PART B

ANALYSIS OF NUMERICAL OUTCOMES

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

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

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

This page intentionally left blank

CHAPTER 4

Means, standard deviations and standard errors

4.1 INTRODUCTION

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

4.2 MEAN, MEDIAN AND MODE

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

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

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

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

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

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

Example 4.1

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

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

(a) $n = 8$

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

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

(b) Rearranging the measurements in increasing order gives:

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

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

$=$ average of 4th and 5th values

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

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

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

4.3 MEASURES OF VARIATION

Range and interquartile range

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

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

Range = highest value − lowest value

Interquartile range = upper quartile − lower quartile

Variance

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

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

Degrees of freedom

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

Standard deviation

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

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

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

or equivalently

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

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

Example 4.2

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

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

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

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

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

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

Interpretation of the standard deviation

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

Change of units

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

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

Coefficient of variation

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

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

4.4 CALCULATING THE MEAN AND STANDARD DEVIATION FROM A FREQUENCY DISTRIBUTION

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

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

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

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

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

The average number of previous pregnancies is, therefore:

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

In the same way:

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

The standard deviation is, therefore:

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

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

4.5 SAMPLING VARIATION AND STANDARD ERROR

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

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

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

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

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

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

Example 4.3

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

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

Understanding standard deviations and standard errors

Example 4.4

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

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

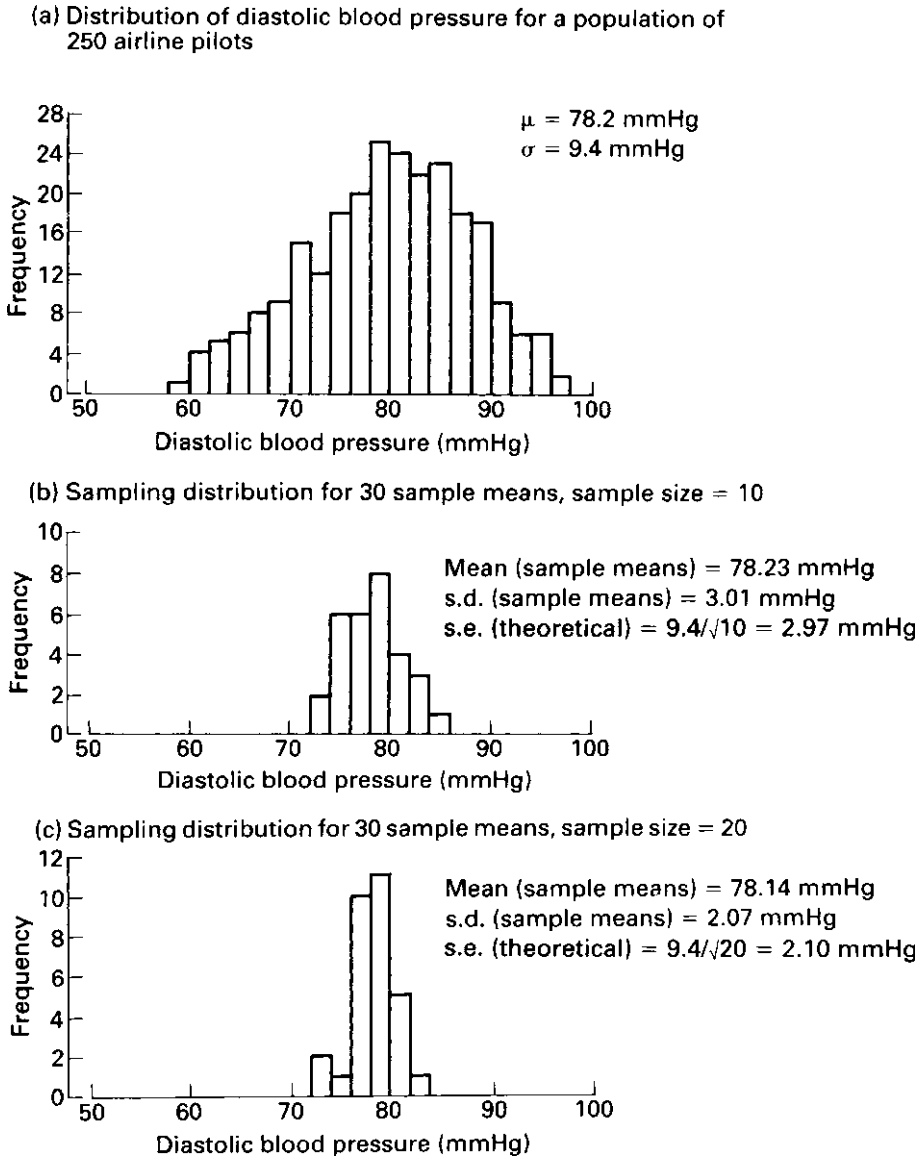


Fig. 4.1 Results of a game played to illustrate the concepts of sampling variation, the sampling distribution, and the standard error.

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

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

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

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

CHAPTER 5

The normal distribution

5.1 INTRODUCTION

Frequency distributions and their various shapes were discussed in Chapter 3. In practice it is found that a reasonable description of many variables is provided by the **normal distribution**, sometimes called the **Gaussian distribution** after its discoverer, Gauss. Its frequency distribution (defined by the **normal curve**) is symmetrical about the mean and bell-shaped; the bell is tall and narrow for small standard deviations and short and wide for large ones. Figure 5.1 illustrates the normal curve describing the distribution of heights of adult men in the United Kingdom. Other examples of variables that are approximately normally distributed are blood pressure, body temperature, and haemoglobin level. Examples of variables that are not normally distributed are triceps skinfold thickness and income, both of which are positively skewed. Sometimes *transforming* a variable, for example by

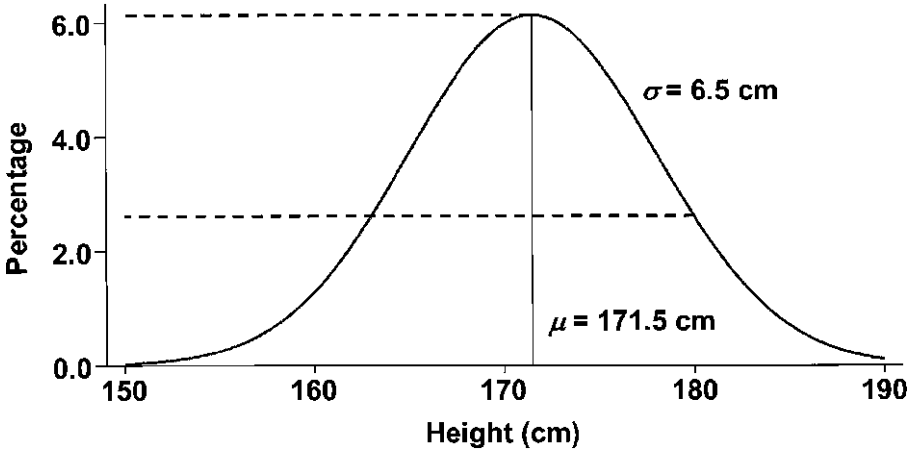


Fig. 5.1 Diagram showing the approximate normal curve describing the distribution of heights of adult men.

taking logarithms, will make its distribution more normal. This is described in Chapter 13, and methods to assess whether a variable is normally distributed are discussed in Chapter 12.

5.2 WHY THE NORMAL DISTRIBUTION IS IMPORTANT

The normal distribution is important not only because it is a good empirical description of the distribution of many variables, but because it occupies a central role in statistical analysis. This is because it can be shown that *the sampling distribution of a mean is normal*, even when the individual observations are not normally distributed, provided that the sample size is not too small. In other words, sample means will be normally distributed around the true population mean. A practical demonstration of this property can easily be had by carrying out a sampling game like Example 4.4, but with the 250 blood pressures replaced by a non-normally distributed variable, such as triceps skinfold thickness. The larger the sample selected in the game, the closer the sample mean will be to being normally distributed. The number needed to give a close approximation to normality depends on how non-normal the variable is, but in most circumstances a sample size of 15 or more is enough.

This finding is based on a remarkable and very useful result known as the **central limit theorem**. It means that calculations based on the normal distribution are used to derive confidence intervals, which were mentioned in Chapter 4, are defined fully in Chapter 6 and used throughout subsequent chapters. The normal distribution also underlies the calculation of *P*-values, which are used to test hypotheses and which are introduced in Chapter 7. The normal distribution is not only important in the analysis of numerical outcomes; we will see in parts C and D that statistical methods for proportions and rates are also based on approximations to the normal distribution.

For these reasons it is important to describe the principles of how to use the normal distribution in some detail before proceeding further. The precise mathematical equation which defines the normal distribution is included in the next section for reference only; this section can be skipped by the majority of readers. In practical terms, calculations are carried out either by a statistical package, or by using standard tables.

5.3 THE EQUATION OF THE NORMAL CURVE

The value of the normal curve with mean μ and standard deviation σ is:

$$y = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$$

where y gives the height of the curve, x is any value on the horizontal axis, exp( ) is the exponential function (see Section 13.2 for an explanation of the exponential function) and $\pi = 3.14159$. The normal curve value y is expressed as a proportion and the total area under the curve sums to 1, corresponding to the whole population.

The vertical axis can be expressed as a percentage, as in Figure 5.1, by multiplying y by 100. The area under the curve then sums to 100%.

Example 5.1

The following give two examples of calculating the height of the curve in Figure 5.1, where $\mu = 171.5$ and $\sigma = 6.5$ cm.

1 When height $x = 171.5$ cm (the mean value) then $(x - \mu) = 0$. This means that the expression inside the bracket is zero. As $\exp(0) = 1$, the height of the curve is given by

$$y = \frac{1}{\sqrt{2\pi} \times 6.5^2} = 0.0614, \text{ or } 6.14\%$$

2 When height $x = 180$ cm, the exponential part of the equation is

$$\exp\left(-\frac{(180 - 171.5)^2}{2 \times 6.5^2}\right) = 0.4253$$

and the height of the curve is given by

$$y = \frac{0.4253}{\sqrt{2\pi} \times 6.5^2} = 0.0261, \text{ or } 2.61\%$$

These values are indicated by the horizontal dashed lines on the normal curve in Figure 5.1.

5.4 THE STANDARD NORMAL DISTRIBUTION

If a variable is normally distributed then a change of units does not affect this. Thus, for example, whether height is measured in centimetres or inches it is normally distributed. Changing the mean simply moves the curve along the horizontal axis, while changing the standard deviation alters the height and width of the curve.

In particular, by a suitable change of units any normally distributed variable can be related to the **standard normal distribution** whose mean is zero and whose standard deviation is 1. This is done by subtracting the mean from each observation and dividing by the standard deviation. The relationship is:

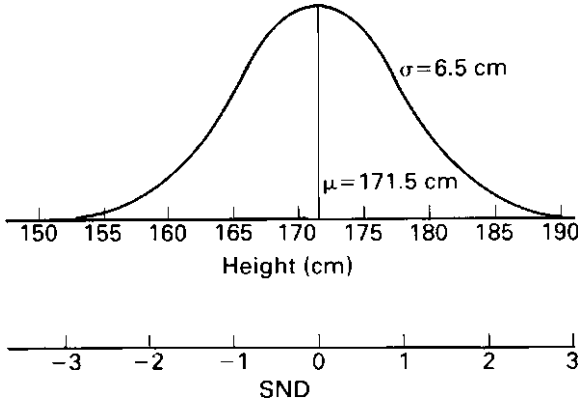


Fig. 5.2 Relationship between normal distribution in original units of measurement and in standard normal deviates. SND = (height − 171.5)/6.5. Height = 171.5 + (6.5 × SND).

$$\text{SND, } z = \frac{x - \mu}{\sigma}$$

where x is the original variable with mean μ and standard deviation σ, and z is the corresponding **standard normal deviate** (SND), alternatively called the ***z*-score**. This is illustrated for the distribution of adult male heights in Figure 5.2. The equation of the **standard normal distribution** is:

$$y = \frac{\exp(-z^2/2)}{\sqrt{2\pi}}$$

The possibility of converting any normally distributed variable into an SND means that calculations based on the standard normal distribution may be converted to corresponding calculations for any values of the mean and standard deviation. These calculations may be done either by using a computer, or by consulting tables of probability values for the normal distribution. The two most commonly provided sets of tables are (i) the area under the frequency distribution curve, and (ii) the so-called percentage points.

5.5 AREA UNDER THE CURVE OF THE NORMAL DISTRIBUTION

The standard normal distribution can be used to determine the proportion of the population that has values in some specified range or, equivalently, the probability that an individual observation from the distribution will lie in the specified range.

This is done by calculating the *area under the curve*. Calculation of areas under the normal curve requires a computer. It can be shown that the area under the whole of the normal curve is exactly 1; in other words the probability that an observation lies somewhere in the whole of the range is 1, or 100%.

Calculation of the proportion of the population in different ranges will be illustrated for the distribution shown in Figure 5.1 of the heights of adult men in the United Kingdom, which is approximately normal with mean $\mu = 171.5$ cm and standard deviation $\sigma = 6.5$ cm.

Area in upper tail of distribution

The proportion of men who are taller than 180 cm may be derived from the proportion of the area under the normal frequency distribution curve that is above 180 cm. The corresponding SND is:

$$z = \frac{180 - 171.5}{6.5} = 1.31$$

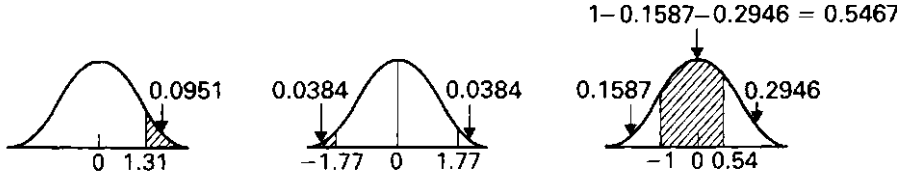
so that the proportion may be derived from the proportion of the area of the standard normal distribution that is above 1.31. This area is illustrated in Figure 5.3(a) and can be found from a computer or from Table A1 in the Appendix. The rows of the table refer to z to one decimal place and the columns to the second decimal place. Thus the area above 1.31 is given in row 1.3 and column 0.01 and is 0.0951. We conclude that a fraction 0.0951, or equivalently 9.51%, of adult men are taller than 180 cm.

Area in lower tail of distribution

The proportion of men shorter than 160 cm, for example, can be similarly estimated:

$$z = \frac{160 - 171.5}{6.5} = -1.77$$

The required area is illustrated in Figure 5.3(b). As the standard normal distribution is symmetrical about zero the area below $z = -1.77$ is equal to



(a) Area above $z = 1.31$ found directly from Table A1

(b) Area below $z = -1.77$ equals area above $z = 1.77$ by symmetry

(c) Area between $z = -1$ and $z = 0.54$ found by subtraction

Fig. 5.3 Examples of the calculation of areas of the standard normal distribution.

the area above $z = 1.77$ and is 0.0384. Thus 3.84% of men are shorter than 160 cm.

Area of distribution between two values

The proportion of men with a height between, for example, 165 cm and 175 cm is estimated by finding the proportions of men shorter than 165 cm and taller than 175 cm and subtracting these from 1. This is illustrated in Figure 5.3(c).

1 SND corresponding to 165 cm is:

$$z = \frac{165 - 171.5}{6.5} = -1$$

Proportion below this height is 0.1587.

2 SND corresponding to 175 cm is:

$$z = \frac{175 - 171.5}{6.5} = 0.54$$

Proportion above this height is 0.2946.

3 Proportion of men with heights between 165 cm and 175 cm
$= 1 -$ proportion below 165 cm $-$ proportion above 175 cm
$= 1 - 0.1587 - 0.2946 = 0.5467$ or 54.67%

Value corresponding to specified tail area

Table A1 can also be used the other way round, that is starting with an area and finding the corresponding z value. For example, what height is exceeded by 5% or 0.05 of the population? Looking through the table the closest value to 0.05 is found in row 1.6 and column 0.04 and so the required z value is 1.64. The corresponding height is found by inverting the definition of SND to give:

$$x = \mu + z\sigma$$

and is $171.5 + 1.64 \times 6.5 = 182.2$ cm.

5.6 PERCENTAGE POINTS OF THE NORMAL DISTRIBUTION, AND REFERENCE RANGES

The SND expresses the value of a variable in terms of the number of standard deviations it is away from the mean. This is shown on the scale of the original variable in Figure 5.4. Thus, for example, $z = 1$ corresponds to a value which is

one standard deviation above the mean and $z = -1$ to one standard deviation below the mean. The areas above $z = 1$ and below $z = -1$ are both 0.1587 or 15.87%. Therefore 31.74% ($2 \times 15.87\%$) of the distribution is further than one standard deviation from the mean, or equivalently 68.26% of the distribution lies within one standard deviation of the mean. Similarly, 4.55% of the distribution is further than two standard deviations from the mean, or equivalently 95.45% of the distribution lies within two standard deviations of the mean. This is the justification for the practical interpretation of the standard deviation given in Section 4.3.

Exactly 95% of the distribution lies between −1.96 and 1.96 (Fig 5.5a). Therefore the z value 1.96 is said to be the 5% **percentage point** of the normal distribution, as 5% of the distribution is further than 1.96 standard deviations from the mean (2.5% in each tail). Similarly, 2.58 is the 1% percentage point. The commonly used percentage points are tabulated in Table A2. Note that they could also be found from Table A1 in the way described above.

The percentage points described here are known as **two-sided** percentage points, as they cover extreme observations in both the upper and lower tails of the distribution. Some tables give **one-sided** percentage points, referring to just one tail of the distribution. The one-sided a% point is the same as the two-sided $2a$%

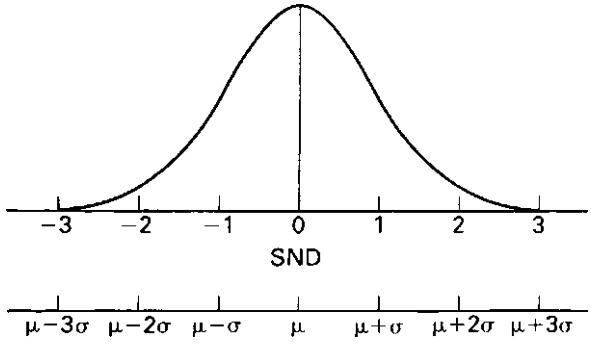
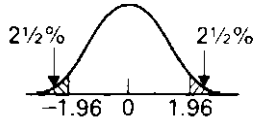
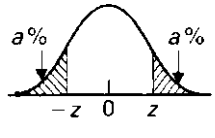


Fig. 5.4 Interpretation of SND in terms of a scale showing the number of standard deviations from the mean.



(a) 1.96 is the one-sided 2½% point or the two-sided 5% point



(b) z is the one-sided a% point or the two-sided $2a$% point

Fig. 5.5 Percentage points of the normal distribution.

point (Figure 5.5b). For example, 1.96 is the one-sided 2.5% point, as 2.5% of the standard normal distribution is above 1.96 (or equivalently 2.5% is below −1.96) and it is the two-sided 5% point. This difference is discussed again in Section 7.3 in the context of hypothesis tests.

These properties mean that, for a normally distributed population, we can derive the range of values within which a given proportion of the population will lie. The 95% **reference range** is given by the mean −1.96 s.d. to mean + 1.96 s.d., since 95% of the values in a population lie in this range. We can also define the 90% reference range and the 99% reference range in the same way, as mean −1.64 s.d. to mean +1.64 s.d. and mean −2.58 s.d. to mean +2.58 s.d., respectively.

5.7 USING *Z*-SCORES TO COMPARE DATA WITH REFERENCE CURVES

SNDs and ***z*-scores** are also used as a way of comparing the values of a variable with those of **reference curves**. The analysis is then carried out using the *z*-scores rather than the original values. For example, this is commonly carried out for anthropometric data, where **growth charts** are used to assess where an individual's weight (or height) lies compared to standard values for their age and sex, and the analysis is in terms of weight-for-age, height-for-age or weight-for-height *z*-scores. This use of *z*-scores is described in Section 13.4, in the chapter on transformations.

CHAPTER 6

Confidence interval for a mean

6.1 INTRODUCTION

In Chapter 4 we explained the idea of sampling variation and the *sampling distribution* of the mean. We showed that the mean of this sampling distribution equals the population mean, μ, and its standard deviation equals $\sigma/\sqrt{n}$, where σ is the population standard deviation, and n is the sample size. We introduced the concept that this standard deviation, which is called the *standard error* of the sample mean, measures how precisely the population mean is estimated by the sample mean. We now describe how we can use the sample mean and its standard error to give us a range of likely values for the population mean, which we wish to estimate.

6.2 LARGE SAMPLE CASE (NORMAL DISTRIBUTION)

In Chapter 4, we stated that approximately 95% of the sample means in the distribution obtained by repeated sampling would lie within two standard errors above or below the population mean. By drawing on the finding presented in Chapter 5, that provided that the sample size is not too small, this sampling distribution is a **normal distribution**, *whether or not* the underlying population distribution is normal, we can now be more precise. We can state that 95% of the sample means would lie within 1.96 standard errors above or below the population mean, since 1.96 is the two-sided 5% point of the standard normal distribution. This means that there is a 95% probability that a particular sample mean ($\bar{x}$) lies within 1.96 standard errors above or below the population mean (μ), which we wish to estimate:

$$\text{Prob}(\bar{x} \text{ is in the range } \mu - 1.96 \times \text{s.e. to } \mu + 1.96 \times \text{s.e.}) = 95\%$$

In practice, this result is used to estimate from the observed sample mean ($\bar{x}$) and its standard error (s.e.) a range within which the population mean is likely to lie. The statement:

'$\bar{x}$ is in the range $\mu - 1.96 \times$ s.e. to $\mu + 1.96 \times$ s.e.'

is equivalent to the statement:

'μ is in the range $\bar{x} - 1.96 \times$ s.e. to $\bar{x} + 1.96 \times$ s.e.'

Therefore there is a 95% probability that the interval between $\bar{x} - 1.96 \times$ s.e. and $\bar{x} + 1.96 \times$ s.e. contains the (unknown) population mean. This interval is called a 95% **confidence interval** (CI) for the population mean, and $\bar{x} - 1.96 \times$ s.e. and $\bar{x} + 1.96 \times$ s.e. are called upper and lower 95% **confidence limits** for the population mean, respectively.

When the sample is large, say n greater than 60, not only is the sampling distribution of sample means well approximated by the normal distribution, but the *sample standard deviation, s, is a reliable estimate of the population standard deviation*, σ, which is usually also not known. The standard error of the sample mean, $\sigma/\sqrt{n}$, can therefore be estimated by $s/\sqrt{n}$.

$$\text{Large-sample 95\% CI} = \bar{x} - (1.96 \times s/\sqrt{n}) \text{ to } \bar{x} + (1.96 \times s/\sqrt{n})$$

Confidence intervals for percentages other than 95% are calculated in the same way using the appropriate percentage point, z', of the standard normal distribution in place of 1.96 (see Chapter 5). For example:

$$\text{Large-sample 90\% CI} = \bar{x} - (1.64 \times s/\sqrt{n}) \text{ to } \bar{x} + (1.64 \times s/\sqrt{n})$$

$$\text{Large-sample 99\% CI} = \bar{x} - (2.58 \times s/\sqrt{n}) \text{ to } \bar{x} + (2.58 \times s/\sqrt{n})$$

Example 6.1

As part of a malaria control programme it was planned to spray all the 10 000 houses in a rural area with insecticide and it was necessary to estimate the amount that would be required. Since it was not feasible to measure all 10 000 houses, a random sample of 100 houses was chosen and the sprayable surface of each of these was measured.

The mean sprayable surface area for these 100 houses was 24.2 m^2 and the standard deviation was 5.9 m^2. It is unlikely that the mean surface area of this sample of 100 houses ($\bar{x}$) exactly equals the mean surface area of all 10 000 houses (μ). Its precision is measured by the standard error $\sigma/\sqrt{n}$, estimated by $s/\sqrt{n} = 5.9/\sqrt{100} = 0.6$ m^2. There is a 95% probability that the sample mean of 24.2 m^2 differs from the population mean by less than 1.96 s.e. $= 1.96 \times 0.6 = 1.2$ m^2. The 95% confidence interval is:

$$95\%\ \text{CI} = \bar{x} - 1.96 \times \text{s.e. to } \bar{x} + 1.96 \times \text{s.e.}$$
$$= 24.2 - 1.2 \text{ to } 24.2 + 1.2 = 23.0 \text{ to } 25.4\ \text{m}^2$$

It was decided to use the upper 95% confidence limit in budgeting for the amount of insecticide required as it was preferable to overestimate rather than underestimate the amount. One litre of insecticide is sufficient to spray 50 m^2 and so the amount budgeted for was:

$$10\,000 \times 25.4/50 = 5080 \text{ litres}$$

There is still a possibility, however, that this is too little insecticide. The interval 23.0 to 25.4 m^2 gives the likely range of values for the mean surface area of all 10 000 houses. There is a 95% probability that this interval contains the population mean but a 5% probability that it does not, with a 2.5% probability ($0.5 \times 5\%$) that the estimate based on the upper confidence limit is too small. A more cautious estimate for the amount of insecticide required would be based on a wider confidence interval, such as 99%, giving a smaller probability (0.5%) that too little would be estimated.

6.3 INTERPRETATION OF CONFIDENCE INTERVALS

We stated in Chapter 2 that our aim in many statistical analyses is to use the sample to make inferences about the population from which it was drawn. Confidence intervals provide us with a means of doing this (see Fig. 6.1).

It is tempting to interpret a 95% CI by saying that 'there is a 95% probability that the population mean lies within the CI'. Formally, this is not quite correct because the population mean (μ) is a fixed unknown number: it is the confidence

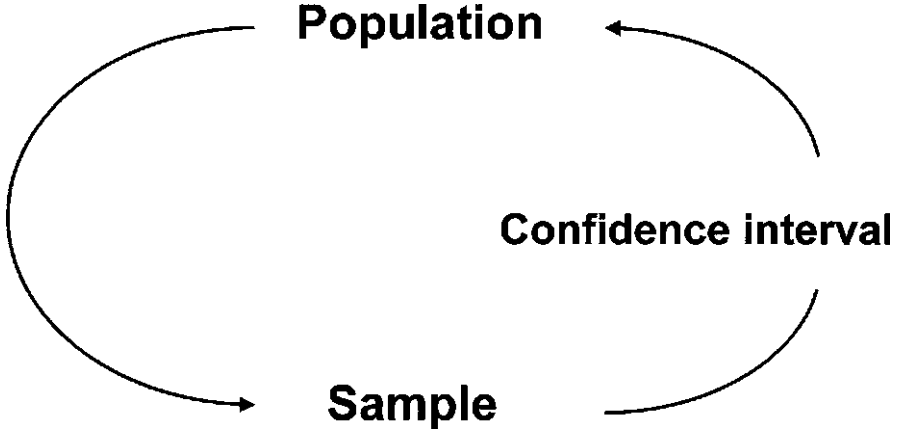


Fig. 6.1 Use of confidence intervals to make inferences about the population from which the sample was drawn.

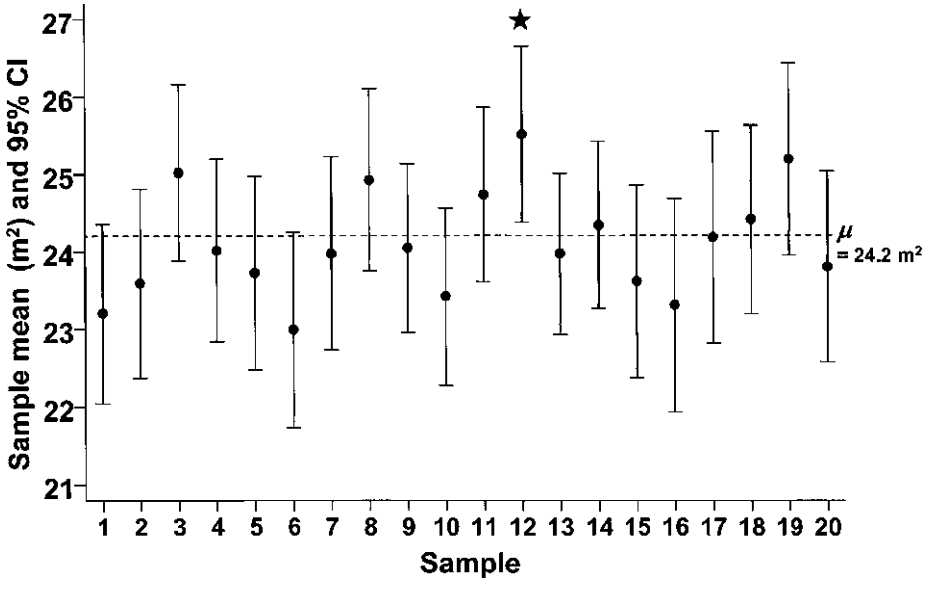


Fig. 6.2 Mean sprayable areas, with 95% confidence intervals, from 20 samples of 100 houses in a rural area. The star indicates that the CI does not contain the population mean.

interval that will vary between samples. In other words, if we were to draw several independent, random samples from the same population and calculate 95% confidence intervals from each of them, then on average 19 of every 20 (95%) such confidence intervals would contain the true population mean, and one of every 20 (5%) would not.

Example 6.2

A further 19 samples, each of 100 houses, were taken from the 10 000 houses described in Example 6.1. The mean sprayable surface and its standard error were calculated from each sample, and these were used to derive 95% confidence intervals. The means and 95% CIs from all 20 samples are shown in Figure 6.2. The mean in the whole population ($\mu = 24.2\,m^2$) is shown by a horizontal dashed line. The sample means vary around the population mean μ, and one of the twenty 95% confidence intervals (indicated by a star) does not contain μ.

6.4 SMALLER SAMPLES

In the calculation of confidence intervals so far described the sample size (n) has been assumed to be large (greater than 60). When the sample size is not large, two aspects may alter:

1 the sample standard deviation, s, which is itself subject to sampling variation, may not be a reliable estimate for σ;

2 when the distribution in the population is not normal, the distribution of the sample mean may also be non-normal.

The second of these effects is of practical importance only when the sample size is very small (less than, say, 15) and when the distribution in the population is extremely non-normal. Because of the central limit theorem (see Chapter 5), it is usually only the first point, the sampling variation in s, which invalidates the use of the normal distribution in the calculation of confidence intervals. Instead, a distribution called the t distribution is used. Strictly speaking, this is valid only if the population is normally distributed, but the use of the t distribution has been shown to be justified, except where the population is extremely non-normal. (This property is called **robustness**.) What to do in cases of severe non-normality is described later in this chapter.

Confidence interval using *t* distribution

The earlier calculation of a confidence interval using the normal distribution was based on the fact that $(\bar{x} - \mu)/(\sigma/\sqrt{n})$ is a value from the standard normal distribution, and that for large samples we could use s in place of σ. In fact, $(\bar{x} - \mu)/(s/\sqrt{n})$ is a value not from the standard normal distribution but from a distribution called the ***t* distribution** with $(n - 1)$ **degrees of freedom**. This distribution was introduced by W. S. Gossett, who used the pen-name 'Student', and is often called Student's t distribution. Like the normal distribution, the t distribution is a symmetrical bell-shaped distribution, but it is more spread out, having longer tails (Figure 6.3).

The exact shape of the t distribution depends on the degrees of freedom (d.f.), $n - 1$, of the sample standard deviation s; the fewer the degrees of freedom, the more the t distribution is spread out. The percentage points are tabulated for various degrees of freedom in Table A3 in the Appendix. For example, if the sample size is 8, the degrees of freedom are 7 and the two-sided 5% point is 2.36. In this case the 95% confidence interval using the sample standard deviation s would be

$$95\%\ \text{CI} = x - 2.36\ s/\sqrt{n} \text{ to } x + 2.36\ s/\sqrt{n}$$

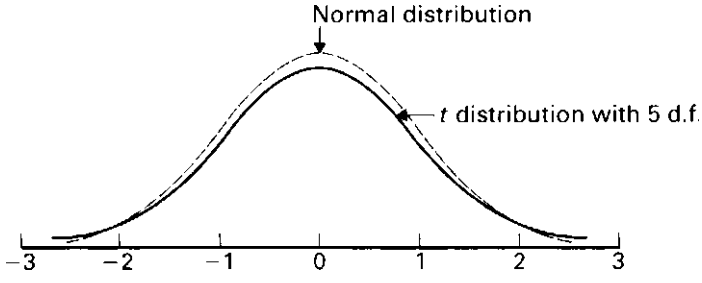


Fig. 6.3 t distribution with 5 degrees of freedom compared to the normal distribution.

In general a confidence interval is calculated using t', the appropriate percentage point of the t distribution with $(n-1)$ degrees of freedom.

$$\text{Small-sample CI} = \bar{x} - (t' \times s/\sqrt{n}) \text{ to } \bar{x} + (t' \times s/\sqrt{n})$$

For small degrees of freedom the percentage points of the t distribution are appreciably larger in value than the corresponding percentage points of the normal distribution. This is because the sample standard deviation s may be a poor estimate of the population value σ, and when this uncertainty is taken into account the resulting confidence interval is considerably wider than if σ were reliably known. For large degrees of freedom the t distribution is almost the same as the standard normal distribution, since s is a good estimate of σ. The bottom row of Table A3 in the Appendix gives the percentage points for the t distribution with an infinite number (∞) of degrees of freedom and it may be seen by comparison with Table A2 that these are the same as for the normal distribution.

Example 6.3

The following are the numbers of hours of relief obtained by six arthritic patients after receiving a new drug:

$$2.2, 2.4, 4.9, 2.5, 3.7, 4.3 \text{ hours}$$
$$\bar{x} = 3.3 \text{ hours}, s = 1.13 \text{ hours}, n = 6, \text{d.f.} = n - 1 = 5$$
$$s/\sqrt{n} = 0.46 \text{ hours}$$

The 5% point of the t distribution with 5 degrees of freedom is 2.57, and so the 95% confidence interval for the average number of hours of relief for arthritic patients in general is:

$$3.3 - 2.57 \times 0.46 \text{ to } 3.3 + 2.57 \times 0.46 = 3.3 - 1.2 \text{ to } 3.3 + 1.2 = 2.1 \text{ to } 4.5 \text{ hours}$$

Severe non-normality

When the distribution in the population is markedly non-normal (see Section 12.2), it may be desirable to **transform** the scale on which the variable x is measured so as to make its distribution on the new scale more normal (see Chapter 13). An alternative is to calculate a **non-parametric** confidence interval or to use **bootstrap** methods (see Chapter 30).

6.5 SUMMARY OF ALTERNATIVES

Table 6.1 summarizes which procedure should be used in constructing a confidence interval. There is no precise boundary between approximate normality and non-normality but, for example, a reverse J-shaped distribution (Fig. 3.6b) is

Table 6.1 Recommended procedures for constructing a confidence interval. (z' is the percentage point from the *normal* distribution, and t' the percentage point from the t distribution with $(n-1)$ degrees of freedom.)

(a) Population standard deviation σ unknown.

	Population distribution	
Sample size	Approximately normal	Severely non-normal*
60 or more	$\bar{x} - (z' \times s/\sqrt{n})$ to $\bar{x} + (z' \times s/\sqrt{n})$	$\bar{x} - (z' \times s/\sqrt{n})$ to $\bar{x} + (z' \times s/\sqrt{n})$
Less than 60	$\bar{x} - (t' \times s/\sqrt{n})$ to $\bar{x} + (t' \times s/\sqrt{n})$	see Chapter 30

(b) Population standard deviation σ known.

	Population distribution	
Sample size	Approximately normal	Severely non-normal*
15 or more	$\bar{x} - (z' \times \sigma/\sqrt{n})$ to $\bar{x} + (z' \times \sigma/\sqrt{n})$	$\bar{x} - (z' \times \sigma/\sqrt{n})$ to $\bar{x} + (z' \times \sigma/\sqrt{n})$
Less than 15	$\bar{x} - (z' \times \sigma/\sqrt{n})$ to $\bar{x} + (z' \times \sigma/\sqrt{n})$	see Chapter 30

*It may be preferable to transform the scale of measurement to make the distribution more normal (see Chapter 13).

severely non-normal, and a skewed distribution (Fig. 3.5b or c) is moderately non-normal.

In rare instances the population standard deviation, σ, is known and therefore not estimated from the sample. When this occurs the standard normal distribution percentage points are used to give the confidence interval regardless of sample size, provided the population distribution is not severely non-normal (in which case see the preceding paragraph).

6.6 CONFIDENCE INTERVALS AND REFERENCE RANGES

It is important to understand the distinction between the **reference range** (which was defined in Section 5.6) and confidence intervals, defined in this chapter. Although they are often confused, each has a different use and a different definition.

A 95% reference range is given by:

$$95\% \text{ reference range} = \mu - 1.96 \times \text{s.d. to } \mu + 1.96 \times \text{s.d.}$$

where μ is the mean of the distribution and s.d. is its standard deviation. A large sample 95% confidence interval is given by:

$$95\% \text{ CI} = \bar{x} - 1.96 \times \text{s.e. to } \bar{x} + 1.96 \times \text{s.e.}$$

where s.e. is the standard error of the distribution: s.e. $= \text{s.d.}/\sqrt{n}$.

The reference range tells us about the variability between individual observations in the population: providing that the distribution is approximately normal

95% of individual observations will lie within the reference range. In contrast, as explained earlier in this chapter, the 95% CI tells us a range of plausible values for the population mean, given the sample mean. Since the sample size n must be > 1, the confidence interval will always be narrower than the reference range.

CHAPTER 7

Comparison of two means: confidence intervals, hypothesis tests and *P*-values

7.1 INTRODUCTION

In Chapter 6 we described how to use a sample mean and its standard error to give us a range of likely values, called a *confidence interval*, for the corresponding population mean. We now extend these ideas to situations where we wish to compare the mean outcomes in two **exposure** (or **treatment**) **groups**. We will label the two groups 0 and 1, and the two means $\bar{x}_0$ and $\bar{x}_1$, with group 1 denoting individuals *exposed* to a risk factor, and group 0 denoting those *unexposed*. In clinical trials, group 1 will denote the *treatment* group and group 0 the *control* group. For example:

- In a study of the determinants of birthweight, we may wish to compare the mean birthweight of children born to smokers (the exposed group, 1) with that for children born to non-smokers (the unexposed group, 0).
- In a clinical trial of a new anti-hypertensive drug, the comparison of interest might be mean systolic blood pressure after 6 months of treatment, between patients allocated to receive the new drug (the treatment group, 1) and those allocated to receive standard therapy (the control group, 0).

The two group means, $\bar{x}_1$ and $\bar{x}_0$, are of interest not in their own right, but for what they tell us more generally about the effect of the exposure on the outcome of interest (or in the case of a clinical trial, of the treatment), in the population from which the groups are drawn. More specifically, we wish to answer the following related questions.

1 What does the *difference* between the two group means in our sample ($\bar{x}_1$ and $\bar{x}_0$) tell us about the difference between the two group means in the population? In other words, what can we say about how much better (or worse) off are exposed individuals compared to unexposed? This is addressed by calculating a

confidence interval for the range of likely values for the difference, following a similar approach to that used for a single mean (see Chapter 6).

2 Do the data provide evidence that the exposure actually affects the outcome, or might the observed difference between the sample means have arisen by chance? In other words, are the data consistent with there being zero difference between the means in the two groups in the population? We address this by carrying out a **hypothesis** (or ***significance*) test** to give a ***P*-value**, which is the probability of recording a difference between the two groups at least as large as that in our sample, if there was no effect of the exposure in the population.

In this chapter we define the sampling distribution of the difference in means comparing the two groups, and then describe how to use this to calculate a confidence interval for the true difference, and how to calculate the test statistic and *P*-value for the related hypothesis test. The methods used are based on either the *normal* or *t* distributions. The rules for which distribution to use are similar to those for the one-sample case. For large samples, or known standard deviations, we use the normal distribution, and for small samples we use the *t* distribution.

The majority of this chapter is concerned with comparing mean outcomes measured in two separate groups of individuals. In some circumstances, however, our data consist instead of *pairs* of outcome measurements. How to compare **paired measurements** is covered in Section 7.6. For example:

- We might wish to carry out a study where the assessment of an anti-hypertensive drug is based on comparing blood pressure measurements in a group of hypertensive men, before and after they received treatment. For each man, we therefore have a pair of outcome measures, blood pressure after treatment and blood pressure before treatment. It is important to take this pairing in the data into account when assessing how much on average the treatment has affected blood pressure.
- Another example would be data from a matched case–control study (see Section 21.4), in which the data consist of case–control pairs rather than of two independent groups of cases and controls, with a control specifically selected to match each case on key variables such as age and sex.

7.2 SAMPLING DISTRIBUTION OF THE DIFFERENCE BETWEEN TWO MEANS

Before we can construct a confidence interval for the difference between two means, or carry out the related hypothesis test, we need to know the sampling distribution of the difference. The difference, $\bar{x}_1 - \bar{x}_0$, between the mean outcomes in the exposed and unexposed groups in our sample provides an estimate of the underlying difference, $\mu_1 - \mu_0$, between the mean outcomes in the exposed and unexposed groups in the population. Just as discussed for a single mean (see Chapter 6), this sample difference will not be exactly equal to the population difference. It is subject to *sampling variation*, so that a different sample from the

same population would give a different value of $\bar{x}_1 - \bar{x}_0$. *Providing that each of the means, $\bar{x}_1$ and $\bar{x}_0$, is normally distributed*, then:

1 the sampling distribution of the difference $(\bar{x}_1 - \bar{x}_0)$ is normally distributed;
2 the mean of this sampling distribution is simply the difference between the two population means, $\mu_1 - \mu_0$;
3 the standard error of $(\bar{x}_1 - \bar{x}_0)$ is based on a combination of the standard errors of the individual means:

$$\text{s.e.} = \sqrt{(s.e._1^2 + s.e._0^2)} = \sqrt{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_0^2}{n_0}\right)}$$

This is estimated using the sample standard deviations, s_1 and s_0. Note that when we calculate the difference between the means in the two groups we *combine* the uncertainty in $\bar{x}_1$ with the uncertainty in $\bar{x}_0$.

7.3 METHODS BASED ON THE NORMAL DISTRIBUTION (LARGE SAMPLES OR KNOWN STANDARD DEVIATIONS)

Confidence interval

When both groups are large (say, greater than 30), or in the rare instances when the population standard deviations are known, then methods for comparing means are based on the normal distribution. We calculate 95% confidence intervals for the difference in the population as:

Large samples

$$\text{CI} = (\bar{x}_1 - \bar{x}_0) - (z' \times \text{s.e.}) \text{ to } (\bar{x}_1 - \bar{x}_0) + (z' \times \text{s.e.})$$

$$\text{s.e.} = \sqrt{(s_1^2/n_1 + s_0^2/n_0)}$$

or

Known σ′s

$$\text{CI} = (\bar{x}_1 - \bar{x}_0) - (z' \times \text{s.e.}) \text{ to } (\bar{x}_1 - \bar{x}_0) + (z' \times \text{s.e.})$$

$$\text{s.e.} = \sqrt{(\sigma_1^2/n_1 + \sigma_0^2/n_0)}$$

In these formulae z' is the appropriate percentage point of the normal distribution. For example, when calculating a 95% confidence interval we use $z' = 1.96$.

Example 7.1

To investigate whether smoking reduces lung function, forced vital capacity (FVC, a test of lung function) was measured in 100 men aged 25–29, of whom 36 were smokers and 64 non-smokers. Results of the study are shown in Table 7.1.

Table 7.1 Results of a study to investigate the association between smoking and lung function.

Group	Number of men	Mean FVC (litres)	s	s.e. of mean FVC
Smokers (1)	$n_1 = 36$	$\bar{x}_1 = 4.7$	$s_1 = 0.6$	s.e.$_1 = 0.6/\sqrt{36} = 0.100$
Non-smokers (0)	$n_0 = 64$	$\bar{x}_0 = 5.0$	$s_0 = 0.6$	s.e.$_0 = 0.6/\sqrt{64} = 0.075$

The mean FVC in smokers was 4.7 litres compared with 5.0 litres in non-smokers. The difference in mean FVC, $\bar{x}_1 - \bar{x}_0$, is therefore $4.7 - 5.0$, that is -0.3 litres. The s.d. in both groups was 0.6 litres. The standard error of the difference in mean FVC is calculated from the individual standard errors, which are shown in the right hand column of the table, as follows:

$$\text{s.e.} = \sqrt{(\text{s.e.}_1^2 + \text{s.e.}_0^2)} = \sqrt{(0.1^2 + 0.075^2)} = 0.125 \text{ litres}$$

The 95% confidence interval for the population difference in mean FVC is therefore:

$$\begin{aligned} 95\%\ \text{CI} &= -0.3 - (1.96 \times 0.125) \text{ to } -0.3 + (1.96 \times 0.125) \\ &= -0.545 \text{ litres to } -0.055 \text{ litres} \end{aligned}$$

Both the lower and upper confidence limits are negative, and both therefore correspond to a reduced FVC among smokers compared to non-smokers. With 95% confidence, the reduction in mean FVC in smokers, compared to non-smokers, lies between 0.055 litres (a relatively small reduction) and 0.545 litres (a reduction likely to have obvious effects).

z-test

The confidence interval gives a range of likely values for the difference in mean outcome between exposed and unexposed groups in the population. With reference to Example 7.1, we now address the related issue of whether the data provide evidence that the exposure (smoking) actually affects the mean outcome (FVC), or whether they are consistent with smoking having no effect. In other words, might the *population* difference between the two groups be zero? We address this issue by carrying out a **hypothesis** (or **significance**) test.

A hypothesis test begins by postulating that, *in the population*, mean FVC is the same in smokers and non-smokers, so that any observed difference between the sample means is due to sampling variation. This is called the **null hypothesis**. The next step is to calculate the probability, *if the null hypothesis were true*, of getting a difference between the two group means as large or larger than the difference than that was observed. This probability is called a ***P*-value**. The idea is that the *smaller* the *P*-value, the *stronger* is the evidence against the null hypothesis.

We use the fact that the sampling distribution of $(\bar{x}_1 - \bar{x}_0)$ is *normal* to derive the *P*-value. If the null hypothesis is true, then the mean of the sampling distribution, $\mu_1 - \mu_0$, is zero. Our **test statistic** is the z-score, or **standard normal deviate** (see Chapter 5) corresponding to the observed difference between the means:

$$z = \frac{\text{difference in means}}{\text{standard error of difference in means}} = \frac{\bar{x}_1 - \bar{x}_0}{\text{s.e.}}$$

The formulae for the z-test are as follows:

Large samples

$$z = \frac{\bar{x}_1 - \bar{x}_0}{\text{s.e.}} = \frac{\bar{x}_1 - \bar{x}_0}{\sqrt{(s_1^2/n_1 + s_0^2/n_0)}}$$

or

Known σ′s

$$z = \frac{\bar{x}_1 - \bar{x}_0}{\text{s.e.}} = \frac{\bar{x}_1 - \bar{x}_0}{\sqrt{(\sigma_1^2/n_1 + \sigma_0^2/n_0)}}$$

The **test statistic** z measures by how many standard errors the mean difference $(\bar{x}_1 - \bar{x}_0)$ differs from the null value of 0. In this example,

$$z = \frac{-0.3}{0.125} = -2.4$$

The difference between the means is therefore 2.4 standard errors below 0, as illustrated in Figure 7.1. The probability of getting a difference of −2.4 standard errors or less (the area under the curve to the left of −2.4) is found using a computer or using Table A1; it is 0.0082. This probability is known as the **one-sided *P*-value**. By convention, we usually use **two-sided *P*-values**; our assessment of the probability that the result is due to chance is based on how extreme the *size* of the departure is from the null hypothesis, and not its direction. We therefore include the probability that the difference might (by chance) have been in the opposite direction: mean FVC might have been greater in smokers than non-smokers. Because the normal distribution is symmetrical, this probability is also 0.0082. The 'two-sided' *P*-value is thus found to be 0.0164 (= 0.0082 + 0.0082), as shown in Figure 7.1.

This means that the probability of observing a difference at least as extreme as 2.4, if the null hypothesis of no difference is correct, is 0.0164, or 1.64%. In other words, if the null hypothesis were true, then sampling variation would yield such a large difference in the mean FVC between smokers and non-smokers in only about 16 in every 1000 similar-sized studies that might be carried out. Such a *P*-value provides evidence *against* the null hypothesis, and suggests that smoking affects FVC.

At this point, you may wish to skip forward to Chapter 8, which gives a fuller description of how to interpret P-values, and how to use P-values and confidence intervals to interpret the results of statistical analyses.

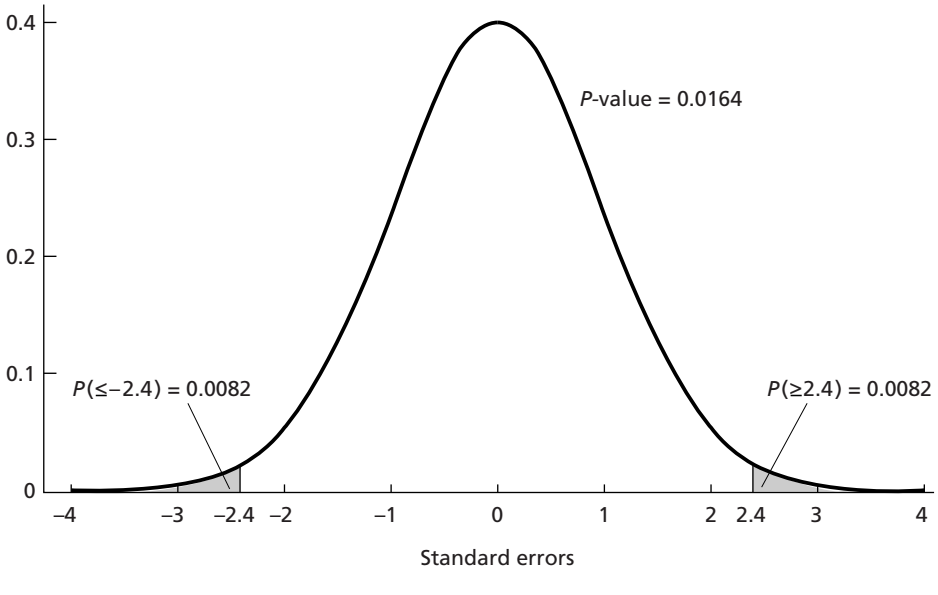


Fig. 7.1 Probability that the size of a standard normal deviate (*z*) is 2.4 standard errors or larger.

7.4 METHODS BASED ON THE *t* DISTRIBUTION (SMALL SAMPLES, EQUAL STANDARD DEVIATIONS)

We saw in Chapter 6 that for small samples we must also allow for the sampling variation in the standard deviation, s, when deriving a confidence interval for a mean. Similar considerations arise when we wish to compare means between small samples. Methods based on the t distribution rather than the normal distribution are used. These require that the population distributions are normal but, as with confidence intervals for a single mean, they are robust against departures from this assumption. When comparing two means, the validity of these methods also depends on the *equality* of the two population standard deviations. In many situations it is reasonable to assume this equality. If the sample standard deviations are very different in size, however, say if one is more than twice as large as the other, then an alternative must be used. This is discussed below in Section 7.5.

Confidence interval

The formula for the standard error of the difference between the means is simplified to:

$$\text{s.e.} = \sqrt{(\sigma^2/n_1 + \sigma^2/n_0)} \text{ or } \sigma\sqrt{(1/n_1 + 1/n_0)}$$

where σ is the common standard deviation. There are two sample estimates of σ from the two samples, s_1 and s_0 and these are combined to give a common estimate, s, of the population standard deviation, with degrees of freedom equal to $(n_1 - 1) + (n_0 - 1) = n_1 + n_0 - 2$.

$$s = \sqrt{\left[\frac{(n_1 - 1)s_1^2 + (n_0 - 1)s_0^2}{(n_1 + n_0 - 2)}\right]}$$

This formula gives greater weight to the estimate from the larger sample as this will be more reliable. The standard error of the difference between the two means is estimated by:

$$\text{s.e.} = s\sqrt{(1/n_1 + 1/n_0)}$$

The confidence interval is calculated using t', the appropriate percentage point of the t distribution with $(n_1 + n_0 - 2)$ degrees of freedom:

$$\text{CI} = (\bar{x}_1 - \bar{x}_0) - (t' \times \text{s.e.}) \text{ to } (\bar{x}_1 - \bar{x}_0) + (t' \times \text{s.e.}),$$
$$\text{d.f.} = (n_1 + n_0 - 2)$$

Example 7.2

Table 7.2 shows the birth weights of children born to 14 heavy smokers (group 1) and to 15 non-smokers (group 0), sampled from live births at a large teaching hospital. The calculations needed to derive the confidence interval are:

difference between the means, $\bar{x}_1 - \bar{x}_0 = 3.1743 - 3.6267 = -0.4524$

standard deviation, $s = \sqrt{\left[\dfrac{13 \times 0.4631^2 + 14 \times 0.3584^2}{15 + 14 - 2}\right]} = 0.4121 \text{ kg}$

standard error of the difference, s.e. $= 0.4121 \times \sqrt{(1/14 + 1/15)} = 0.1531$ kg

degrees of freedom, d.f. $= 14 + 15 - 2 = 27$; $t' = 2.05$

The 5% percentage point of the *t* distribution with 27 degrees of freedom is 2.05, and so the 95% confidence interval for the difference between the mean birth weights is:

$$-0.4524 - (2.05 \times 0.1531) \text{ to } -0.4524 + (2.05 \times 0.1531) = -0.77 \text{ to } -0.14 \text{ kg}$$

Table 7.2 Comparison of birth weights (kg) of children born to 14 heavy smokers with those of children born to 15 non-smokers.

Heavy smokers (group 1)	Non-smokers (group 0)
3.18	3.99
2.74	3.89
2.90	3.60
3.27	3.73
3.65	3.31
3.42	3.70
3.23	4.08
2.86	3.61
3.60	3.83
3.65	3.41
3.69	4.13
3.53	3.36
2.38	3.54
2.34	3.51
	2.71
$\bar{x}_1 = 3.1743$	$\bar{x}_0 = 3.6267$
$s_1 = 0.4631$	$s_0 = 0.3584$
$n_1 = 14$	$n_0 = 15$

With 95% confidence, mean birth weight is between 0.14 and 0.77 kg lower for children born to heavy smokers than for those born to non-smokers.

t test

In small samples we allow for the sampling variation in the standard deviations by using the t distribution for our test of the null hypothesis. This is called a ***t* test**, sometimes also known as an **unpaired *t* test**, to distinguish it from the **paired *t* test** for paired measurements, described in Section 7.6. The t value is calculated as:

$$t = \frac{\bar{x}_1 - \bar{x}_0}{\text{s.e.}} = \frac{\bar{x}_1 - \bar{x}_0}{s\sqrt{(1/n_1 + 1/n_0)}}, \text{ d.f.} = n_1 + n_0 - 2$$

where, as before

$$s = \sqrt{\left[\frac{(n_1 - 1)s_1^2 + (n_0 - 1)s_0^2}{(n_1 + n_0 - 2)}\right]}$$

The corresponding P-value is derived in exactly the same way as for the z distribution. This is best done using a computer, rather than tables, as it is impractical to have sets of tables for all the different possible degrees of freedom. However, an approximate P-value corresponding to different values of the test statistic t may be derived from Table A4 (see Appendix), which tabulates this for a selection of degrees of freedom. It can be seen that unless the number of degrees of freedom is small the P-value based on the normal distribution (right hand column) does not differ greatly from that based on the t distribution (main part of table).

Example 7.2 (continued)

The calculations for the t-test to compare the birth weights of children born to 14 heavy smokers with those of children born to 15 non-smokers, as shown in Table 7.2, are as follows:

$$t = \frac{(3.1743 - 3.6267)}{0.4121\sqrt{(1/14 + 1/15)}} = -\frac{0.4524}{0.1531} = -2.95,$$
$$\text{d.f.} = 14 + 15 - 2 = 27, \; P = 0.0064$$

As the test is two-sided, the *P*-value corresponding to *minus* 2.95 is the same as that corresponding to *plus* 2.95. Table A4 shows that the *P*-value corresponding to $t = 3.0$ with 25 degrees of freedom is 0.006. The *precise* *P*-value of 0.0064 was derived using a computer. As explained in more detail in Chapter 8, a *P*-value of 0.0064 provides fairly strong evidence against the null hypothesis. These data therefore suggest that smoking during pregnancy reduces the birthweight of the baby.

7.5 SMALL SAMPLES, UNEQUAL STANDARD DEVIATIONS

When the population standard deviations of the two groups are different, and the sample size is not large, the main possibilities are:

1. seek a suitable change of scale (a *transformation*, see Chapter 13) which makes the standard deviations similar so that methods based on the *t* distribution can be used. For example, if the standard deviations seem to be proportional in size to the means, then taking logarithms of the individual values may be appropriate;
2. use *non-parametric* methods based on ranks (*see* Section 30.2);
3. use either the Fisher–Behrens or the Welch tests, which allow for unequal standard deviations (consult Armitage & Berry 2002);
4. estimate the difference between the means using the original measurements, but use bootstrap methods to derive confidence intervals (see Section 30.3).

7.6 PAIRED MEASUREMENTS

In some circumstances our data consist of *pairs* of measurements, as described in the introduction to the chapter. These pairs may be two outcomes measured on the same individual under different exposure (or treatment) circumstances. Alternatively, the pairs may be two individuals matched during sample selection to share certain key characteristics such as age and sex, for example in a matched case–control study or in a clinical trial with matched controls (see Chapter 21). Our analysis needs to take this pairing in the data into account: this is done by considering the *differences* between each pair of outcome observations. In other words we turn our data of pairs of outcomes into a single sample of differences.

Confidence interval

The confidence interval for the mean of these differences is calculated using the methods explained for a single mean in Chapter 6, and depending on the sample size uses either the normal or the *t* distribution. In brief, the confidence interval for the difference between the means is:

Large samples (60 or more pairs)

$$\text{CI} = \bar{x}_d - (z' \times \text{s.e.}) \text{ to } \bar{x}_d + (z' \times \text{s.e.})$$

or

Small samples (less than 60 pairs)

$$\text{CI} = \bar{x}_d - (t' \times \text{s.e.}) \text{ to } \bar{x}_d + (t' \times \text{s.e.})$$

where for large samples z' is the chosen percentage point of the normal distribution and for small samples t' is the chosen percentage point of the t distribution with $n - 1$ degrees of freedom. (See Table 6.1 for more details.)

Example 7.3

Consider the results of a clinical trial to test the effectiveness of a sleeping drug in which the sleep of ten patients was observed during one night with the drug and one night with a placebo. The results obtained are shown in Table 7.3. For each patient a pair of sleep times, namely those with the drug and with the placebo, was recorded and the difference between these calculated. The average number of additional hours slept with the drug compared with the placebo was $\bar{x}_d = 1.08$, and the standard deviation *of the differences* was $s = 2.31$ hours. The standard error of the differences is $s/\sqrt{n} = 2.31/\sqrt{10} = 0.73$ hours.

Table 7.3 Results of a placebo-controlled clinical trial to test the effectiveness of a sleeping drug.

	Hours of sleep		
Patient	Drug	Placebo	Difference
1	6.1	5.2	0.9
2	6.0	7.9	−1.9
3	8.2	3.9	4.3
4	7.6	4.7	2.9
5	6.5	5.3	1.2
6	5.4	7.4	−2.0
7	6.9	4.2	2.7
8	6.7	6.1	0.6
9	7.4	3.8	3.6
10	5.8	7.3	−1.5
Mean	$\bar{x}_A = 6.66$	$\bar{x}_B = 5.58$	$\bar{x}_d = 1.08$

Since we have only ten pairs we use the t distribution with 9 degrees of freedom. The 5% point is 2.26, and so the 95% confidence interval is:

$$95\%\ \text{CI} = 1.08 - (2.26 \times 0.73) \text{ to } 1.08 + (2.26 \times 0.73) = -0.57 \text{ to } 2.73 \text{ hours.}$$

With 95% confidence, we therefore estimate the drug to increase average sleeping times by between −0.51 and 2.73 hours. This small study is thus consistent with an effect of the drug which ranges from a small reduction in mean sleep time to a substantial increase in mean sleep time.

Note that the mean of the differences ($\bar{x}x$) is the same as the difference between the means ($\bar{x}x_1 - \bar{x}x_0$). However, the standard error of $\bar{x}x$ will be smaller than the standard error of ($\bar{x}x_1 - \bar{x}x_0$) because we have cancelled out the variation between individuals in their underlying sleep times by calculating *within-person* differences. In other words, we have accounted for the *between-person* variation (see Section 31.4), and so our confidence interval is narrower than if we had used an unpaired design of a similar size.

Hypothesis test

Hypothesis testing of paired means is carried out using either a paired z test or paired t test, depending on the same criteria as laid out for confidence intervals. We calculate the mean of the paired differences, and the test statistic is:

Large sample

$$z = \frac{\bar{x}x}{\text{s.e.}} = \frac{\bar{x}x}{s/\sqrt{n}}$$

or

Small sample

$$t = \frac{\bar{x}x}{\text{s.e.}} = \frac{\bar{x}x}{s/\sqrt{n}},\ \text{d.f.} = n - 1$$

where $\bar{x}x$ is the mean of the paired differences, and n is the number of pairs.

Example 7.3 (continued)

In the above example in Table 7.3 the mean difference in sleep time is 1.08 hours and the standard error is 0.73 hours. A paired t test gives:

$$t = 1.08/0.73 = 1.48,\ \text{d.f.} = 9$$

The probability of getting a t value as large as this in a t distribution with 9 degrees of freedom is 0.17, so there is no evidence against the null hypothesis that the drug does not affect sleep time. This is consistent with the interpretation of

the 95% CI given earlier. An approximate *P*-value can be found from Table A4 (see Appendix), which shows that if the test statistic is 1.5 with 9 degrees of freedom then the *P*-value is 0.168. Further examples of the use of confidence intervals and *P*-values to interpret the results of statistical analyses are given in the next chapter.

CHAPTER 8

Using *P*-values and confidence intervals to interpret the results of statistical analyses

8.1 INTRODUCTION

In Chapter 7 we described how statistical methods may be used to examine the difference between the mean outcome in two exposure groups We saw that we present the results of analyses in two related ways, by reporting a *confidence interval* which gives a range of likely values for the difference in the population, and a *P-value* which addresses whether the observed difference in the sample could arise because of chance alone, if there were no difference in the population.

Throughout this book, we will repeat this process. That is, we will:

1 estimate the magnitude of the difference in disease outcome between exposure groups;
2 derive a confidence interval for the difference; and
3 derive a *P*-value to test the null hypothesis that there is no association between exposure and disease in the population.

In this chapter, we consider how to use *P*-values and confidence intervals to interpret the results of statistical analyses. We discuss hypothesis tests in more detail, explain how to interpret *P*-values and describe some common errors in their interpretation. We conclude by giving examples of the interpretation of the results of different studies.

8.2 TESTING HYPOTHESES

Suppose we believe that everybody who lives to age 90 or more is a non-smoker. We could investigate this hypothesis in two ways:

1 ***Prove the hypothesis*** by finding every single person aged 90 or over and checking that they are all non-smokers.
2 ***Disprove the hypothesis*** by finding just one person aged 90 or over who is a smoker.

In general, it is much easier to find evidence *against* a hypothesis than to be able to prove that it is correct. In fact, one view of science (put forward by the philosopher

Karl Popper) is that it is a process of *disproving* hypotheses. For example, Newton's laws of mechanics were accepted until Einstein showed that there were circumstances in which they did not work.

Statistical methods formalize this idea by looking for evidence against a very specific form of hypothesis, called a **null hypothesis**: that there is *no difference* between groups or *no association* between variables. Relevant data are then collected and assessed for their consistency with the null hypothesis. Links between exposures and outcomes, or between treatments and outcomes, are assessed by examining the strength of the evidence *against* the null hypothesis, as measured by a ***P*-value** (see Section 8.3).

Examples of null hypotheses might be:

- Treatment with beta-interferon has no effect on mean quality of life in patients with multiple sclerosis.
- Performing radical surgery on men aged 55 to 75 diagnosed with prostate cancer does not improve their subsequent mortality.
- Living close to power lines does not affect a child's risk of developing leukaemia.

In some circumstances, statistical methods are not required in order to reject the null hypothesis. For example, before 1990 the most common treatment for stomach ulcers was surgery. A pathologist noticed a particular organism (now known as *Helicobacter pylori*) was often present in biopsy samples taken from stomach ulcers, and grew the organism in culture. He then swallowed a glassful, following which he experienced acute gastritis, and found that the organism progressed to a chronic infection. No statistical analysis of this experiment was necessary to confidently deduce this causal link and reject the *null hypothesis* of no association (B.J. Marshall *et al*. 1985, *Med J Australia* **142**; 436–9), although this was confirmed through antibiotic trials showing that eradicating *H. pylori* cured stomach ulcers.

Similarly, when penicillin was first used as a treatment for pneumonia in the 1940s the results were so dramatic that no formal trial was necessary. Unfortunately such examples, where the results 'hit you straight between the eyes', are rare in medical research. This is because there is rarely such a one-to-one link between exposures and outcomes; there is usually much more inherent *variability* from person to person. Thus although we know that smoking causes lung cancer, we are aware that some heavy smokers will live to an old age, and also that some non-smokers will die prematurely. In other words, smoking increases the risk, but it does not by itself determine death; the outcome is *unpredictable* and is influenced by many other factors.

Statistical methods are used to assess the strength of evidence against a null hypothesis, taking into account this person-to-person variability. Suppose that we want to evaluate whether a new drug reduces cholesterol levels. We might study a group of patients treated with the new drug (the *treatment* group) and a comparable group treated with a *placebo* (the *control* group), and discover that cholesterol levels were on average 5 mg per decilitre lower among patients in the treatment group compared to those in the control group. Before concluding that the drug is

effective, we would need to consider whether this could be a chance finding. We address this question by calculating a *test statistic* and its corresponding *P-value* (also known as a *significance level*). This is the probability of getting a difference of at least 5 mg between the mean cholesterol levels of patients in the treatment and control groups if the drug really has no effect. The *smaller* the *P*-value, the *stronger* the evidence against the null hypothesis that the drug has no effect on cholesterol levels.

8.3 GENERAL FORM OF CONFIDENCE INTERVALS AND TEST STATISTICS

Note that in all cases the **confidence interval** is constructed as the sample estimate (be it a mean, a difference between means or any of the other measures of exposure effect introduced later in the book), plus or minus its standard error multiplied by the appropriate percentage point. Unless the sample size is small, this percentage point is based on the normal distribution (e.g. 1.96 for 95% confidence intervals). The **test statistic** is simply the sample estimate divided by its standard error.

$$95\%\ \text{CI} = \text{estimate} - (1.96 \times \text{s.e.}) \text{ to estimate} + (1.96 \times \text{s.e.})$$

$$\text{Test statistic} = \frac{\text{estimate}}{\text{s.e.}}$$

The standard error is *inversely* related to the sample size. Thus the larger the sample size, the smaller will be the standard error. Since the standard error determines the width of the confidence interval and the size of the test statistic, this also implies the following: for any particular size of difference between the two groups, the *larger* the sample size, the *smaller* will be the confidence interval and the *larger* the test statistic.

The *test statistic* measures by how many standard errors the estimate differs from the null value of zero. As illustrated in Figure 7.1, the test statistic is used to derive a ***P*-value**, which is defined as the probability of getting a difference at least as big as that observed if the null hypothesis is true. By convention, we usually use **two-sided *P*-values**; we include the possibility that the difference could have been of the same size but in the opposite direction. Figure 8.1 gives some examples of how the *P*-value decreases as the test statistic z gets further away from zero. The *larger* the test statistic, the *smaller* is the *P*-value. This can also be seen by examining the one-sided *P*-values (the areas in the upper tail of the standard normal distribution), which are tabulated for different values of z in Table A1 in the Appendix.

Note that we will meet other ways of deriving test statistics later in the book. For example, we introduce chi-squared tests for association in contingency tables

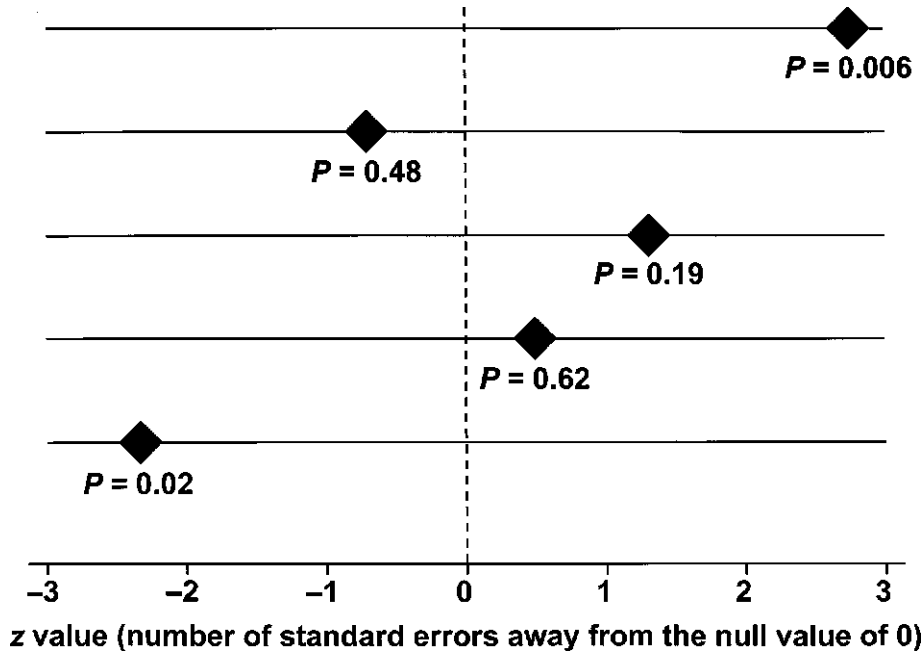


Fig. 8.1 Different *P*-values corresponding to the distance from the null value to the sample mean (expressed as standard errors). Adapted from original by Dr K. Tilling, with thanks.

in Chapter 17, and likelihood ratio tests for testing hypotheses in regression models in Chapters 28 and 29. The interpretation of *P*-values is the same, no matter how they are derived.

8.4 INTERPRETATION OF *P*-VALUES

The smaller the *P*-value, the lower the chance of getting a difference as big as the one observed if the null hypothesis were true. In other words, *the smaller the P-value, the stronger the evidence against the null hypothesis*, as illustrated in Figure 8.2. If the *P*-value is large (more than 0.1, say) then the data do not provide evidence against the null hypothesis, since there is a reasonable chance that the observed difference could simply be the result of sampling variation. If the *P*-value is small (less than 0.001, say) then a difference as big as that observed would be very unlikely to occur if the null hypothesis were true; there is therefore strong evidence against the null hypothesis.

It has been common practice to interpret a *P*-value by examining whether it is smaller than particular threshold values. In particular *P*-values less than 0.05 are often reported as '**statistically significant**' and interpreted as being small enough to justify rejection of the null hypothesis. This is why hypothesis tests have often been called **significance tests**. The 0.05 threshold is an arbitrary one that became commonly used in medical and psychological research, largely because *P*-values

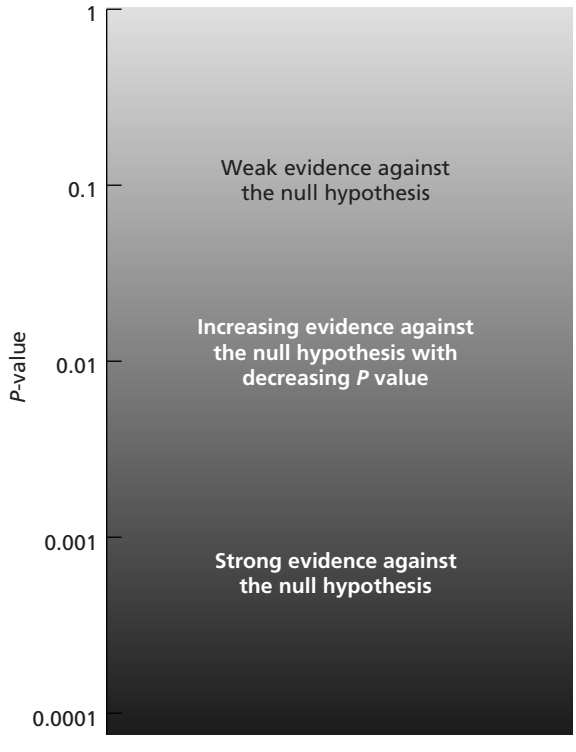


Fig. 8.2 Interpretation of *P*-values.

were determined by comparing the test statistic against tabulations of specific percentage points of distributions such as the z and t distributions, as for example in Table A3 (see Appendix). These days most statistical computer packages will report the precise *P*-value rather than simply whether it is less than 0.05, 0.01, etc. In reporting the results of a study, we recommend this precise *P*-value should be reported together with the 95% confidence interval, and the results of the analyses should be interpreted in the light of both. This is illustrated in Section 8.5.

It should be acknowledged that the 95% confidence level is based on the same arbitrary value as the 0.05 threshold: a z value of 1.96 corresponds to a *P*-value of 0.05. This means that if $P < 0.05$ then the 95% confidence interval will not contain the null value. However, interpretation of a confidence interval should not focus on whether or not it contains the null value, but on the range and potential importance of the different values in the interval.

It is also important to appreciate that the size of the *P*-value depends on the size of the sample, as discussed in more detail in Section 8.5. Three common and serious mistakes in the interpretation of *P*-values are:

1 Potentially medically important differences observed in small studies, for which the *P*-value is more than 0.05, are denoted as non-significant and ignored. To

protect ourselves against this error, we should always consider the range of possible values for the difference shown by the confidence interval, as well as the *P*-value.

2 All statistically significant ($P < 0.05$) findings are assumed to result from real treatment effects, whereas by definition an average of one in 20 comparisons in which the null hypothesis is true will result in $P < 0.05$.

3 All statistically significant ($P < 0.05$) findings are assumed to be of medical importance whereas, given a sufficiently large sample size, even an extremely small difference in the population will be detected as different from the null hypothesis value of zero.

These issues are discussed in the context of examples in the following section and in the context of sample size and power in Chapter 35.

8.5 USING *P*-VALUES AND CONFIDENCE INTERVALS TO INTERPRET THE RESULTS OF A STATISTICAL ANALYSIS

We have now described two different ways of making inferences about differences in mean outcomes between two exposure (or treatment) groups in the target population from the sample results.

1 A confidence interval gives us the range of values within which we are reasonably confident that the population difference lies.

2 The *P*-value tells us the strength of the evidence against the null hypothesis that the true difference in the population is zero.

Since both confidence intervals and *P*-values are derived from the size of the difference and its standard error, they are of course closely related. For example, if the 95% confidence interval does not contain the null value, then we know the *P*-value must be smaller than 0.05. And vice versa; if the 95% confidence interval does include the null value, then the *P*-value will be greater than 0.05. Similarly if the 99% confidence interval does not contain the null value, then the *P*-value is less than 0.01. Because the standard error decreases with increasing sample size, the width of the confidence interval and the size of the *P*-value are as dependent on the sample size as on the underlying population difference. For a particular size of difference in the population, the *larger* the sample size the *narrower* will be the confidence interval, the *larger* the test statistic and the *smaller* the *P*-value.

Both confidence intervals and *P*-values are helpful in interpreting the results of medical research, as shown in Figure 8.3.

Example 8.1

Table 8.1 shows the results of five controlled trials of three different drugs to lower cholesterol levels in middle-aged men and women considered to be at high risk of a heart attack. In each trial patients were randomly assigned to receive either the drug (drug group) or an identical placebo (control group). The number of patients was the same in the treatment and control groups. Drugs A and B are relatively cheap, while drug C is an expensive treatment. In each case cholesterol levels were measured after 1 year, and the mean cholesterol in the control group was

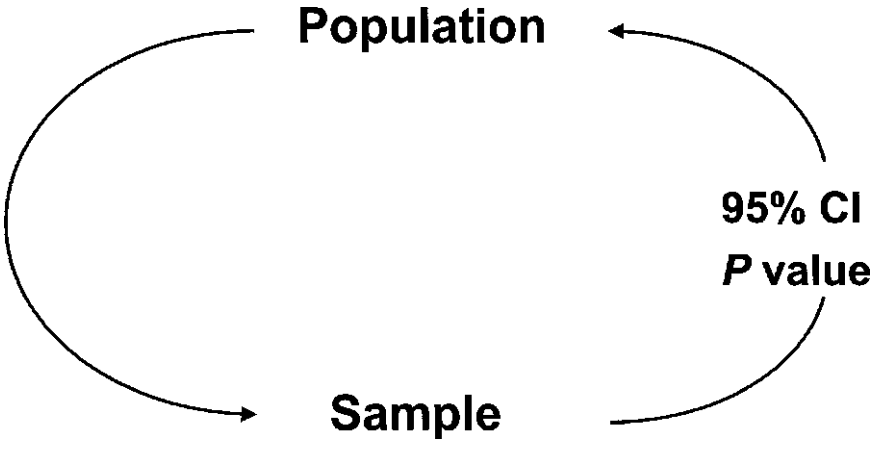


Fig. 8.3 Statistical methods to make inferences about the population from the sample.

Table 8.1 Results of five trials of drugs to lower serum cholesterol.

Trial	Drug	Cost	No. of patients per group	Mean cholesterol (mg/decilitre) in drug group	Mean cholesterol (mg/decilitre) in control group	Reduction (mg/decilitre)
1	A	Cheap	30	140	180	40
2	A	Cheap	3000	140	180	40
3	B	Cheap	40	160	180	20
4	B	Cheap	4000	178	180	2
5	C	Expensive	5000	175	180	5

180 mg/decilitre. The effect of treatment, measured by the difference in the mean cholesterol levels in the drug and control groups, varied markedly between the trials. We will assume that a mean reduction of 40 mg/decilitre confers substantial protection against subsequent heart disease, while a reduction of 20 mg/decilitre confers moderate protection.

What can we infer from these five trials about the effects of the drugs in the population? Table 8.2 shows the effects (measured by the difference in mean

Table 8.2 Results of five trials of drugs to lower serum cholesterol, presented as mean difference (drug group minus control group), s.e. of the difference, 95% confidence interval and *P*-value.

Trial	Drug	Cost	No. of patients per group	Difference in mean cholesterol (mg/decilitre)	s.e. of difference	95% CI for difference	*P*-value
1	A	Cheap	30	−40	40	−118.4 to 38.4	0.32
2	A	Cheap	3000	−40	4	−47.8 to −32.2	< 0.001
3	B	Cheap	40	−20	33	−84.7 to 44.7	0.54
4	B	Cheap	4000	−2	3.3	−8.5 to 4.5	0.54
5	C	Expensive	5000	−5	2	−8.9 to −1.1	0.012

cholesterol between the drug and control groups), together with the standard error of the difference, the 95% confidence interval and the *P*-value.

Note that it is sufficient to display *P*-values accurate to two significant figures (e.g. 0.32 or 0.012). It is common practice to display *P*-values less than 1 in 1000 as '$P < 0.001$' (although other lower limits such as < 0.0001 would be equally acceptable).

- In **trial 1 (drug A)**, mean cholesterol was reduced by 40 mg/decilitre. However, there were only 30 patients in each group. The 95% confidence interval shows us that the results of the trial are consistent with a difference ranging from an *increase* of 38.4 mg/decilitre (corresponding to an adverse effect of the drug) to a very large decrease of 118.4 mg/decilitre. The *P*-value shows that there is no evidence against the null hypothesis of no effect of drug A.
- In **trial 2 (also drug A)**, mean cholesterol was also reduced by 40 mg/decilitre. This trial was much larger, and the *P*-value shows that there was strong evidence against the null hypothesis of no treatment effect. The 95% confidence interval suggests that the effect of drug A in the population is a reduction in mean cholesterol of between 32.2 and 47.8 mg/decilitre. Given that drug A is cheap, this trial strongly suggests that it should be used routinely.

Note that the estimated effect of drug A was the same (a mean reduction of 40 mg/decilitre) in trials 1 and 2. However because trial 1 was small it provided no evidence against the null hypothesis of no treatment effect. This illustrates an extremely important point: in small studies *a large P-value does not mean that the null hypothesis is true*. This is summed up in the phrase '*Absence of evidence is not evidence of absence*'.

Because large studies have a better chance of detecting a given treatment effect than small studies, we say that they are *more powerful*. The concept of power is discussed in more detail in Chapter 35, on choice of sample size.

- In **trial 3 (drug B)**, the reduction in mean cholesterol was 20 mg/decilitre, but because the trial was small the 95% confidence interval is wide (from a reduction of 84.7 mg/decilitre to an increase of 44.7 mg/decilitre). The *P*-value is 0.54: there is no evidence against the null hypothesis that drug B has no effect on cholesterol levels.
- In **trial 4 (also drug B)**, mean cholesterol was reduced by only 2 mg/decilitre. Because the trial was large the 95% confidence interval is narrow (from a reduction of 8.5 mg/decilitre to an increase of 4.5 mg/decilitre). This trial therefore excludes any important effect of drug B. The *P*-value is 0.54: there is no evidence against the null hypothesis that drug B has no effect on cholesterol levels.

Note that there was no effect of drug B in either trial 3 or trial 4, and the *P*-values for the two trials were the same. However, examining the confidence

intervals reveals that they provide very different information about the effect of drug B. Trial 3 (the small trial) is consistent with either a substantial benefit or a substantial harmful effect of drug B while trial 4 (the large trial) *excludes* any substantial effect of drug B (because the lower limit of the confidence interval corresponds to a reduction of only 8.5 mg per decilitre).

- Finally, **trial 5 (drug C)**, was a very large trial in which there was a 5 mg/decilitre reduction in mean cholesterol in the drug group, compared to the control group. The *P*-value shows that there was evidence against the null hypothesis of no effect of drug C. However, the 95% confidence interval suggests that the reduction in mean cholesterol in the population is at most 8.9 mg/decilitre, and may be as little as 1.1 mg/decilitre. Even though we are fairly sure that drug C would reduce cholesterol levels, it is very unlikely that it would be used routinely since it is expensive and the reduction is not of the size required clinically.

Even when the *P*-value shows strong evidence against the null hypothesis, it is vital to examine the confidence interval to ascertain the range of values for the difference between the groups that is consistent with our data. The *medical importance* of the estimated effect should always be considered, even when there is good statistical evidence against the null hypothesis.

For further discussion of these issues see Sterne and Davey Smith (2001), and Chapter 35 on choice of appropriate sample size.

CHAPTER 9

Comparison of means from several groups: analysis of variance

9.1 INTRODUCTION

When our exposure variable has more than two categories, we often wish to compare the mean outcomes from each of the groups defined by these categories. For example, we may wish to examine how haemoglobin measurements collected as part of a community survey vary with age and sex, and to see whether any sex difference is the same for all age groups. We can do this using **analysis of variance**. In general this will be done using a computer package, but we include details of the calculations for the simplest case, that of one-way analysis of variance, as these are helpful in understanding the basis of the methods. Analysis of variance may be seen as a generalization of the methods introduced in Chapters 6 to 8, and is in turn a special case of **multiple regression**, which is described in Chapter 11.

We start with one-way analysis of variance, which is appropriate when the subgroups to be compared are defined by just one exposure, for example in the comparison of means between different socioeconomic or ethnic groups. Two-way analysis of variance is also described and is appropriate when the subdivision is based on two factors such as age and sex. The methods can be extended to the comparison of subgroups cross-classified by more than two factors.

An exposure variable may be chosen for inclusion in an analysis of variance either in order to examine its effect on the outcome, or because it represents a source of variation that it is important to take into account. This is discussed in more detail in the context of multiple regression (Chapter 11).

This chapter may be omitted at a first reading.

9.2 ONE-WAY ANALYSIS OF VARIANCE

One-way analysis of variance is used to compare the mean of a numerical outcome variable in the groups defined by an exposure level with two or more categories. It is called one-way as the exposure groups are classified by just one variable. The method is based on assessing how much of the overall variation in the outcome is attributable to differences between the exposure group means:

hence the name analysis of variance. We will explain this in the context of a specific example.

Example 9.1

Table 9.1(a) shows the mean haemoglobin levels of patients according to type of sickle cell disease. We start by considering the variance of all the observations, ignoring their subdivision into groups. Recall from Chapter 4 that the variance is the square of the standard deviation, and equals the sum of squared deviations of the observations about the overall mean divided by the degrees of freedom:

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

One-way analysis of variance partitions this **sum of squares** (SS $= \Sigma(x - \bar{x})^2$) into two distinct components.

1 The sum of squares due to differences between the group means.
2 The sum of squares due to differences between the observations within each group. This is also called the **residual** sum of squares.

The total degrees of freedom $(n - 1)$ are similarly divided. The between-groups SS has $(k - 1)$ d.f., and the residual SS has $(n - k)$ d.f., where k is the number of groups. The calculations for the sickle cell data are shown in Table 9.1(b) and the results laid out in an analysis of variance table in Table 9.1(c). Note that the subscript i refers to the group number so that n_1, n_2 and n_3 are the number of observations in each of the three groups, $\bar{x}_1$, $\bar{x}_2$ and $\bar{x}_3$ are their mean haemoglobin levels and s_1, s_2, and s_3 their standard deviations. Of the total sum of squares (= 137.85), 99.89 (72.5%) is attributable to between-group variation.

The fourth column of the table gives the amount of variation per degree of freedom, and this is called the **mean square** (MS). The test of the null hypothesis that the mean outcome does not differ between exposure groups is based on a comparison of the between-groups and within-groups mean squares. If the observed differences in mean haemoglobin levels for the different types of sickle cell disease were simply due to chance, the variation between these group means would be about the same size as the variation between individuals with the same type, while if they were real differences the between-groups variation would be larger. The mean squares are compared using the ***F* test**, sometimes called the **variance-ratio** test.

$$F = \frac{\text{Between-groups MS}}{\text{Within-groups MS}}, \quad \begin{aligned} \text{d.f.} &= \text{d.f.}_{\text{Between-groups}}, \ \text{d.f.}_{\text{Within-groups}} \\ &= k - 1, \ n - k \end{aligned}$$

where n is the total number of observations and k is the number of groups.

Table 9.1 One-way analysis of variance: differences in steady-state haemoglobin levels between patients with different types of sickle cell disease. Data from Anionwu *et al*. (1981) *British Medical Journal* **282:** 283–6.

(a) Data.

Type of sickle cell disease	No. of patients (n_i)	Haemoglobin (g/decilitre)		
		Mean ($\bar{x}_i$)	s.d. (s_i)	Individual values (x)
Hb SS	16	8.7125	0.8445	7.2, 7.7, 8.0, 8.1, 8.3, 8.4, 8.4, 8.5, 8.6, 8.7, 9.1, 9.1, 9.1, 9.8, 10.1, 10.3
Hb S/β-thalassaemia	10	10.6300	1.2841	8.1, 9.2, 10.0, 10.4, 10.6, 10.9, 11.1, 11.9, 12.0, 12.1
Hb SC	15	12.3000	0.9419	10.7, 11.3, 11.5, 11.6, 11.7, 11.8, 12.0, 12.1, 12.3, 12.6, 12.6, 13.3, 13.3, 13.8, 13.9

(b) Calculations.

$n = \Sigma n_i = 16 + 10 + 15 = 41$, no. of groups $(k) = 3$

$\Sigma x = 7.2 + 7.7 + \ldots + 13.8 + 13.9 = 430.2$

$\Sigma x^2 = 7.2^2 + 7.7^2 + \ldots + 13.8^2 + 13.9^2 = 4651.80$

Total: $SS = \Sigma(x - \bar{x})^2 = \Sigma x^2 - (\Sigma x)^2/n = 4651.80 - 430.2^2/41 = 137.85$

d.f. $= n - 1 = 40$

Between groups: $SS = \Sigma n_i(\bar{x}_i - \bar{x})^2$, more easily calculated as $\Sigma n_i \bar{x}_i^2 - (\Sigma x)^2/n$

$= 16 \times 8.7125^2 + 10 \times 10.6300^2 + 15 \times 12.3000^2 - 430.2^2/41 = 99.89$

d.f. $= k - 1 = 2$

Within groups: $SS = \Sigma(n_i - 1)s_i^2$

$= 15 \times 0.8445^2 + 9 \times 1.2841^2 + 14 \times 0.9419^2 = 37.96$

d.f. $= n - k = 41 - 3 = 38$

(c) Analysis of variance.

Source of variation	SS	d.f.	MS = SS/d.f.	$F = \frac{\text{Between-groups MS}}{\text{Within-groups MS}}$
Between groups	99.89	2	49.94	49.9, $P < 0.001$
Within groups	37.96	38	1.00	
Total	137.85	40		

F should be about 1 if there are no real differences between the groups and larger than 1 if there are differences. Under the null hypothesis that the between-group differences are simply due to chance, this ratio follows an ***F* distribution** which, in contrast to most distributions, is specified by a pair of degrees of freedom: $(k - 1)$ degrees of freedom in the numerator and $(n - k)$ in the denominator. P-values for the corresponding test of the null hypothesis (that mean haemoglobin levels do not differ according to type of sickle-cell disease) are reported by statistical computer packages.

In Table 9.1(c), $F = 49.94/1.00 = 49.9$ with degrees of freedom (2,38): the corresponding P-value is < 0.001. There is thus strong evidence that mean steady-

state haemoglobin levels differ between patients with different types of sickle cell disease, the mean being lowest for patients with Hb SS disease, intermediate for patients with Hb S/β-thalassaemia, and highest for patients with Hb SC disease.

Assumptions

There are two assumptions underlying the analysis of variance and corresponding *F* test. The first is that the outcome is normally distributed. The second is that the population value for the standard deviation between individuals is the same in each exposure group. This is estimated by the square root of the within-groups mean square. Moderate departures from normality may be safely ignored, but the effect of unequal standard deviations may be serious. In the latter case, transforming the data may help (see Chapter 13).

Relationship with the unpaired *t* test

When there are only two groups, the one-way analysis of variance gives exactly the same results as the *t* test. The *F* statistic (with 1, $n - 2$ degrees of freedom) exactly equals the square of the corresponding *t* statistic (with $n - 2$ degrees of freedom), and the corresponding *P*-values are identical.

9.3 TWO-WAY ANALYSIS OF VARIANCE

Two-way analysis of variance is used when the data are classified in two ways, for example by age-group and sex. The data are said to have a **balanced design** if there are equal numbers of observations in each group and an **unbalanced design** if there are not. Balanced designs are of two types, **with replication** if there is more than one observation in each group and **without replication** if there is only one. Balanced designs were of great importance before the widespread availability of statistical computer packages, because they can be analysed using simple and elegant mathematical formulae. They also allow a division of the sum of squares into different components. However, they are of less importance now that calculations for analysis of variance are done using a computer.

Balanced design with replication

Example 9.2

Table 9.2 shows the results from an experiment in which five male and five female rats of each of three strains were treated with growth hormone. The aims were to find out whether the strains responded to the treatment to the same extent, and whether there was any sex difference. The measure of response was weight gain after seven days.

These data are classified in two ways, by strain and by sex. The design is balanced with replication because there are five observations in each strain–sex

Table 9.2 Differences in response to growth hormone for five male and five female rats from three different strains.

(a) Mean weight gains in grams with standard deviations in parentheses ($n = 5$ for each group).

	Strain		
Sex	A	B	C
Male	11.9 (0.9)	12.1 (0.7)	12.2 (0.7)
Female	12.3 (1.1)	11.8 (0.6)	13.1 (0.9)

(b) Two-way analysis of variance: balanced design with replication.

Source of variation	SS	d.f.	MS	$F = \frac{\text{MS effect}}{\text{MS residual}}$
Main effects				
Strain	2.63	2	1.32	1.9, $P = 0.17$
Sex	1.16	1	1.16	1.7, $P = 0.20$
Interaction				
Strain × sex	1.65	2	0.83	1.2, $P = 0.32$
Residual	16.86	24	0.70	
Total	22.30	29		

group. Two-way analysis of variance divides the total sum of squares into four components:

1 The sum of squares due to *differences between the strains*. This is said to be the **main effect** of the factor, strain. Its associated degrees of freedom are one less than the number of strains and equal 2.
2 The sum of squares due to *differences between the sexes*, that is the main effect of sex. Its degrees of freedom equal 1, one less than the number of sexes.
3 The sum of squares due to the **interaction** between strain and sex. An interaction means that the strain differences are not the same for both sexes and, equivalently, that the sex difference is not the same for the three strains. The degrees of freedom equal the product of the degrees of freedom of the two main effects, which is $2 \times 1 = 2$. The use of regression models to examine interaction between the effects of exposure variables is discussed in Section 29.5.
4 The *residual sum of squares* due to differences between the rats within each strain–sex group. Its degrees of freedom equal 24, the product of the number of strains (3), the number of sexes (2) and one less than the number of observations in each group (4).

The null hypotheses of no main effect of the two exposures and of no interaction are examined by using the F test to compare their mean squares with the residual mean square, as described for one-way analysis of variance. No evidence of any association was obtained in this experiment.

Balanced design without replication

In a balanced design without replication there is no residual sum of squares in the analysis of variance, since there is only one observation in each cell of the table showing the cross-classification of the two exposures. In such a case, it is assumed that there is no interaction between the effects of the two exposures, and the interaction mean square is used as an estimate of the residual mean square for calculating F statistics for the main effects. The two-way analysis of variance for a balanced design without replication is an extension of the **paired *t* test**, comparing the values of more than two variables measured on each individual. The two approaches give the same results when just two variables are measured, and the F value equals the square of the t value.

Unbalanced design

When the numbers of observations in each cell are not equal the design is said to be unbalanced. The main consequence, apart from the additional complexity of the calculations, is that it is not possible to disentangle the effects of the two exposures on the outcome. Instead, the *additional* sum of squares due to the effect of one variable, allowing for the effect of the other, may be calculated. These issues are explained in more detail in Chapter 11, which describes multiple linear regression.

Unbalanced data are common, and unavoidable, in survey investigations. The interpretation of clinical trials and laboratory experiments will be simplified if they have a balanced design, but even when a balanced design is planned this will not always succeed as, for example, people may withdraw or move out of the area half-way through a trial, or animals may die during the course of an experiment.

9.4 FIXED AND RANDOM EFFECTS

The effect of exposures can be defined in two ways, as **fixed effects** or as **random effects**. Factors such as sex, age-group and type of sickle cell disease are all *fixed effects* since their individual levels have specific values; sex is always male or female. In contrast, the individual levels of a *random* effect are not of intrinsic interest but are a sample of levels representative of a source of variation. For example, consider a study to investigate the variation in sodium and sucrose concentrations of home-prepared oral rehydration solutions, in which ten persons were each asked to prepare eight solutions. In this case, the ten persons are of interest only as representatives of the variation between solutions prepared by different persons. Persons is then a *random* effect. The method of analysis is the same for fixed and random effects in one-way designs and in two-way designs without replication, but not in two-way designs with replication (or in higher level designs). In the latter, if both effects are fixed, their mean squares are compared with the residual mean square as described above. If, on the other hand, both

effects are random, their mean squares are compared with the interaction rather than the residual mean square. If one effect is random and the other fixed, it is the other way round; the random effect mean square is compared with the residual mean square, and the fixed effect mean square with the interaction. Analyses with random effects are described in more detail in Chapter 31.

CHAPTER 10

Linear regression and correlation

10.1 INTRODUCTION

Previous chapters have concentrated on the association between a numerical outcome variable and a categorical exposure variable with two or more levels. We now turn to the relationship between a numerical outcome and a *numerical* exposure. The method of linear regression is used to estimate the best-fitting straight line to describe the association. The method also provides an estimate of the correlation coefficient, which measures the closeness (strength) of the linear association. In this chapter we consider *simple* **linear regression** in which only one exposure variable is considered. In the next chapter we introduce *multiple regression* models for the effect of more than one exposure on a numerical outcome.

10.2 LINEAR REGRESSION

Example 10.1

Table 10.1 shows the body weight and plasma volume of eight healthy men. A **scatter plot** of these data (Figure 10.1) shows that high plasma volume tends to be

Table 10.1 Plasma volume, and body weight in eight healthy men. Sample size $n = 8$, mean body weight $\bar{x}x = 66.875$, mean plasma volume $\bar{y}y = 3.0025$.

Subject	Body weight (kg)	Plasma volume (litres)
1	58.0	2.75
2	70.0	2.86
3	74.0	3.37
4	63.5	2.76
5	62.0	2.62
6	70.5	3.49
7	71.0	3.05
8	66.0	3.12

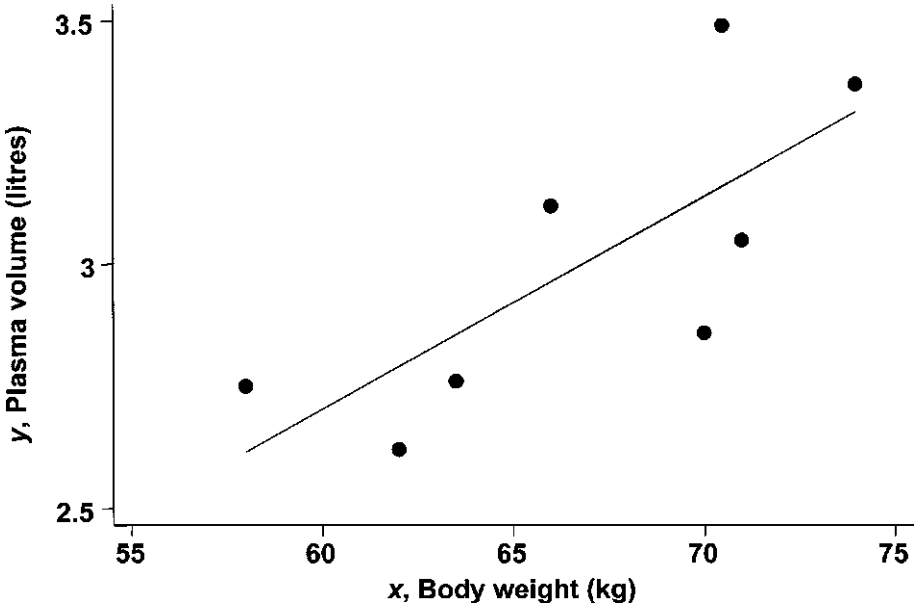


Fig. 10.1 Scatter diagram of plasma volume and body weight showing the best-fitting linear regression line.

associated with high weight and vice versa. Note that it is conventional to plot the exposure on the horizontal axis and the outcome on the vertical axis. In this example, it is obviously the dependence of plasma volume on body weight that is of interest, so plasma volume is the outcome variable and body weight is the exposure variable. Linear regression gives the equation of the straight line that best describes how the outcome y increases (or decreases) with an increase in the exposure variable x. The equation of the **regression line** is:

$$y = \beta_0 + \beta_1 x$$

where β is the Greek letter beta. We say that β_0 and β_1 are the **parameters** or **regression coefficients** of the linear regression: β_0 is the **intercept** (the value of y when $x = 0$), and β_1 the **slope** of the line (the increase in y for every unit increase in x; see Figure 10.2).

Estimation of the regression parameters

The best-fitting line is derived using the method of **least squares**: by finding the values for the parameters β_0 and β_1 that minimize the sum of the squared vertical distances of the points from the line (Figure 10.3). The parameters β_0 and β_1 are are estimated using the following formulae:

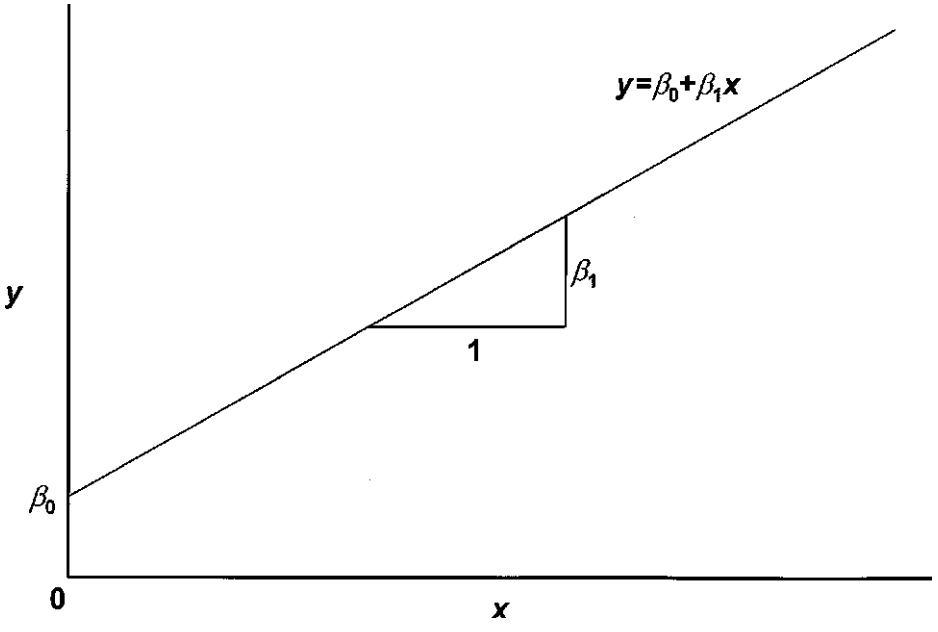


Fig. 10.2 The intercept and slope of the regression equation, $y = \beta_0 + \beta_1 x$. The intercept, β_0, is the point where the line crosses the y axis and gives the value of y for $x = 0$. The slope, β_1, is the increase in y corresponding to a unit increase in x.

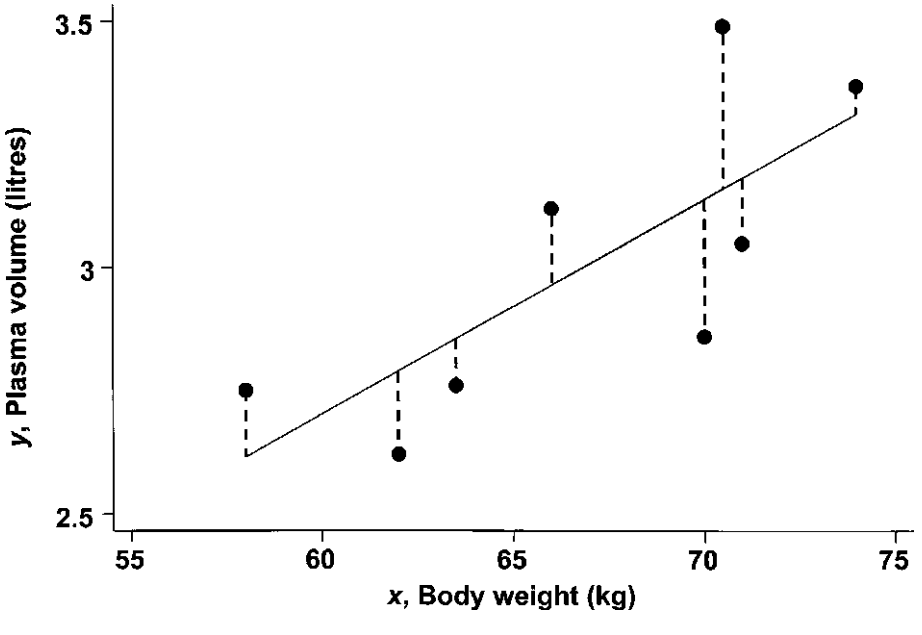


Fig. 10.3 Linear regression line, $y = \beta_0 + \beta_1 x$, fitted by least squares. β_0 and β_1 are calculated to minimize the sum of squares of the vertical deviations (shown by the dashed lines) of the points about the line; each deviation equals the difference between the observed value of y and the corresponding point on the line, $\beta_0 + \beta_1 x$.

$$\beta_1 = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(x - \bar{x})^2} \quad \text{and} \quad \beta_0 = \bar{y} - \beta_1 \bar{x}$$

Regression coefficients are sometimes known as 'beta-coefficients', and are labelled in this way by some statistical software packages. When the slope $\beta_1 = 0$ this corresponds to a horizontal line at a height of $\bar{y}$ and means that there is no association between x and y.

In this example:

$$\Sigma(x - \bar{x})(y - \bar{y}) = 8.96 \quad \text{and} \quad \Sigma(x - \bar{x})^2 = 205.38$$

So:

$$\beta_1 = 8.96/205.38 = 0.043615$$

and:

$$\beta_0 = 3.0025 - 0.043615 \times 66.875 = 0.0857$$

Thus the best-fitting straight line describing the association of plasma volume with body weight is:

$$\text{Plasma volume} = 0.0857 + 0.0436 \times \text{weight}$$

which is shown in Figures 10.1 and 10.3.

The regression line is drawn by calculating the co-ordinates of two points which lie on it. For example:

$$x = 60, \; y = 0.0857 + 0.0436 \times 60 = 2.7$$

and

$$x = 70, \; y = 0.0857 + 0.0436 \times 70 = 3.1$$

As a check, the line should pass through the point $(\bar{x}, \bar{y}) = (66.9, \; 3.0)$. Statistical software packages will usually allow the user to include the regression line in scatter plots.

The calculated values for β_0 and β_1 are estimates of the population values of the intercept and slope and are, therefore, subject to sampling variation. As with estimated differences between exposure group means (see Chapter 7) their precision is measured by their standard errors.

$$\text{s.e.}(\beta_0) = s\sqrt{\left[\frac{1}{n} + \frac{\bar{x}^2}{\Sigma(x-\bar{x})^2}\right]} \quad \text{and} \quad \text{s.e.}(\beta_1) = \frac{s}{\sqrt{\Sigma(x-\bar{x})^2}}$$

$$s = \sqrt{\left[\frac{\Sigma(y-\bar{y})^2 - \beta_1^2\Sigma(x-\bar{x})^2}{(n-2)}\right]}$$

s is the **standard deviation of the points about the line**. It has $(n-2)$ degrees of freedom (the sample size minus the number of regression coefficients). In this example $\Sigma(y-\bar{y})^2 = 0.6780$ and so:

$$s = \sqrt{\frac{0.6780 - 0.0436^2 \times 205.38}{6}} = 0.2189$$

$$\text{s.e.}(\beta_0) = 0.2189\sqrt{\left[\frac{1}{8} + \frac{66.9^2}{205.38}\right]} = 1.0237$$

and

$$\text{s.e.}(\beta_1) = \frac{0.2189}{\sqrt{205.38}} = 0.0153$$

Computer output

Linear regression models are usually estimated using a statistical computer package. Table 10.2 shows typical output; for our example, *plasvol* and *weight* were the names of the outcome and exposure variables respectively in the computer file. The output should be interpreted as follows.

1. The regression coefficient for *weight* is the same as the estimate of β_1 calculated earlier while the regression coefficient labelled 'Constant' corresponds to the estimate of the intercept (β_0).

   Note that in this example the intercept is not a meaningful number: its literal interpretation is as the estimated mean plasma volume when weight = 0. The intercept can be made meaningful by **centring** the exposure variable: subtracting its mean so that the new exposure variable has mean = 0. The intercept in a linear regression with a centred exposure variable is equal to the mean outcome.
2. The standard errors also agree with those calculated above.
3. The t statistics in the fourth column are the values of each regression coefficient divided by its standard error. Each t statistic may be used to test the null hypothesis that the corresponding regression coefficient is equal to zero. The degrees

Table 10.2 Computer output for the linear regression of plasma volume on body weight (data in Table 10.1).

Plasvol	Coefficient	Std err	t	$P > \|t\|$	95% CI
Weight	0.0436	0.0153	2.857	0.029	0.0063 to 0.0810
Constant	0.0857	1.024	0.084	0.936	−2.420 to 2.591

of freedom are the sample size minus the number of regression coefficients, $n - 2$. The corresponding P-values are in the fifth column. In this example, the P-value for *weight* is 0.029: there is some evidence against the null hypothesis that there is no association between body weight and plasma volume. The P-value for the intercept tests the null hypothesis that the intercept is equal to zero: this is not usually an interesting null hypothesis but is reported because computer packages tend to present their output in a uniform manner.

4 The 95% confidence intervals are calculated as:

$$\text{CI} = \text{regression coefficient} - t' \times \text{s.e. to regression coefficient} + t' \times \text{s.e.}$$

where t' is the relevant percentage point of the t distribution with $n - 2$ degrees of freedom. In this example the 5% point of the t distribution with 6 d.f. is 2.45, and so (for example) the lower limit of the 95% CI for β_1 is $0.0436 - 2.45 \times 0.0153 = 0.0063$. In large samples the 5% point of the normal distribution (1.96) is used (d.f. $= \infty$ in Table A3, Appendix).

Assumptions

There are two assumptions underlying linear regression. The first is that, for any value of x, y is normally distributed. The second is that the magnitude of the scatter of the points about the line is the same throughout the length of the line. This scatter is measured by the standard deviation, s, of the points about the line as defined above. More formally, we assume that:

$$y = \beta_0 + \beta_1 x + e$$

where the **error**, e, is normally distributed with mean zero and standard deviation σ, which is estimated by s (the standard deviation of the points about the line). The vertical deviations (shown by the dotted lines) in Figure 10.3 are the estimated errors, known as **residuals**, for each pair of observations.

A change of scale may be appropriate if either of the two assumptions does not hold, or if the relationship seems non-linear (see Sections 11.5 and 29.6). It is important to examine the scatter plot to check that the association is approximately

linear *before* proceeding to fit a linear regression. Ways to check the assumptions made in a linear regression are discussed in more detail in Section 12.3.

Prediction

In some situations it may be useful to use the regression equation to predict the value of y for a particular value of x, say x'. The **predicted value** is:

$$y' = \beta_0 + \beta_1 x'$$

and its standard error is:

$$\text{s.e.}(y') = s\sqrt{\left[1 + \frac{1}{n} + \frac{(x' - \bar{x})^2}{\Sigma(x - \bar{x})^2}\right]}$$

This standard error is least when x' is close to the mean, $\bar{x}$In general, one should be reluctant to use the regression line for predicting values outside the range of x in the original data, as the linear relationship will not necessarily hold true beyond the range over which it has been fitted.

Example 10.1 (continued)
In this example, the measurement of plasma volume is time-consuming and so, in some circumstances, it may be convenient to predict it from the body weight. For instance, the predicted plasma volume for a man weighing 66 kg is:

$$0.0832 + 0.0436 \times 66 = 2.96 \text{ litres}$$

and its standard error equals:

$$0.2189\sqrt{\left[1 + \frac{1}{8} + \frac{(66 - 66.9)^2}{205.38}\right]} = 0.23 \text{ litres}$$

10.3 CORRELATION

As well as estimating the best-fitting straight line we may wish to examine the strength of the linear association between the outcome and exposure variables. This is measured by the **correlation coefficient**, r, which is estimated as:

$$r = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\sqrt{\left[\Sigma(x - \bar{x})^2 \Sigma(y - \bar{y})^2\right]}}$$

where x denotes the exposure, y denotes the outcome, and $\bar{x}x$and $\bar{y}y$are the corresponding means. Scatter plots illustrating different values of the correlation coefficient are shown in Figure 10.4. The correlation coefficient is always a number between -1 and $+1$, and equals zero if the variables are not associated. It is positive if x and y tend to be high or low together, and the larger its value the closer the association. The maximum value of 1 is obtained if the points in the scatter plot lie exactly on a straight line. Conversely, the correlation coefficient is negative if high values of y tend to go with low values of x, and vice versa. The correlation coefficient has the same sign as the regression coefficient β_1. When there is no correlation β_1 equals zero, corresponding to a horizontal regression line at height $\bar{y}y$(no association between x and y).

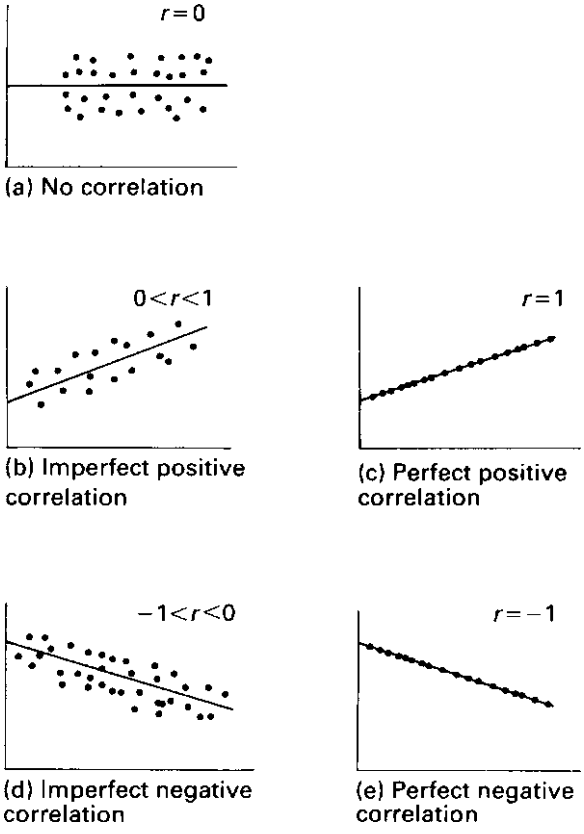


Fig. 10.4 Scatter plots illustrating different values of the correlation coefficient. Also shown are the regression lines.

Example 10.1 (continued)

In this example:

$$r = \frac{8.96}{\sqrt{(205.38 \times 0.6780)}} = 0.7591$$

Table 10.3 Computer output for the linear regression of the derived variable *stdplasvol* on *stdweight* (plasma volume and body weight divided by their standard deviations).

stdplasvol	Coefficient	Std err	t	$P > t$	95% CI
stdweight	0.7591	0.2657	2.86	0.029	0.1089 to 1.4094
Constant	0.2755	3.2904	0.08	0.936	−7.7759 to 8.3268

A useful interpretation of the correlation coefficient is that it is the *number of standard deviations that the outcome y changes for a standard deviation change in the exposure x*. In larger studies (sample size more than about 100), this provides a simple way to derive a confidence interval for the correlation coefficient, using standard linear regression. In this example, the standard deviation of body weight was 5.42 kg, and the standard deviation of plasma volume was 0.31 litres. If we divide each variable by its standard deviation we can create new variables, each of which has a standard deviation of 1. We will call these variables *stdplasvol* and *stdweight*: a change of 1 in these variables therefore corresponds to a change of one standard deviation in the original variables. Table 10.3 shows computer output from the regression of *stdplasvol* on *stdweight*. The regression coefficient for *stdweight* is precisely the same as the *correlation* coefficient calculated earlier. Note also that the *P*-values are identical to those in Table 10.2: the null hypothesis that the correlation $r = 0$ is identical to the null hypothesis that the regression coefficient $\beta_1 = 0$.

For large samples the confidence interval corresponding to the regression coefficient for the modified exposure variable (*stdweight* in Table 10.3) may be interpreted as a confidence interval for the correlation coefficient. In this very small study, however, the upper limit of the 95% CI is 1.4094, whereas the maximum possible value of the correlation is 1. For studies whose sample size is less than about 100, confidence intervals for the correlation coefficient can be derived using **Fisher's transformation**:

$$z_r = \frac{1}{2}\log_e\left(\frac{1+r}{1-r}\right)$$

See Section 13.2 for an explanation of logarithms and the exponential function. The standard error of the transformed correlation z_r is approximately $1/\sqrt{(n-3)}$, and so a 95% confidence interval for z_r is:

$$95\%\ \text{CI} = z_r - 1.96/\sqrt{(n-3)}\ \text{to}\ z_r + 1.96/\sqrt{(n-3)}$$

This can then be transformed back to give a confidence interval for r using the inverse of Fisher's transformation:

$$r = \frac{\exp(2z_r) - 1}{\exp(2z_r) + 1}$$

In this example, the transformed correlation between weight and plasma volume is $z_r = 0.5 \log_e (1.7591/0.2409) = 0.9941$. The standard error of z_r is $1/\sqrt{(8–3)} = 0.4472$. The 95% CI for z_r is:

$$\begin{aligned} 95\%\ \text{CI for } z_r &= 0.9941 - 1.96 \times 0.4472 \text{ to } 0.9941 + 1.96 \times 0.4472 \\ &= 0.1176 \text{ to } 1.8706 \end{aligned}$$

Applying the inverse of Fisher's transformation to the upper and lower confidence limits gives a 95% CI for the correlation:

$$95\%\ \text{CI for } r = 0.1171 \text{ to } 0.9536$$

10.4 ANALYSIS OF VARIANCE APPROACH TO SIMPLE LINEAR REGRESSION

We stated earlier that the regression coefficients β_0 and β_1 are calculated so as to minimize the sum of squared deviations of the points about the regression line. This can be compared to the overall variation in the outcome variable, measured by the **total sum of squares**

$$SS_{Total} = \Sigma(y - \bar{y})^2$$

This is illustrated in Figure 10.5 where the deviations about the line are shown by the dashed vertical lines and the deviations about the mean, $(y - \bar{y})$, are shown by the solid vertical lines. The sum of squared deviations about the best-fitting regression line is called the **residual sum of squares** ($SS_{Residual}$). This is less than SS_{Total} by an amount which is called the sum of squares *explained by the regression* of plasma volume on body weight, or simply the **regression sum of squares**

$$SS_{Regression} = SS_{Total} - SS_{Residual}$$

This splitting of the overall variation into two parts can be laid out in an analysis of variance table (see Chapter 9).

Example 10.1 (continued)

The analysis of variance results for the linear regression of plasma volume on body weight are presented in Table 10.4. There is 1 degree of freedom for the regression and $n - 2 = 6$ degrees of freedom for the residual.

If there were no association between the variables, then the regression mean square would be about the same size as the residual mean square, while if the variables were associated it would be larger. This is tested using an F test, with degrees

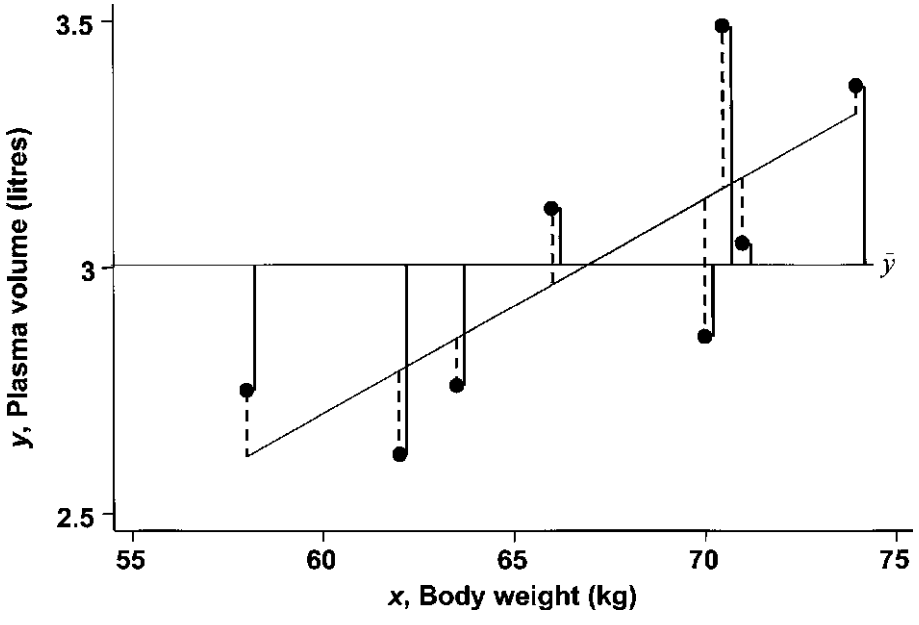


Fig. 10.5 Deviations in the outcome y about the regression line (dashed vertical lines) and about the mean $\bar{y}y$(solid vertical lines).

Table 10.4 Analysis of variance for the linear regression of plasma volume on body weight ($n = 8$).

Source of variation	Sum of squares (SS)	Degrees of freedom (d.f.)	Mean square (MS = SS/d.f.)	$F = \frac{\text{MS regression}}{\text{MS residual}}$
Regression	0.3907	1	0.3907	8.16, $P = 0.029$
Residual	0.2873	6	0.0479	
Total	0.6780	7	0.0969	

of freedom (1, $n - 2$), as described in Chapter 9. The resulting P-value is identical to that from the t statistic in the linear regression output presented in Table 10.2.

10.5 RELATIONSHIP BETWEEN CORRELATION COEFFICIENT AND ANALYSIS OF VARIANCE TABLE

The analysis of variance table gives an alternative interpretation of the correlation coefficient. The square of the correlation coefficient, r^2, equals the regression sum of squares divided by the total sum of squares ($0.76^2 = 0.5763 = 0.3907/0.6780$). It is thus the *proportion of the total variation in plasma volume that has been explained by the regression*. In Example 10.1, we can say that body weight accounts for 57.63% of the total variation in plasma volume.

CHAPTER 11

Multiple regression

11.1 INTRODUCTION

Situations frequently occur in which we wish to examine the dependency of a numerical outcome variable on *several* exposure variables, not just one. This is done using **multiple linear regression**, a generalization of the methods for linear regression that were introduced in Chapter 10.

In general, there are two reasons for including extra exposure variables in a multiple regression analysis. The first is to estimate an exposure effect after allowing for the effects of other variables. For example, if two exposure groups differed in respect to other factors, such as age, sex, socioeconomic status, which were known to affect the outcome of interest, then it would be important to adjust for these differences before attributing any difference in outcome between the exposure groups to the exposure. This is described in Section 11.2 below, and is an example of the control of **confounding** factors, explained in more detail in Chapter 18. The second reason is that inclusion of exposure variables that are strongly associated with the outcome variable will reduce the residual variation and hence decrease the standard error of the regression coefficients for other exposure variables. This means that it will increase both the accuracy of the estimation of the other regression coefficients, and the likelihood that the related hypothesis tests will detect any real effects that exist. This latter attribute is called the power of the test and is described in detail in Chapter 35 ('Calculation of required sample size'). This second reason applies only when the outcome variable is numerical (and not, for example, when we use logistic regression to analyse the association of one or more exposure variables with a binary outcome variable, see Chapters 19 and 20).

Multiple regression can be carried out with any number of variables, although it is recommended that the number be kept reasonably small, as with larger numbers

the interpretation becomes increasingly more complex. These issues are discussed in more detail in the chapters on regression modelling (Chapter 29) and strategies for analysis (Chapter 38).

11.2 MULTIPLE REGRESSION WITH TWO EXPOSURE VARIABLES

Example 11.1

All the methods will be illustrated using a study of lung function among 636 children aged 7 to 10 years living in a deprived suburb of Lima, Peru. The maximum volume of air that the children could breathe out in 1 second (Forced Expiratory Volume in 1 second, denoted as FEV_1) was measured using a spirometer. The age and height of the children were recorded, and their carers were asked about respiratory symptoms that the children had experienced in the last year.

Consider first the relationship of lung function (FEV_1) with the two exposure variables: age and height of the child. It seems likely that FEV_1 will increase with both height and age, and this is confirmed by scatter plots, which suggest that the relationship of FEV_1 with each of these is linear (Figure 11.1). The output from separate linear regression models for the association between FEV_1 and each of these two exposure variables is shown in Table 11.1.

As is apparent from the scatter plots, there is a strong association between FEV_1 and both age and height. The regression coefficients tell us that FEV_1 increases by 0.2185 litres for every year of age, and by 0.0311 litres for every centimetre of height. The regression lines are shown on the scatter plots in Figure 11.1. The correlations of FEV_1 with age and height are 0.5161 and 0.6376, respectively.

As might be expected, there is also a strong association between age and height (correlation = 0.5946). We may therefore ask the following questions:

- what is the association between age and FEV_1, having taken the association between height and FEV_1 into account?
- what is the association between height and FEV_1, having taken the association between age and FEV_1 into account?

Table 11.1 Computer output for two separate linear regression models for FEV_1.

(a) FEV_1 and age.

FEV_1	Coefficient	Std err	t	$P > \|t\|$	95% CI
Age	0.2185	0.0144	15.174	0.000	0.1902 to 0.2467
Constant	−0.3679	0.1298	−2.835	0.005	−0.6227 to −0.1131

(b) FEV_1 and height.

FEV_1	Coefficient	Std err	t	$P > \|t\|$	95% CI
Height	0.0311	0.00149	20.840	0.000	0.0282 to 0.0341
Constant	−2.2658	0.1855	−12.216	0.000	−2.6300 to −1.9016

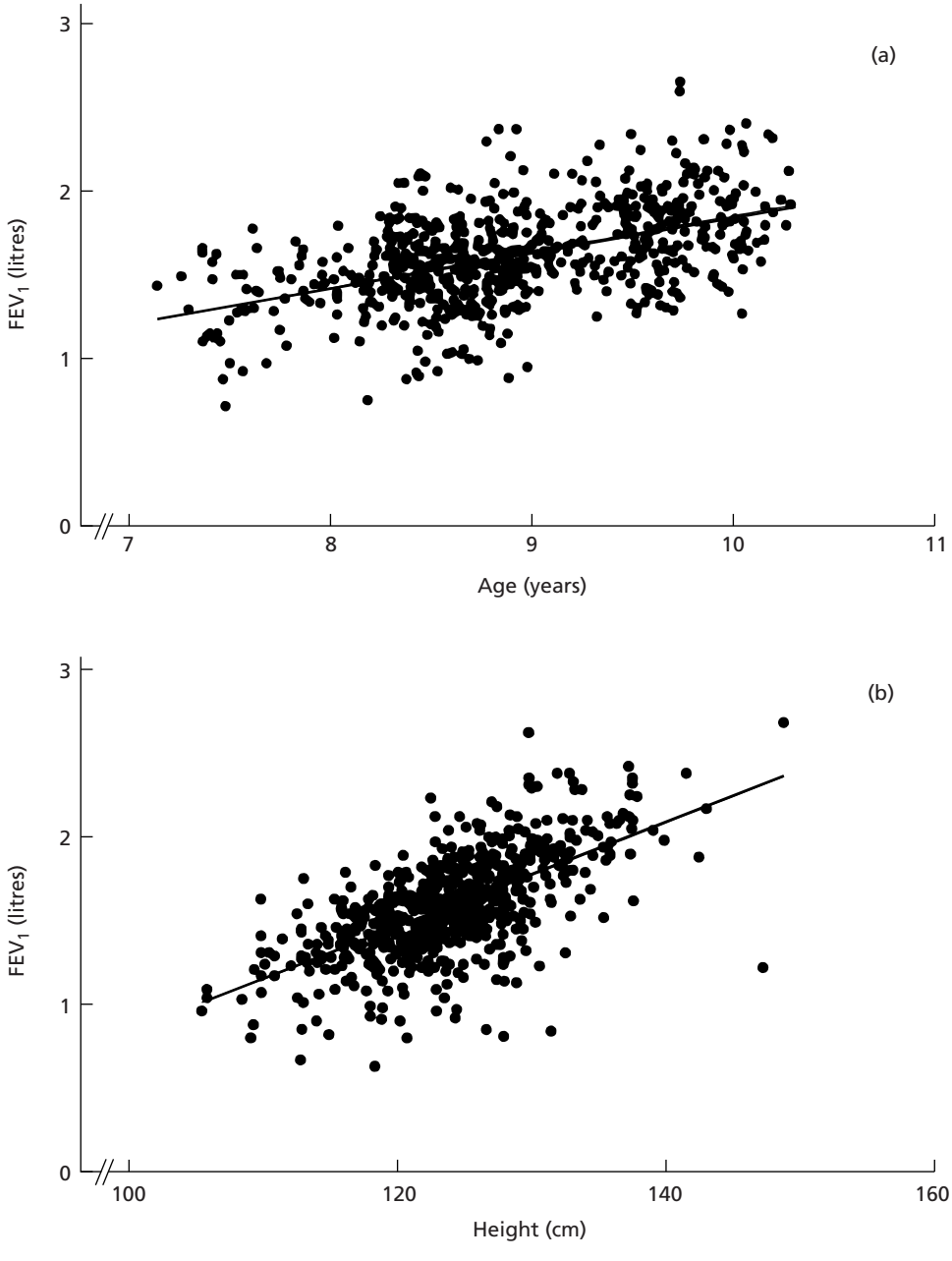


Fig. 11.1 Scatter plots showing the relationship of FEV_1 with (a) age and (b) height in 636 Peruvian children. Analyses and displays by kind permission of Dr M.E. Penny.

Often, we talk of the effect of a variable having **adjusted** or **controlled** for the effects of the other variable(s) in the model.

These questions may be answered by fitting a **multiple regression** model for the effects of height and age on FEV_1. The general form of a multiple regression model for the effects of two exposure variables (x_1 and x_2) on an outcome variable (y) is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

The **intercept** β_0 is the value of the outcome y when both exposure variables x_1 and x_2 are zero. In this example:

$$\text{FEV}_1 = \beta_0 + \beta_1 \times \text{age} + \beta_2 \times \text{height}$$

This model assumes that for any age, FEV_1 is linearly related to height, and correspondingly that for any height, FEV_1 is linearly related to age. Note that β_1 and β_2 will be different to the regression coefficients from the simple linear regressions on age and height separately, unless the two exposure variables are unrelated.

The way in which the regression coefficients are estimated is the same as for linear regression with a single exposure variable: the values of β_0, β_1 and β_2 are chosen to minimize the sum of squares of the differences $[y - (\beta_0 + \beta_1 x_1 + \beta_2 x_2)]$ or, in other words, the variation about the regression. In this example each observed FEV_1 is compared with $(\beta_0 + \beta_1 \times \text{age} + \beta_2 \times \text{height})$. The estimated regression coefficients are shown in Table 11.2.

The regression output tells us that the best-fitting model is:

$$\text{FEV}_1 = -2.3087 + 0.0897 \times \text{age} + 0.0250 \times \text{height}$$

After controlling for the association between FEV_1 and height, the regression coefficient for age is much reduced (from 0.2185 litres/year to 0.0897 litres/year). There is a smaller reduction in the regression coefficient for height: from 0.0311 litres/cm to 0.0250 litres/cm. The t statistics and corresponding P-values for age and height test the null hypotheses that, respectively, there is no association of

Table 11.2 Computer output showing the estimated regression coefficients from the multiple regression relating FEV_1 to age and height.

FEV_1	Coefficient	Std err	t	$P > \|t\|$	95% CI
Age	0.0897	0.0157	5.708	0.000	0.0588 to 0.1206
Height	0.0250	0.0018	13.77	0.000	0.0214 to 0.0285
Constant	−2.3087	0.1812	−12.743	0.000	−2.6645 to −1.9529

FEV_1 with age having controlled for its association with height, and no association of FEV_1 with height having controlled for its association with age.

Note that the *P*-values in this analysis are not really zero; they are simply too small to be displayed using the precision chosen by the software package. In this case the *P*-values should be interpreted and reported as < 0.001. There is thus strong evidence that age and height are each associated with FEV_1 after controlling for one another.

Analysis of variance for multiple regression

Example 11.1 (continued)

We can examine the extent to which the joint effects of age and height explain the variation in FEV_1 in an analysis of variance table (Table 11.3). There are now 2 degrees of freedom for the regression as there are two exposure variables. The *F* test for this regression is 244.3 with (2,633) degrees of freedom ($P < 0.0001$).

The regression accounts for 43.56% (25.6383/58.8584) of the total variation in FEV_1. This proportion equals R^2, where $R = \sqrt{0.4356} = 0.66$ is defined as the **multiple correlation coefficient**. *R* is always positive as no direction can be attached to a correlation based on more than one variable.

The sum of squares due to the regression of FEV_1 on both age and height comprises the sum of squares explained by age (= 15.6802, derived from the simple linear regression $FEV_1 = \beta_0 + \beta_1 \times$ age) plus the *extra* sum of squares explained by height after controlling for age (Table 11.4). This provides an alternative means of testing the null hypothesis that there is no association of FEV_1 with height having controlled for its association with age. We derive an *F* statistic using the residual mean square from the multiple regression:

$$F = 9.9581/0.05248 = 189.75,\ \text{d.f} = (1,633),\ P < 0.0001$$

Again, there is clear evidence of an association of FEV_1 with height having controlled for its association with age. Note that the *t* statistic for height presented in the computer output shown in Table 11.2 is exactly the square root of the *F* statistic: $\sqrt{189.75} = 13.77$.

Reversing the order in which the variables are entered into the model allows us to test the null hypothesis that there is no association with age having controlled for height: this gives an *F* statistic 32.58, d.f = (1,633), $P < 0.0001$. Again this corresponds to the *t* statistic in Table 11.2: $\sqrt{32.58} = 5.708$.

Table 11.3 Analysis of variance for the multiple regression relating FEV_1 to age and height.

Source of variation	SS	d.f.	MS	$F = \frac{\text{MS regression}}{\text{MS residual}}$
Regression on age and height of child	25.6383	2	12.8192	244.3, $P < 0.0001$
Residual	33.2201	633	0.05248	
Total	58.8584	635	0.09269	

Table 11.4 Individual contributions of age and height of the child to the multiple regression including both variables, when age is entered into multiple regression first.

Source of variation	SS	d.f.	MS	$F = \frac{\text{MS regression}}{\text{MS residual}}$
Age	15.6802	1	15.6082	
Height adjusting for age	9.9581	1	9.9581	189.75, $P < 0.0001$
Age and height	25.6383	2		

Note that these two orders of breaking down the combined regression sum of squares from Table 11.3 into the separate sums of squares do not give the same component sums of squares because the exposure variables (age and height) are themselves correlated. However, the regression coefficients and their corresponding standard errors in Table 11.2 are unaffected by the order in which the exposure variables are listed.

11.3 MULTIPLE REGRESSION WITH CATEGORICAL EXPOSURE VARIABLES

Until now, we have included only continuous exposure variables in regression models. In fact, it is straightforward to estimate the effects of binary or other categorical exposure variables in regression models. We now show how to do this, and how the results relate to methods introduced in previous chapters.

Regression with binary exposure variables

We start by considering a **binary exposure variable**, coded as 0 (unexposed) or 1 (exposed) in the dataset.

Example 11.1 (continued)

A variable that takes only the values 0 and 1 is known as an **indicator variable** because it indicates whether the individual possesses the characteristic or not. Computer output from the linear regression of FEV_1 on variable *male* in the data on lung function in Peruvian children is shown in Table 11.5. The interpretation of such output is straightforward.

1. The regression coefficient for the indicator variable is the difference between the mean in boys (variable *male* coded as 1) and the mean in girls (variable *male* coded as 0). The value of the *t* statistic (and corresponding *P*-value) for this coefficient is identical to that derived from the *t* test of the null hypothesis that the mean in girls is the same as in boys (see Chapter 7), and the confidence interval is identical to the confidence interval for the difference in means, also presented in Chapter 7.
2. The regression coefficient for the constant term is the mean in girls (the group for which the indicator variable is coded as 0).

To see why this is the case, consider the equation for this regression model. This states that on average:

Table 11.5 Computer output for the linear regression of FEV_1 on gender of the child.

FEV_1	Coefficient	Std err	t	$P > \|t\|$	95% CI
Male	0.1189	0.0237	5.01	0.000	0.0723 to 0.1655
Constant	1.5384	0.0163	94.22	0.000	1.5063 to 1.5705

$$\text{FEV}_1 = \beta_0 + \beta_1 \times \text{male}$$

Thus in girls, mean $\text{FEV}_1 = \beta_0 + \beta_1 \times 0 = \beta_0$ and so the estimated value of the intercept β_0 (the regression coefficient for the constant term) is the mean FEV_1 in girls. In boys, mean $\text{FEV}_1 = \beta_0 + \beta_1 \times 1 = \beta_0 + \beta_1$. Therefore:

$$\beta_1 = \text{mean FEV}_1 \text{ in boys} - \text{mean FEV}_1 \text{ in girls}$$

We may wish to ask whether the difference in mean FEV_1 between boys and girls is accounted for by differences in their age or height. This is done by including the three exposure variables together in a multiple regression model. The regression equation is:

$$\text{FEV}_1 = \beta_0 + \beta_1 \times \text{age} + \beta_2 \times \text{height} + \beta_3 \times \text{male}$$

Output for this model is shown in Table 11.6. The regression coefficient for variable male (β_3) estimates the difference in mean FEV_1 in boys compared to girls, *having allowed for the effects of age and height*. This is slightly increased compared to the mean difference before the effects of age and height were taken into account.

Table 11.6 Computer output for the multiple regression of FEV_1 on age, height and gender of the child.

FEV_1	Coefficient	Std err	t	$P > \|t\|$	95% CI
Age	0.0946	0.0152	6.23	0.000	0.0648 to 0.1244
Height	0.0246	0.0018	14.04	0.000	0.0211 to 0.0280
Male	0.1213	0.0176	6.90	0.000	0.0868 to 0.1559
Constant	−2.360	0.1750	−13.49	0.000	−2.704 to −2.0166

Regression with exposure variables with more than two categories

The effects of categorical exposures with more than two levels (for example age-group or extent of exposure to cigarette smoke) are estimated by introducing a series of indicator variables to describe the differences. First we choose a **baseline** group to which the other groups are to be compared: often this is the lowest coded value of the variable or the group representing the unexposed category. If the variable has k levels, $k - 1$ indicator variables are then included, corresponding to each non-baseline group. This is explained in more detail in the context of logistic regression, in the box in Section 19.3. The regression coefficients for the indicator

variables then equal the differences in mean outcome, comparing each non-baseline group with the baseline.

11.4 GENERAL FORM OF THE MULTIPLE REGRESSION MODEL

The general form of a **multiple regression model** for the effects of p exposure variables is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_p x_p + e$$

The quantity, $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_p x_p$, on the right-hand side of the equation is known as the **linear predictor** of the outcome y, given particular values of the exposure variables x_1 to x_p. The *error*, e, is normally distributed with mean zero and standard deviation σ, which is estimated by the square root of the residual mean square.

11.5 MULTIPLE REGRESSION WITH NON-LINEAR EXPOSURE VARIABLES

It is often found that the relationship between the outcome variable and an exposure variable is non-linear. There are three possible ways of incorporating such an exposure variable in the multiple regression equation. The first method is to redefine the variable into distinct subgroups and include it as a categorical variable using indicator variables, as described in Section 11.3, rather than as a numerical variable. For example, age could be divided into five-year age-groups. The relationship with age would then be based on a comparison of the means of the outcome variable in each age-group (assuming that mean outcome is approximately constant in each age group) but would make no other assumption about the form of the relationship of mean outcome with age. At the initial stages of an analysis, it is often useful to include an exposure variable in both forms, as a numerical variable and grouped as a categorical variable. The difference between the two associated sums of squares can then be used to assess whether there is an important non-linear component to the relationship. For most purposes, a subdivision into 3–5 groups, depending on the sample size, is adequate to investigate non-linearity of the relationship. See Section 29.6 for more detail.

A second possibility is to find a suitable transformation for the exposure variable. For example, in a study of women attending an antenatal clinic conducted to identify variables associated with the birth weight of their baby, it was found that birth weight was linearly related to the logarithm of family income rather than to family income itself. The use of transformations is discussed more fully in Chapter 13. The third possibility is to find an algebraic description of the relationship. For example, it may be quadratic, in which case both the variable (x) and its square (x^2) would be included in the model. This is described in more detail in Section 29.6.

11.6 RELATIONSHIP BETWEEN MULTIPLE REGRESSION AND ANALYSIS OF VARIANCE

Analysis of variance is simply a special case of multiple regression. The two approaches give identical results. A regression model test of the null hypothesis that there is no difference in mean response between k exposure groups uses an F test with $(k-1,\ n-k)$ degrees of freedom. This is identical to the F statistic derived using a one-way analysis of variance (see Chapter 9). Similarly, inclusion of two categorical variables (using indicator variables) in a multiple regression model will give identical results to a two-way analysis of variance. Analysis of variance can also be extended to examine differences between groups adjusted for the effects of numerical exposure variables, as described for multiple regression above, when the difference in FEV_1 between males and females was adjusted for age and height (Table 11.6). In this context it is sometimes called **analysis of covariance** (Armitage and Berry 2002), and the numerical exposure variables are called **covariates**.

11.7 MULTIVARIATE ANALYSIS

Multiple regression, and other regression models (see Chapters 19–21, 24 and 27) are often referred to as *multivariate* methods, since they investigate how an outcome variable is related to more than one exposure variable. A better term for such models is to call them **multivariable** regression models. In the strict statistical sense, **multivariate analysis** means the study of how several outcome variables vary together. The three methods most relevant to medical research will briefly be described. For more detail see Armitage and Berry (2002) and Everitt and Dunn (2001).

Principal component analysis is a method used to find a few combinations of variables, called components, that adequately explain the overall observed variation, and thus to reduce the complexity of the data. **Factor analysis** is a related method commonly used in the analysis of psychological tests. It seeks to explain how the responses to the various test items may be influenced by a number of underlying factors, such as emotion, rational thinking, etc. Finally, **cluster analysis** is a method that examines a collection of variables to see if individuals can be formed into any natural system of groups. Techniques used include those of **numerical taxonomy**, principal components and **correspondence analysis**.

CHAPTER 12

Goodness of fit and regression diagnostics

12.1 INTRODUCTION

In this chapter we discuss how to assess whether the distribution of an observed set of data agrees with that expected under a particular theoretical model. We start by considering how to assess whether the distribution of a variable conforms with the normal distribution, as assumed in the statistical methods described in this part of the book. We then consider how to check the assumptions made in fitting linear and multiple regression models. The final part of the chapter is more general. It describes the chi-squared goodness of fit test for testing whether an observed frequency distribution differs from the distribution predicted by a theoretical model.

12.2 GOODNESS OF FIT TO A NORMAL DISTRIBUTION

The assumption of normality underlies the linear regression, multiple regression and analysis of variance methods introduced earlier in this section. It can be checked by comparing the shape of the observed frequency distribution with that of the normal distribution. Formal significance testing is rarely necessary, as in general we are only interested in detecting marked departures from normality; the methods are robust against moderate departures so that parameter estimates, confidence intervals and hypothesis tests remain valid. If the sample size is large, visual assessment of the frequency distribution is often adequate.

The main problem with departures from normality is that the standard errors of parameter estimates may be wrong. In Chapter 13 we describe how to transform variables to make them more normally distributed, and in Chapter 30 we see how to check for this problem by deriving alternative standard errors (for example using bootstrapping or robust standard errors).

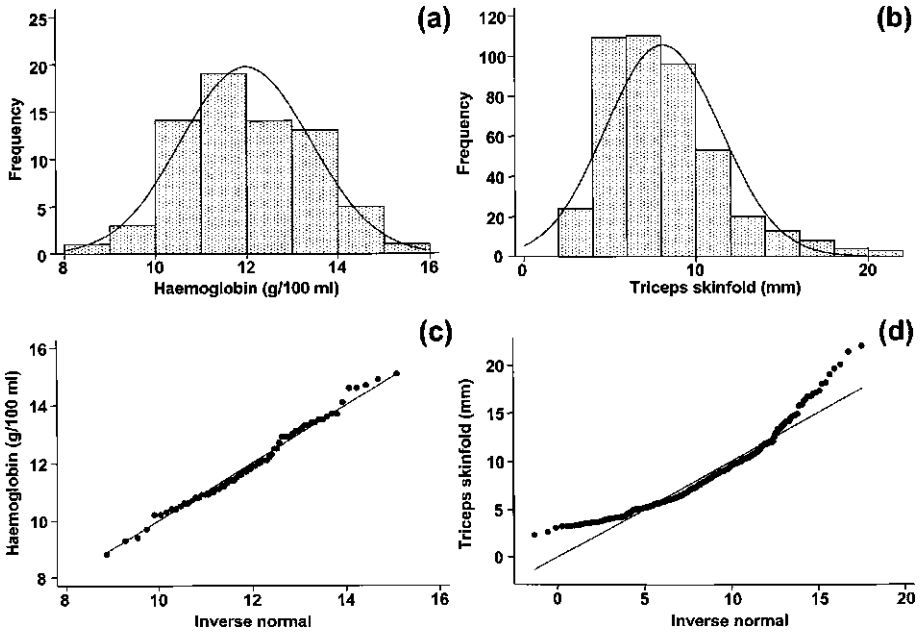


Fig. 12.1 Frequency distributions with inverse normal plots to assess the normality of the data. (a) and (c) Haemoglobin levels of 70 women (normally distributed, inverse normal plot linear). (b) and (d) Triceps skinfold measurements of 440 men (positively skewed, inverse normal plot non-linear).

Example 12.1

In Table 3.2 we presented measurements of haemoglobin (g/100 ml) in 70 women. The distribution of these measurements will be compared with that of triceps skinfold measurements made in 440 men. Histograms of these variables, together with the corresponding normal distribution curves with the same means and standard deviations, are shown in Figure 12.1(a) and (b). For haemoglobin the shape seems reasonably similar to that of the normal distribution, while that for triceps skinfold is clearly positively (right-) skewed.

Inverse normal plots

The precise shape of the histogram depends on the choice of groups, and it can be difficult to tell whether or not the bars at the extreme of the distribution are consistent with the normal distribution. A graphical technique that avoids these problems is the **inverse normal plot**. This is a scatter plot comparing the values of the observed distribution with the corresponding points of the normal distribution. The inverse normal plot is linear if the data are normally distributed and curved if they are not. The plot is constructed as follows:

1 The measurements are arranged in order, and the corresponding quantiles of the distribution are calculated as $1/(n+1)$, $2/(n+1)$, $\ldots$ $n/(n+1)$. Table 12.1 illustrates the calculations for the haemoglobin data. It shows the

Table 12.1 Calculations of points for inverse normal plot of 70 haemoglobin measurements.

Observation no.	Haemoglobin (g/100 ml)	Quantile	Probit	Inverse normal = 11.98 + probit × 1.41
1	8.8	1/71 = 1.4%	−2.195	8.88
34	11.8	34/71 = 49.3%	−0.018	11.96
35	11.9	35/71 = 50.7%	0.018	12.01
70	15.1	70/71 = 98.6%	2.195	15.09

minimum (1st), median (34th and 35th) and maximum (70th) haemoglobin measurements, together with their corresponding quantiles.

2 For each measurement, the **probit** (the value of the standard normal distribution corresponding to its quantile) is derived using Table A6 in the Appendix or (more commonly) using a computer. For example, the value of the standard normal distribution corresponding to a quantile of 1.4% is −2.195, since 1.4% of the standard normal distribution lies *below* this value.

3 The corresponding points of the normal distribution with the same standard deviation and mean as the data are found by multiplying the probit by the standard deviation, then adding the mean. This is called the **inverse normal**:

$$\text{Inverse normal} = \text{mean} + \text{probit} \times \text{s.d.}$$

For the haemoglobin data, the mean is 11.98, and the standard deviation is 1.41 g/100 ml.

4 Finally, the original values are plotted against their corresponding inverse normal points. Figure 12.1(c) shows the haemoglobin levels plotted against their corresponding inverse normal points. If haemoglobin levels are normally distributed then they should lie along the line of identity (the line where $y = x$) shown on the plot. The plot is indeed linear, confirming the visual impression from the histogram that the haemoglobin data are normally distributed.

In contrast, Figure 12.1(d) shows the non-linear inverse normal plot corresponding to the positively skewed distribution of triceps skinfold measurements shown in Figure 12.1(b). The line is clearly curved, and illustrates the deficit of observations on the left and corresponding excess on the right.

Skewness and kurtosis

We now introduce two measures that can be used to assess departures from normality. In Chapter 4 we saw that the variance is defined as the average of the squared differences between each observation and the mean:

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

Because the variance is based on the sum of the *squared* (power 2) differences between each observation and the sample mean, it is sometimes called the **second moment**, $m_2 = s^2$. The **third and fourth moments** of a distribution are defined in a similar way, based on the third and fourth powers of the differences:

$$\text{Third moment } m_3 = \frac{\Sigma(x - \bar{x})^3}{n} \quad \text{and} \quad \text{Fourth moment } m_4 = \frac{\Sigma(x - \bar{x})^4}{n}$$

The **coefficients of skewness** and **kurtosis** of a distribution are defined as:

$$\text{skewness} = m_3 m_2^{-\frac{3}{2}} \quad \text{and} \quad \text{kurtosis} = m_4 m_2^{-2}$$

For any symmetrical distribution, the coefficient of skewness is zero: positive values of the coefficient of skewness correspond to a right-skewed distribution while negative values correspond to a left-skewed distribution.

The coefficient of kurtosis measures how spread out are the values of a distribution. For the normal distribution the coefficient of kurtosis is 3. If the distribution is more spread out than the normal distribution then the coefficient of kurtosis will be greater than 3. For example, Figure 6.3 shows that compared to the normal distribution, the t distribution with 5 degrees of freedom is more spread out. The kurtosis of the t distribution with 5 d.f. is approximately 7.6.

Example 12.1 (continued)

For the 70 measurements of haemoglobin (g/100 ml) the coefficients of skewness and kurtosis were 0.170 and 2.51 respectively. This distribution shows little evidence of asymmetry, since the coefficient of skewness is close to zero. The coefficient of kurtosis shows that the spread of the observations was slightly less than would have been expected under the normal distribution. For the 440 measurements of triceps skinfold (mm) the coefficients of skewness and kurtosis were 1.15 and 4.68 respectively. This distribution is right-skewed and more spread out than the normal distribution.

Shapiro–Wilk test

We stated at the start of this section that although the assumption of normality underlies most of the statistical methods presented in this part of the book, formal tests of this assumption are rarely necessary. However, the assumption of a normal distribution may be of great importance if we wish to predict ranges within which a given proportion of the population should lie. For example, growth charts for babies and infants include lines within which it is expected that 90%, 99% and even 99.9% of the population will lie. Departures from normality may be very important if we wish to use the data to construct such charts.

The **Shapiro–Wilk test** (Shapiro and Wilk 1965, Royston 1993) is a general test of the assumption of normality, based on comparing the ordered sample values with those which would be expected if the distribution was normal (as done in the inverse normal plots introduced earlier). The mathematics of the test are a little complicated, but it is available in many statistical computer packages.

Example 12.1 (continued)

The *P*-values from the Shapiro–Wilk test were 0.612 for the haemoglobin measurements and < 0.0001 for the triceps measurements. As suggested by the quantile plots and coefficients of skewness and kurtosis, there is strong evidence against the assumption of normality for the triceps measurements, but no evidence against this assumption for the haemoglobin measurements.

12.3 REGRESSION DIAGNOSTICS

Examining residuals

In Chapters 10 and 11 we saw that linear and multiple regression models are fitted by minimizing the **residual sum of squares**:

$$\mathrm{SS}_{\mathrm{residual}} = \Sigma[y - (\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots)]^2$$

The differences $[y - (\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots)]$ between the observed outcome values and those predicted by the regression model (the dashed vertical lines in Figures 10.3 and 10.5) are called the **residuals**. As explained in Chapter 10, it is assumed that the residuals are normally distributed. This assumption can be examined using the methods introduced in the first part of this chapter.

Example 12.2

Figure 12.2(a) shows a histogram of the residuals from the multiple linear regression of FEV_1 on age, height and sex from the data on lung function in schoolchildren from Peru which were introduced in Chapter 11, while Figure 12.2(b) shows the corresponding inverse normal plot. The distribution appears reasonably close

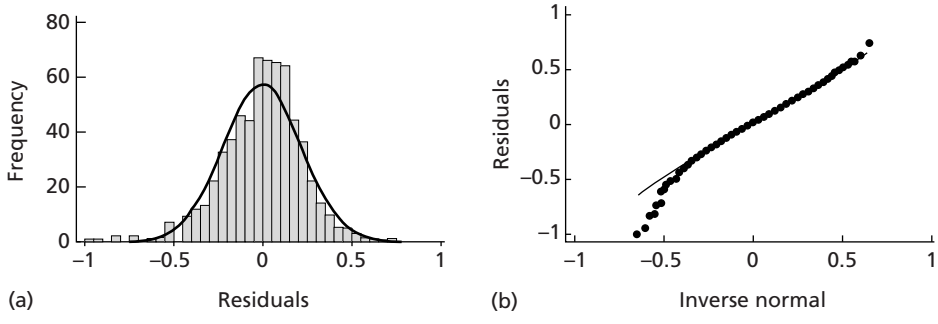


Fig. 12.2 (a) Histogram and (b) inverse normal plot of the residuals from the multiple linear regression of FEV_1 on age, height and sex.

to normal except at the extreme left. The coefficients of skewness and kurtosis are −0.52 and 4.68 respectively, confirming this impression.

The *P*-value from the Shapiro–Wilk test is less than 0.0001 so there is clear evidence that the distribution is not normal. However, Figure 12.2 shows that the departure from normality is fairly modest and is unlikely to undermine the results of the analysis. For fairly large datasets such as this one the Shapiro–Wilk test is extremely sensitive to departures from normality, while the central limit theorem (see Chapter 5) means that the parameter estimates are likely to be normally distributed even though the residuals are not.

A particular use of the residual plot is to detect unusual observations (**outliers**): those for which the observed value of the outcome is a long way from that predicted by the model. For example, we might check the data corresponding to the extreme left of the distribution to make sure that these observations have not resulted from coding errors in either the outcome or exposure variables. In general, however, outliers should not be omitted simply because they are at the extreme of the distribution. Unless we know they have resulted from errors they should be included in our analyses. We discuss how to identify observations with a substantial influence on the regression line later in this section.

Plots of residuals against fitted values

Having estimated the parameters of a regression model we can calculate the fitted values (also called predicted values) for each observation in the data. For example, the fitted values for the regression of FEV_1 on age, height and gender (see Table 11.6) are calculated using the regression equation:

$$FEV_1 = -2.360 + 0.0946 \times \text{age} + 0.0246 \times \text{height} + 0.1213 \times \text{male}$$

where the indicator variable *male* takes the value 0 in girls and 1 in boys. These values can be calculated for every child in the dataset. If the model fits the data well

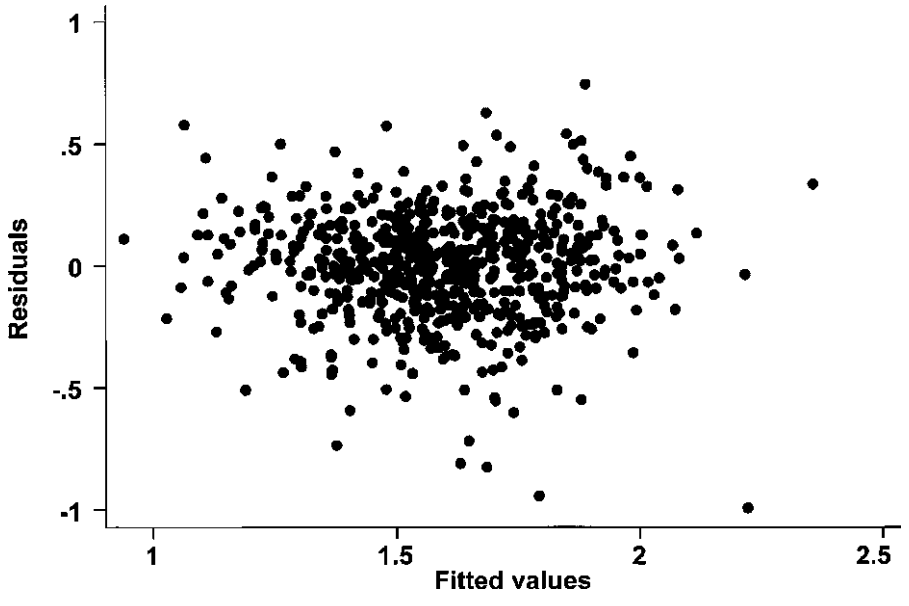


Fig. 12.3 Scatter plot of residuals against fitted values, for the regression of FEV_1 on age, height and gender.

then there should be no association between the fitted values and the residuals. This assumption can be examined in a scatter plot, as shown in Figure 12.3.

There is no strong pattern in Figure 12.3, but it does seem that the variability in the residuals increases a little with increasing fitted values, and that there may be a U-shaped relationship between the residuals and the fitted values. We might investigate this further by examining models which allow for quadratic or other non-linear associations between FEV_1 and age or height (see Section 29.6).

A common problem is that the variability (spread) of the residuals increases with increasing fitted values. This may indicate the need for a **log transformation** of the outcome variable (see Section 13.2).

Influence

A final consideration is whether individual observations have a large **influence** on the estimated regression line. In other words, would the omission of a particular observation make a large difference to the regression?

Example 12.3

Figure 12.4 is a scatter plot of a hypothetical outcome variable y against an exposure x. There appears to be clear evidence of an association between x and y: the slope of the regression line is 0.76, 95% CI = 0.32 to 1.19, $P = 0.004$. However, inspection of the scatter plot leads to the suspicion that the association

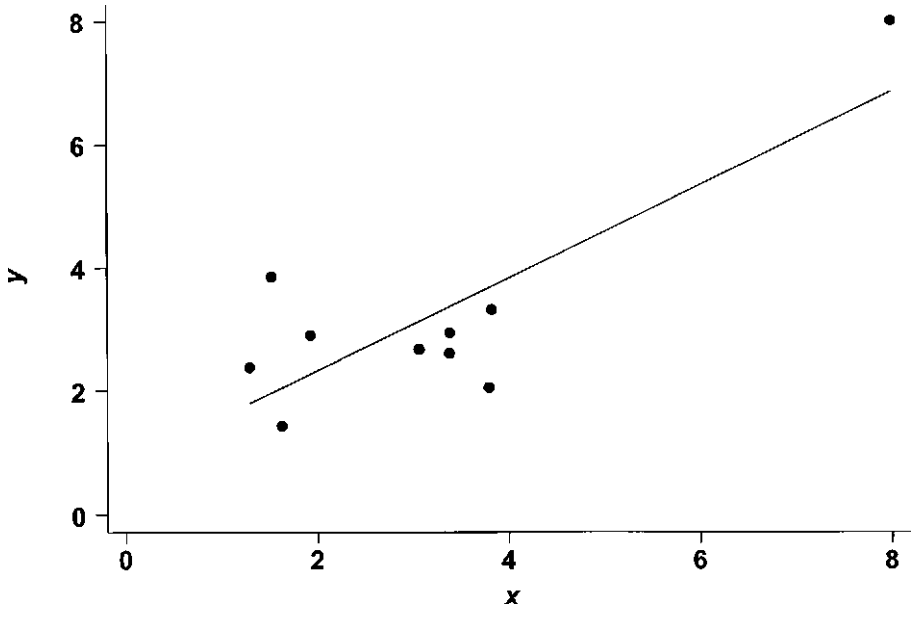


Fig. 12.4 Scatter plot of a hypothetical outcome variable *y* against an exposure *x*, in which there is a highly influential observation at the top right of the graph.

is mainly because of the point at the top right of the graph. The point is close to the regression line, so examining the residuals will not reveal a problem.

To assess the dependence of the regression on individual observations we calculate **influence** statistics. The most commonly used measure of influence is **Cook's D**. These statistics are listed, together with the residuals, in Table 12.2. It can be seen that observation 10 (the point on the top right of the graph) has much greater influence than the other observations. It would be appropriate to check whether this point arose because of an error in coding or data entry, or if there is some

Table 12.2 Data plotted in Figure 12.4, together with the influence statistic and residual for each observation.

Observation	*y*	*x*	Influence (Cook's D)	Residual
1	2.94	3.39	0.01	−0.43
2	3.32	3.83	0.01	−0.38
3	1.44	1.63	0.04	−0.61
4	2.05	3.80	0.15	−1.63
5	2.90	1.94	0.03	0.63
6	2.38	1.30	0.05	0.59
7	2.67	3.07	0.01	−0.45
8	3.85	1.53	0.39	1.89
9	2.60	3.38	0.03	−0.76
10	8.00	8.00	8.25	1.15

clear explanation for it being different from the rest of the population. As discussed earlier, observations should not be omitted from the regression purely because they have large residuals or have a large influence on the results. However, we might check whether similar conclusions are reached if an observation is omitted: and perhaps present results both including and excluding a highly influential observation.

Another useful plot is a scatter plot of influence against residuals (or squared residual) for each observation. Observations with large influence, large residuals or both may lead to further checks on the data, or attempts to fit different regression models. **Standardized residuals**, which are the residual divided by its standard error, are also of use in checking the assumptions made in regression models. These are discussed in more detail in Draper and Smith (1998) and Weisberg (1985).

What to do if the regression assumptions appear to be violated

The more checks we make, the more likely we are to find possible problems with our regression model. Evidence that assumptions are violated in one of the ways discussed here is *not* a reason to reject the whole analysis. It is very important to remember that provided that the sample size is reasonably large the results may well be robust to violation of assumptions. However, possible actions that might be taken include:

- checks for mistakes in data coding or data entry which have led to outlying or influential observations;
- exploration of non-linear relationships between the outcome and exposure variables;
- sensitivity analyses which examine whether conclusions change if influential observations are omitted;
- use of transformations as described in the next chapter;
- use of methods such as bootstrapping to derive confidence intervals independently of the assumptions made in the model about the distribution of the outcome variable. These are discussed in Chapter 30.

12.4 CHI-SQUARED GOODNESS OF FIT TEST

It is sometimes useful to test whether an observed frequency distribution differs significantly from a postulated theoretical one. This may be done by comparing the observed and expected frequencies using a chi-squared test. The form of the test is:

$$\chi^2 = \Sigma \frac{(O - E)^2}{E}$$

This is exactly the same as that for contingency tables, which is introduced in Chapter 17. Like the t distribution, the shape of the chi-squared distribution depends on the **degrees of freedom**. Here, these equal the number of groups in the frequency distribution minus 1, minus the number of parameters estimated from the data. In fitting a normal distribution, two parameters are needed, its mean, μ, and its standard deviation, σ. In some cases no parameters are estimated from the data, either because the theoretical model requires no parameters, as in Example 12.4 below, or because the parameters are specified as part of the model.

$$\text{d.f.} = \begin{array}{c}\text{number of groups}\\ \text{in frequency}\\ \text{distribution}\end{array} - \begin{array}{c}\text{number of}\\ \text{parameters}\\ \text{estimated}\end{array} - 1$$

Calculation of expected numbers

The first step in carrying out a chi-squared goodness of fit test is to estimate the parameters needed for the theoretical distribution from the data. The next step is to calculate the **expected numbers** in each category of the frequency distribution, by multiplying the total frequency by the probability that an individual value falls within the category.

$$\begin{array}{c}\text{Expected}\\ \text{frequency}\end{array} = \begin{array}{c}\text{total}\\ \text{frequency}\end{array} \times \begin{array}{c}\text{probability individual falls}\\ \text{within category}\end{array}$$

For discrete data, the probability is calculated by a straightforward application of the distributional formula. This is illustrated later in the book for the Poisson distribution (see Example 28.3).

Validity

The chi-squared goodness of fit test should not be used if more than a small proportion of the *expected* frequencies are less than 5 or if any are less than 2. This can be avoided by combining adjacent groups in the distribution.

Example 12.4

Table 12.3 examines the distribution of the final digit of the weights recorded in a survey, as a check on their accuracy. Ninety-six adults were weighed and their weights recorded to the nearest tenth of a kilogram. If there were no biases in recording, such as a tendency to record only whole or half kilograms, one would expect an equal number of 0s, 1s, 2s . . . and 9s for the final digit, that is 9.6 of each.

Table 12.3 Check on the accuracy in a survey of recording weight.

Final digit of weight	Observed frequency	Expected frequency	$\frac{(O-E)^2}{E}$
0	13	9.6	1.20
1	8	9.6	0.27
2	10	9.6	0.02
3	9	9.6	0.04
4	10	9.6	0.02
5	14	9.6	2.02
6	5	9.6	2.20
7	12	9.6	0.60
8	11	9.6	0.20
9	4	9.6	3.27
Total	96	96.0	9.84

The agreement of the observed distribution with this can be tested using the chi-squared goodness of fit test. There are ten frequencies and no parameters have been estimated.

$$\chi^2 = \Sigma\frac{(O-E)^2}{E} = 9.84,\ \text{d.f.} = 10 - 0 - 1 = 9,\ P = 0.36$$

The observed frequencies therefore agree well with the theoretical ones, suggesting no recording bias.

CHAPTER 13

Transformations

13.1 INTRODUCTION

The assumption of normality will not always be satisfied by a particular set of data. For example, a distribution may be positively skewed and this will often mean that the standard deviations in different groups will be very different. Or a relationship between the outcome and exposure variable(s) may not be linear, violating the assumptions of the linear and multiple regression methods introduced in this part of the book. We will now describe how such problems can often be overcome simply by transforming the data to a different scale of measurement. By far the most common choice is the logarithmic transformation, which will be described in detail. A summary of the use of other transformations will then be presented.

Finally, in the last section of the chapter, we describe the use of ***z*-scores** to compare data against **reference curves** in order to improve their *interpretability*. In particular, we explain why this is the standard approach for the analysis of **anthropometric data**.

13.2 LOGARITHMIC TRANSFORMATION

When a logarithmic transformation is applied to a variable, each individual value is replaced by its logarithm.

$$u = \log x$$

where x is the original value and u the transformed value. The meaning of logarithms is easiest to understand in reverse. We will start by explaining this for **logarithms to the base 10**.

If $x = 10^u$, then by definition 'u is the logarithm (base 10) of x'

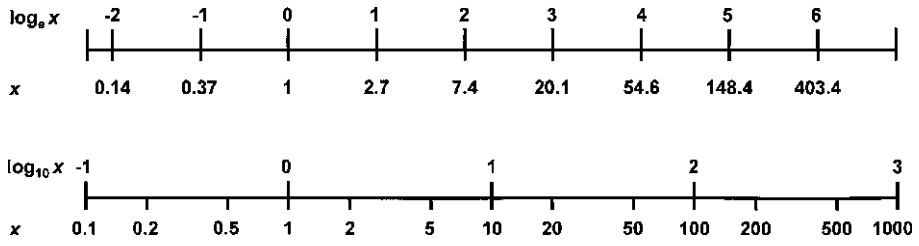


Fig. 13.1 The logarithmic transformation, using base 10 (lower line) and base *e* (upper line).

Thus, for example, since $100 = 10^2$, $2 = \log_{10}(100)$, and since $0.1 = 10^{-1}$, $-1 = \log_{10}(0.1)$. Different values of x and $\log_{10}(x)$ are shown in the lower part of Figure 13.1. The logarithmic transformation has the effect of stretching out the lower part of the original scale, and compressing the upper part. For example, on a logarithmic scale, the distance between 1 and 10 is the same as that between 10 and 100 and as that between 100 and 1000; they are all ten-fold differences.

Although logarithms to base 10 are most easily understood, statistical packages generally use **logarithms to base e**, where e is the 'natural constant':

$$e = 2.7182818$$

The function e^x is called the **exponential function** and is often written as $\exp(x)$.

If $x = e^u$, then by definition 'u is the logarithm (base e) of x'

Logarithms to base e are also known as **natural logarithms**. For example, $7.389 = e^2$ so $2 = \log_e(7.389)$, $20.086 = e^3$ so $3 = \log_e(20.086)$, and $0.3679 = e^{-1}$ so $-1 = \log_e(0.3679)$. Different values of x and $\log_e(x)$ are shown in the upper part of Figure 13.1. Note that logarithms to base 10 are simply logarithms to base e multiplied by a constant amount:

$$\log_{10}(x) = \log_{10}(e) \times \log_e(x) = 0.4343 \times \log_e(x)$$

Throughout this book, we will use logarithms to base e (natural logarithms). We will omit the subscript, and refer simply to $\log(x)$. The notation $\ln(x)$ is also used to refer to natural logarithms. For more on the laws of logarithms see Section 16.5, where we show how logarithmic transformations are used to derive confidence intervals for ratio measures such as risk ratios and odds ratios.

Logarithmic transformations can only be used with positive values, since logarithms of negative numbers do not exist, and the logarithm of zero is minus infinity. There are sometimes instances, however, when a logarithmic transformation is indicated, as in the case of parasite counts, but the data contain some zeros as well as positive numbers. This problem can be solved by adding a constant to each value before transforming, although it must be remembered that the choice of the constant does affect the results obtained. One is a common choice. Note also that 1 must then also be subtracted after the final results have been converted back to the original scale.

Positively skewed distributions

Example 13.1

The logarithmic transformation will tend to normalize positively skewed distributions, as illustrated by Figure 13.2, which is the result of applying a logarithmic transformation to the triceps skinfold data presented in Figure 12.1(b). The histogram is now symmetrical and the inverse normal plot linear, showing that the transformation has removed the skewness and normalized the data. Triceps skinfold is said to have a **lognormal distribution**.

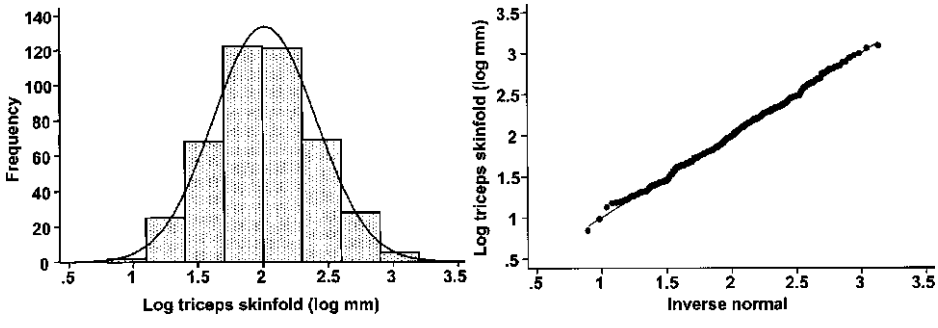


Fig. 13.2 Lognormal distribution of triceps skinfold measurements of 440 men. Compare with Figure 12.1 (b) and (d).

Unequal standard deviations

Example 13.2

The mechanics of using a logarithmic transformation will be described by considering the data of Table 13.1(a), which show a higher mean urinary β-thromboglobulin (β-TG) excretion in 12 diabetic patients than in 12 normal subjects. These means cannot be compared using a *t* test since the standard deviations of the two groups are very different. The right-hand columns of the table show the observations after a logarithmic transformation. For example, $\log_e(4.1) = 1.41$.

The transformation has had the effects both of equalizing the standard deviations (they are 0.595 and 0.637 on the logarithmic scale) and of removing skewness in each group (see Figure 13.3). The *t* test may now be used to examine

Table 13.1 Comparison of urinary β-thromboglobulin (β-TG) excretion in 12 normal subjects and in 12 diabetic patients. Adapted from results by van Oost, B.A., Veldhuyzen, B., Timmermans, A.P.M. & Sixma, J.J. (1983) Increased urinary β-thromboglobulin excretion in diabetes assayed with a modified RIA, Kit-Technique. *Thrombosis and Haemostasis* (Stuttgart) **49** (1): 18–20, with permission.

(a) Original and logged data.

	β-TG (ng/day/100 ml creatinine)		Log β-TG (log ng/day/100 ml creatinine)	
	Normals	Diabetics	Normals	Diabetics
	4.1	11.5	1.41	2.44
	6.3	12.1	1.84	2.49
	7.8	16.1	2.05	2.78
	8.5	17.8	2.14	2.88
	8.9	24.0	2.19	3.18
	10.4	28.8	2.34	3.36
	11.5	33.9	2.44	3.52
	12.0	40.7	2.48	3.71
	13.8	51.3	2.62	3.94
	17.6	56.2	2.87	4.03
	24.3	61.7	3.19	4.12
	37.2	69.2	3.62	4.24
Mean	13.53	35.28	2.433	3.391
s.d.	9.194	20.27	0.595	0.637
n	12	12	12	12

(b) Calculation of *t* test on logged data.

$$s = \sqrt{[(11 \times 0.595^2 + 11 \times 0.637^2)/22]} = 0.616$$

$$t = \frac{2.433 - 3.391}{0.616\sqrt{1/12 + 1/12}} = -3.81, \text{ d.f.} = 22, P = 0.001$$

(c) Results reported in original scale.

	Geometric mean β-TG	95% CI
Normals	exp(2.433) = 11.40	7.81 to 16.63
Diabetics	exp(3.391) = 29.68	19.81 to 44.49

differences in mean log β-TG between diabetic patients and normal subjects. The details of the calculations are presented in Table 13.1(b).

Geometric mean and confidence interval

Example 13.2 (continued)

When using a transformation, all analyses are carried out on the transformed values, u. It is important to note that this includes the calculation of any

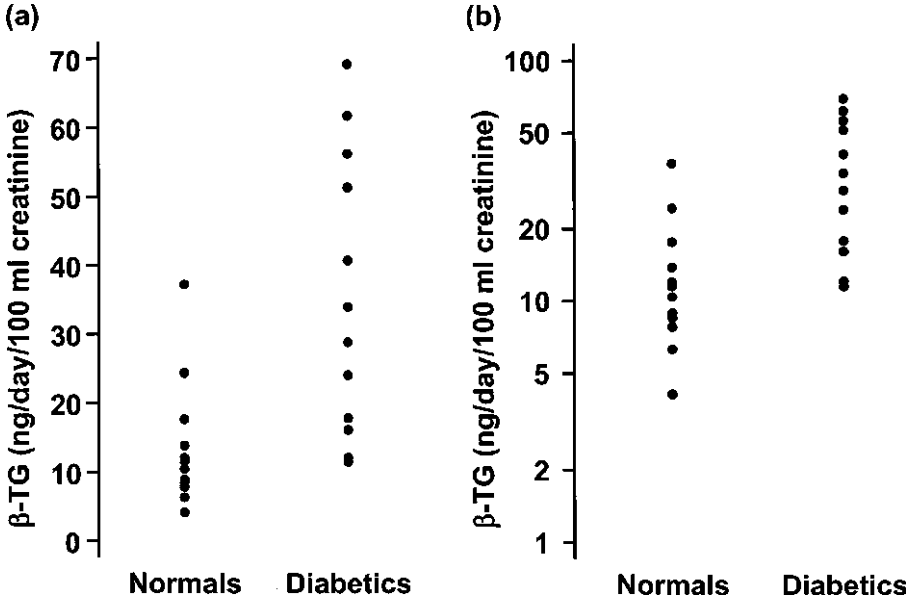


Fig. 13.3 β-Thromboglobulin data (Table 13.1) drawn using (a) a linear scale and (b) a logarithmic scale. Note that the logarithmic scale has been labelled in the original units.

confidence intervals. For example, the mean log β-TG of the normals was 2.433 log ng/day/100 ml. Its 95% confidence interval is:

$$\begin{aligned} 95\%\ \text{CI} &= 2.433 - 2.20 \times 0.595/\sqrt{12} \text{ to } 2.433 - 2.20 \times 0.595/\sqrt{12} \\ &= 2.055 \text{ to } 2.811 \text{ ng/day/100 ml} \end{aligned}$$

Note that 2.20 is the 5% point of the t distribution with 11 degrees of freedom.

When reporting the final results, however, it is sometimes clearer to transform them back into the original units by taking **antilogs** (also known as **exponentiating**), as done in Table 13.1(c). The antilog of the mean of the transformed values is called the **geometric mean**.

$$\text{Geometric mean (GM)} = \text{antilog}(\bar{u}) = \exp(\bar{u}) = e^{\bar{u}}$$

For example, the geometric mean β-GT of the normal subjects is:

$$\text{Antilog}(2.433) = e^{2.433} = 11.39 \text{ ng/day/100 ml}$$

The geometric mean is always smaller than the corresponding arithmetic mean (unless all the observations have the same value, in which case the two measures are equal). Unlike the arithmetic mean, it is not overly influenced by the very large

values in a skewed distribution, and so gives a better representation of the average in this situation.

Its confidence interval is calculated by exponentiating the confidence limits calculated on the log scale. For the normal subjects, the 95% confidence interval for the geometric mean therefore equals:

$$95\%\ \text{CI} = \exp(2.055)\ \text{to}\ \exp(2.811) = 7.81\ \text{to}\ 16.63\ \text{ng/day/100 ml}$$

Note that the confidence interval is not symmetric about the geometric mean. Instead the ratio of the upper limit to the geometric mean, $16.63/11.39 = 1.46$, is the same as the ratio of the geometric mean to the lower limit, $11.39/7.81 = 1.46$. This reflects the fact that a standard deviation on a log scale corresponds to a *multiplicative* rather than an *additive* error on the original scale. For the same reason, the antilog of the standard deviation is not readily interpretable, and is therefore not commonly used.

Non-linear relationship

Example 13.3

Figure 13.4(a) shows how the frequency of 6-thioguanine (6TG) resistant lymphocytes increases with age. The relationship curves upwards and there is greater scatter of the points at older ages. Figure 13.4(b) shows how using a log transformation for the frequency has both linearized the relationship and stabilized the variation.

In this example, the relationship curved upwards and the y variable (frequency) was transformed. The equivalent procedure for a relationship that curves downwards is to take the logarithm of the x value.

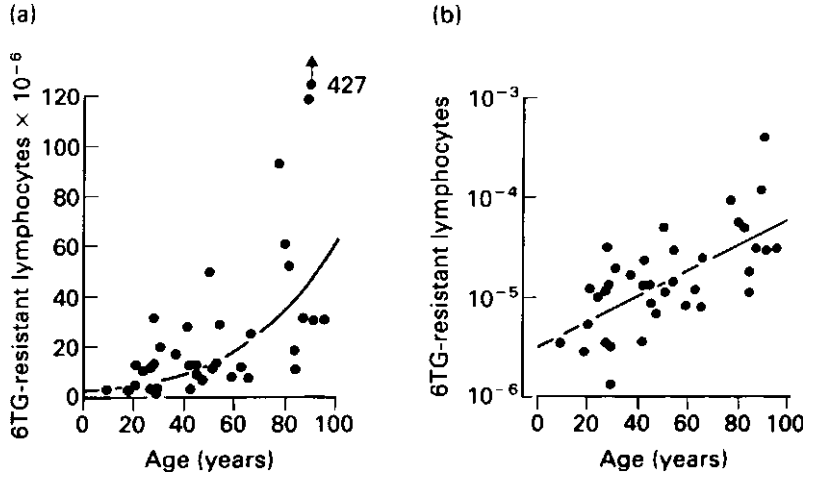


Fig. 13.4 Relationship between frequency of 6TG-resistant lymphocytes and age for 37 individuals drawn using (a) a linear scale, and (b) a logarithmic scale for frequency. Reprinted from Morley *et al*. *Mechanisms of Ageing and Development* **19**: 21–6, copyright (1982), with permission from Elsevier Science.

Analysis of titres

Many serological tests, such as the haemagglutination test for rubella antibody, are based on a series of doubling dilutions, and the strength of the most dilute solution that provides a reaction is recorded. The results are called **titres**, and are expressed in terms of the strengths of the dilutions: 1/2, 1/4, 1/8, 1/16, 1/32, etc. For convenience, we will use the terminology more loosely, and refer instead to the reciprocals of these numbers, namely 2, 4, 8, 16, 32, etc., as titres. Titres tend to be positively skewed, and are therefore best analysed using a logarithmic transformation. This is accomplished most easily by replacing the titres with their corresponding dilution numbers. Thus titre 2 is replaced by dilution number 1, titre 4 by 2, titre 8 by 3, titre 16 by 4, titre 32 by 5, and so on. This is equivalent to taking logarithms to the base 2 since, for example, $8 = 2^3$ and $16 = 2^4$.

$$u = \text{dilution number} = \log_2 \text{titre}$$

All analyses are carried out using the dilution numbers. The results are then transformed back into the original units by calculating 2 to the corresponding power.

Example 13.4

Table 13.2 shows the measles antibody levels of ten children one month after vaccination for measles. The results are expressed as titres with their corresponding dilution numbers. The mean dilution number is $\bar{u} = 4.4$. We antilog this by calculating $2^{4.4} = 21.1$. The result is the **geometric mean titre** and equals 21.1.

$$\text{Geometric mean titre} = 2^{\text{mean dilution number}}$$

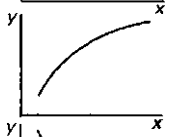
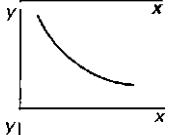
Table 13.2 Measles antibody levels one month after vaccination.

Child no.	Antibody titre	Dilution no.
1	8	3
2	16	4
3	16	4
4	32	5
5	8	3
6	128	7
7	16	4
8	32	5
9	32	5
10	16	4

13.3 CHOICE OF TRANSFORMATION

As previously mentioned, the logarithmic transformation is by far the most frequently applied. It is appropriate for removing positive skewness and is used on a great variety of variables including incubation periods, parasite counts, titres, dose levels, concentrations of substances, and ratios. There are, however, alternative transformations for skewed data as summarized in Table 13.3. For example, the **reciprocal transformation** is stronger than the logarithmic, and would be appropriate if the distribution were considerably more positively skewed than lognormal, while the **square root transformation** is weaker. Negative skewness, on the other hand, can be removed by using a **power transformation**, such as a square or a cubic transformation, the strength increasing with the order of the power.

Table 13.3 Summary of different choices of transformations. Those removing positive skewness are called group A transformations, and those removing negative skewness group B.

Situation	Transformation
Positively skewed distribution (group A)	
Lognormal	Logarithmic ($u = \log x$)
More skewed than lognormal	Reciprocal ($u = 1/x$)
Less skewed than lognormal	Square root ($u = \sqrt{x}$)
Negatively skewed distribution (group B)	
Moderately skewed	Square ($u = x^2$)
More skewed	Cubic ($u = x^3$)
Unequal variation	
s.d. proportional to mean	Logarithmic ($u = \log x$)
s.d. proportional to mean^2	Reciprocal ($u = 1/x$)
s.d. proportional to $\sqrt{}$mean	Square root ($u = \sqrt{x}$)
Non-linear relationship	Transform: y variable and/or x variable
	Group A (y) Group B (x)
	Group B (y) Group A (x)
	Group A (y) Group A (x)
	Group B (y) Group B (x)

There is a similar choice of transformation for making standard deviations more similar, depending on how much the size of the standard error increases with increasing mean. (It rarely decreases.) Thus, the logarithmic transformation is appropriate if the standard deviation increases approximately in proportion to the mean, while the reciprocal is appropriate if the increase is steeper, and the square root if it is less steep.

Table 13.3 also summarizes the different sorts of simple non-linear relationships that might occur. The choice of transformation depends on the shape of the curve and whether the y variable or the x variable is to be transformed.

13.4 *z*-SCORES AND REFERENCE CURVES

In this section we consider a different type of transformation; namely the use of ***z*-scores** to compare data against **reference curves** in order to improve their *interpretability*. Their most common use is for the analysis of **anthropometric data**. For example, an individual's weight and height cannot be interpreted unless they are related to the individual's age and sex. More specifically they need to be compared to the distribution of weights (or heights) for individuals of the same age and sex in an appropriate reference population, such as the NCHS/WHO* growth reference data.

Recall from Section 5.4 that a z-score expresses how far a value is from the population mean, and expresses this difference in terms of the number of standard deviations by which it differs. In the context here, a z-score is used to compare a particular value with the mean and standard deviation for the corresponding reference data:

$$z\text{-}score = \frac{x - \mu}{\sigma}$$

where x is the observed value, μ is the mean reference value† and σ the standard deviation of the corresponding **reference data**. A z-score is therefore a value from the *standard normal distribution*.

*NCHS/WHO growth reference data for height and weight of US children collected by the National Center for Health Statistics and recommended by the World Health Organization for international use.

†The NCHS/WHO reference curves were developed by fitting two separate half normal distributions to the data for each group. Both distributions were centred on the median value for that age. One distribution was fitted so that its upper half matched the spread of values above the median, and the other so that its lower half matched the spread of values below the median. The upper half of the first curve was then joined together at the median with the lower half of the second curve. This means that the z-score calculations use the median value for that age, and the standard deviation corresponding to either the *upper* or the *lower* half of the distribution for that age, depending on whether the observed value is respectively *above* or *below* the median.

The analysis can then be carried out with the calculated *z*-scores as the outcome variable. Such a *z*-score value will have the same interpretation regardless of the age or sex of the individual. Thus, for example, individuals with weight-for-age *z*-scores of −2 or below compare approximately with the bottom 2% of the reference population, since 2.3% of the standard normal curve lies below −2 (see Appendix A1). This interpretation is true whatever the ages of the individuals.

Example 13.5

An example of an analysis based on *z*-scores is given in Figure 13.5, which shows the mean weight-for-age *z*-scores (based on the NCHS/WHO growth curves) during the first 5 years of life for children in the Africa, Asia and Latin America/Caribbean regions. A mean *z*-score of zero would imply that the average weight of children in the region is exactly comparable to the average weight of American children of the same age in the NCHS/WHO reference population. A mean *z*-score above zero would imply that children in the region were on average heavier than their reference counterparts, while a mean *z*-score below zero implies that on average they are lighter. The curves in Figure 13.5 illustrate how in all three regions there is rapid growth faltering that starts between 3 and 6 months of age, and that by one year of age in all three regions the average child is very considerably underweight compared to

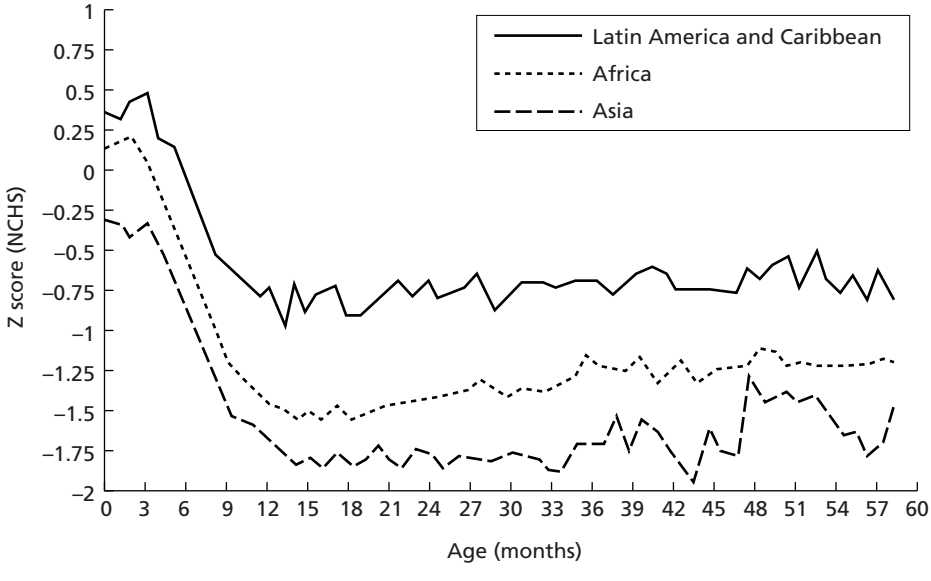


Fig. 13.5 Comparison of weight for age by region for children aged less than 5 years. Reprinted with permission from Shrimpton R, Victora CG, de Onis M, Lima RC, Bloessner M, Clugston G, Worldwide timing of growth faltering. *Pediatrics* 2001; **107**: E75

their counterparts in the reference population. It further shows that the level of disadvantage is most pronounced in Asia and least so in Latin America/Caribbean, with Africa in between.

See the report by the WHO Expert Committee on Physical Status (1995) for a detailed guide to the analysis and interpretation of anthropometric data.