 0.0082 | 0.0164 | 0.9836 |
| **2.58** | **0.9951** | **0.4951** | **0.0049** | **0.0098** | **0.9902** |
| 2.80 | 0.9974 | 0.4974 | 0.0026 | 0.0052 | 0.9948 |
| 3.00 | 0.9986 | 0.4986 | 0.0014 | 0.0028 | 0.9972 |
| 3.20 | 0.9993 | 0.4993 | 0.0007 | 0.0014 | 0.9986 |
| 3.40 | 0.9997 | 0.4997 | 0.0003 | 0.0006 | 0.9994 |
| 3.60 | 0.9999 | 0.4999 | 0.0001 | 0.0002 | 0.9998 |
| 3.80 | 0.9999 | 0.4999 | 0.0001 | 0.0002 | 0.9998 |

*Note:*
The two-tailed area in column (5) refers to both the positive and the negative values of Z.
*Source:* Lindley and Scott, *Statistical Tables*, p. 34 for column (2); other columns as explained in the text.

You will find that authors usually give only one of these five possible proportions, and different authors choose different proportions. This divergence in the way the table is printed is confusing, and means that if you want to use one of these tables you must first establish the form in which the information is presented.

The proportions are normally given to 3 or 4 decimal places. It is often more convenient to think in terms of percentages, and these are obtained by multiplying the proportions by 100 (i.e. moving the decimal point 2 places to the right): for example, 0.500 = 50 per cent.

Once the form in which the data are presented in any particular table is clarified, it is a simple matter to find the proportion up to or beyond any given value of Z. Take, for instance, the student in table 2.6 with a height of 77 inches. This corresponds to a value for Z = 1.0, and is thus 1.0 standard deviations from the mean height of 72 inches. From column (4) of table 2.7 it can be seen that the proportion greater than this value of Z is 0.1587, or 15.9 per cent. Since there are 40 students this indicates that if the heights of this sample are normally distributed there should be approximately six students (15.9 per cent of 40) taller than this one. Table 2.6 in fact has six taller students.

## *Notes*

[1] Note that the precise class intervals depend on the extent to which the underlying data have been rounded. The data in table 2.1 were rounded to 1 decimal place, so the true upper limit of the first class interval in table 2.2 would be 9.94 shillings. All values from 9.90 to 9.94 would have been rounded down to 9.9 shillings, and so would be included in the interval $\geq 5$ but $< 10$. All values from 9.95 to 9.99 would have been rounded up to 10 shillings, and so would be included in the interval $\geq 10$ but $< 15$.

[2] If a perfect die is thrown once the probability of obtaining a six is 1/6, and the probability of obtaining each of the other face values from 1 to 5 would be exactly the same, so these outcomes would represent a rectangular distribution. A pyramid-shaped distribution would be obtained for the probability of each value if the die were thrown twice. With each successive throw beyond two the distribution moves towards the bell-shape of the normal curve. For the corresponding successive distributions when the probabilities relate to tossing increasing numbers of coins, see figure 5.1.

## 2.7 Exercises for chapter 2

**All exercises in this and subsequent chapters may be solved by computer, unless the question states specifically that they should be done *by hand.***

1. Extract the data on INCOME for Cambridgeshire (County 5) by parish from the Boyer relief data set.

   (i) *By hand*, use these data to construct:
   - An array
   - A frequency distribution using 5 class intervals of equal width of £5, beginning at <£20
   - A frequency distribution using 10 class intervals of equal width of £2, beginning at <£20
   - A frequency distribution using 16 class intervals of equal width of £1, beginning at <£20.

   (ii) Which frequency distribution provides the information on Cambridgeshire incomes in the most useful fashion?

   (iii) What difference does it make if you round the income figures to 1 decimal place before constructing the frequency distributions?

   (iv) What difference does it make if you round the income figures to the nearest whole number before constructing the frequency distributions?

   (v) Instruct the computer to construct a histogram with 10 intervals. Are the resulting class intervals the same as for the relative frequency distribution in (i) above? Explain your finding.

2. Take the WEALTH variable for every parish in the Boyer relief data set.

   (i) Calculate the following statistics:
   - Upper, lower, and middle quartiles
   - 10th and 90th percentiles
   - Range
   - Mean, median, and mode
   - Variance
   - Standard deviation
   - Coefficient of variation.

   (ii) Plot a histogram of WEALTH with: (a) 31 intervals and (b) 11 intervals.

   (iii) Plot a bar chart of mean WEALTH in each of 21 counties. (*Hint*: use COUNTY for the category axis.)

Comment on the relationship between the different measures of central tendency and dispersion, and discuss what they tell you about wealth in these parishes.

3. Using Boyer's birth rate data set, calculate the means and standard deviations of INCOME, BRTHRATE, and DENSITY, counting each parish as a single observation. Now calculate the means and standard deviations of the same variables, with each parish weighted by its population (POP). How would you explain the difference between the two measures? Is there any reason to prefer one set of measures to the other?

4. The following table provides data on the percentage of women in England and Wales, aged 10 and above, who worked outside the home in 1911 and 1991, classified by age and marital status. It also shows the percentage of women in each age group who were married in 1911 and 1991, as well as the age structure of all women, single or married, in the two years.

| | 1911 | | | | 1991 | | | |
|---|---|---|---|---|---|---|---|---|
| **Age** | **% single women working** | **% married women working** | **% all women married** | **% women by age** | **% single women working** | **% married women working** | **% all women married** | **% women by age** |
| <15 | 10.4 | 0.0 | 0.0 | 11.8 | – | – | – | – |
| <20 | 69.5 | 13.4 | 1.2 | 11.3 | 54.7 | 47.2 | 2.0 | 6.2 |
| <25 | 77.6 | 12.9 | 24.2 | 11.3 | 75.8 | 62.3 | 22.9 | 9.1 |
| <35 | 73.7 | 10.6 | 63.2 | 21.0 | 75.7 | 62.2 | 62.8 | 18.4 |
| <45 | 65.3 | 10.6 | 75.3 | 16.9 | 75.4 | 72.4 | 78.5 | 16.8 |
| <55 | 53.5 | 10.5 | 70.9 | 12.3 | 73.0 | 70.5 | 79.9 | 13.9 |
| <65 | 36.6 | 8.8 | 58.4 | 8.2 | 35.9 | 38.0 | 72.5 | 12.6 |
| <75 | 18.7 | 5.7 | 37.5 | 5.1 | 4.9 | 5.1 | 53.6 | 11.9 |
| ≥75 | 6.3 | 2.3 | 16.1 | 2.1 | 1.1 | 1.5 | 23.2 | 11.2 |
| ALL | 50.4 | 10.3 | 44.6 | 100.0 | 46.0 | 53.0 | 56.1 | 100.0 |

*Note:*
Single includes single, widowed, and divorced women. Married refers to currently married women.
*Sources:* Census of Population, England and Wales, 1911 and 1991.

(i) What proportion of ALL women were in paid employment in (a) 1911? (b) 1991?
(ii) How much of the overall change in the proportion of ALL women working between 1911 and 1991 was due to:

(a) The changing proportion at work among single and married women. (*Hint*: evaluate the level of female market work if the distribution of married and single women working had remained the same as in 1911.)
(b) The changing age structure of the female population.
(c) The changing age at marriage among women.

Write a brief paragraph setting out what the table and your calculations indicate about the changing patterns of women's involvement in paid employment.

5. Let us suppose that the data on WEALTH for the 311 parishes in the Boyer relief data set have been aggregated in the original source to produce the following table:

| **Average wealth per person (£)** | **Number of parishes** | **Population** |
|---|---|---|
| ⩾0 but <2 | 14 | 26,023 |
| ⩾2 but <3 | 44 | 90,460 |
| ⩾3 but <4 | 83 | 123,595 |
| ⩾4 but <5 | 58 | 76,421 |
| ⩾5 but <6 | 48 | 48,008 |
| ⩾6 but <7 | 25 | 19,811 |
| ⩾7 but <8 | 16 | 11,086 |
| ⩾8 but <9 | 6 | 3,820 |
| ⩾9 but <10 | 9 | 5,225 |
| ⩾10 | 8 | 2,634 |
| *ALL* | *311* | *407,083* |

*Note:*
No other information is available.

(i) Calculate total wealth in each class interval using (a) the number of parishes, and (b) the population, as the weighting system. What assumptions did you make about the level of wealth per person in each class interval and why?
(ii) Calculate the mean, median, mode, upper and lower quartiles, variance, standard deviation, and coefficient of variation for total wealth.
(iii) Change your assumption about the level of wealth per person in the top and bottom classes and recalibrate the measures of central ten-

dency and dispersion. Which measures are most sensitive to changes in the measurement of average wealth? Explain your answers.

6. Using the WEALTH data from Boyer's relief data set:

(i) Calculate the average level of wealth per person for each county. Repeat the exercise of question 5, using both the number of parishes and the number of people in each county to calculate total wealth.
(ii) Compare the measures of central value and dispersion with those produced by question 5. Identify and account for any discrepancies. How do your results in questions 5 and 6 compare to the results of question 2?

Is the loss of information in these various permutations serious? How might we evaluate any loss of information?

7. Instruct the computer to produce histograms of the following variables from the Boyer data set on births:

BRTHRATE
INFTMORT
INCOME
DENSITY

(i) In each case, inspect the shapes of the histograms and assess whether the data are normally distributed, negatively skewed, or positively skewed. Record your findings.
(ii) Now ask the computer to calculate the degree of skewness in each of these variables. Compare these results to your own findings.

8. The LONDON variable in the Boyer data set is measured by assigning to each parish in a given county the same distance from the centre of London as the mid-point of the county, in miles. This is likely to create a bias in the measurement of distance for each parish. Why? How will the bias affect the measured mean and standard deviation of the variable LONDON, relative to its 'true' value? How might you test for the extent of any differences?

9. We are interested in determining which parishes are the most and least generous providers of relief to the unemployed poor. How might we construct such a variable from the data in the RELIEF file?

10. A random sample of 1,000 women was found to have a mean age of first marriage of 23.6 years with a standard deviation of 3 years. The ages can be assumed to be normally distributed. Use the table of the standard normal distribution (table 2.7) to calculate:

(i) How many of the women were first married between 20 and 29 years?
(ii) What was the minimum age of marriage of the oldest 5 per cent of the sample?
(iii) What proportion of the women married at an age that differed from the mean by more than 1.8 standard deviations?