PART E

STATISTICAL MODELLING

Previous parts of the book have discussed methods of analysis according to the different types of outcome (and exposure) variables. An understanding of what statistical method is appropriate given the type of data that have been collected is obviously crucial, but it is also important to realize that different statistical methods have much in common, so that an understanding of one method helps in understanding others. For example, the interpretation of confidence intervals and *P*-values follows the same logic, regardless of the particular situation in which they are derived. We have seen that computer output from different regression models is presented in a similar way, and issues such as testing hypotheses, examining interactions between exposure effects and selection of the most appropriate model also apply to all regression models.

In this part of the book we present statistical methods that apply to many types of exposure and outcome variables. We begin, in Chapter 28, by introducing likelihood: the concept that underlies most commonly used statistical methods. In Chapter 29 we consider general issues in regression modelling, including the use of likelihood ratio tests of hypotheses about the parameters of regression models.

Chapter 30 introduces methods that can be used when the usual model assumptions are violated: these provide a means of checking the robustness of results derived using standard methods. A common situation in which standard assumptions are violated is when data are clustered; that is when observations on individuals within a cluster tend to be *more similar* to each other than to individuals in other clusters. Failure to take account of clustering can lead to confidence intervals that are too narrow, and *P*-values that are too small. Chapter 31 introduces methods that are appropriate for the analysis of such data.

Chapter 32 focuses on how evidence can be summarized on a particular subject in order to make it accessible to medical practitioners and inform the practice of **evidence-based medicine**. In particular it covers systematic reviews of the medical literature, the statistical methods which are used to combine effect estimates from different studies (meta-analysis), and sources of bias in meta-analysis and how these may be detected.

Finally, in Chapter 33 we briefly describe the Bayesian approach to statistical inference.

This page intentionally left blank

CHAPTER 28

Likelihood

28.1 INTRODUCTION

In this chapter, we introduce the concept of **likelihood** and explain how **likelihood theory** provides the basis for a general approach to using data to yield estimates of **parameters** of interest. The idea that we use data to estimate parameters of interest using an underlying probability model is fundamental to statistics. This ranges from:

- simple models to estimate a single parameter of interest, based on assuming a normal, binomial or Poisson distribution for the outcome of interest. For example, estimating the risk of vertical transmission of HIV during pregnancy or childbirth, in HIV-infected mothers given antiretroviral therapy during pregnancy, is based on assuming a binomial distribution for the occurrence (or not) of vertical transmission, or a normal approximation to this binomial distribution;
- to multivariable regression models assuming a particular distribution for the outcome based on the values of a number of exposure variables. Such models relate the probability distribution of the outcome to the levels of the exposure variables via the values of one or more *parameters*. For example, in Example 24.2, we used Poisson regression to compare rates of myocardial infarction according to whether men in the Caerphilly study were current smokers or never/ex-smokers. The regression model had two *parameters*: the *log of the rate* in the never/ex-smokers, and the *log of the rate ratio* comparing current smokers with never/ex-smokers.

In most of the chapter, we will show how likelihood theory can be used to reproduce results that we derived earlier in the book using properties of the

normal distribution, and approximations to the normal distribution. The strength of the likelihood approach, however, lies in the way it can be generalized to any statistical model, for any number of parameters. It provides the basis for fitting logistic, Poisson and Cox regression models. For this reason it is of great importance in modern medical statistics.

This chapter is conceptually fairly sophisticated, and may be skipped at a first reading. An understanding of likelihood is not essential to the conduct of the majority of statistical analysis. However, this chapter does provide insights into understanding how regression models are fitted, the different ways that we can test hypotheses about the parameters of regression models, the meaning of some of the 'small print' items obtained on regression outputs, such as the iteration number, and why problems may be encountered. We recommend Clayton and Hills (1993), for a fuller explanation of the ideas presented here, and Royall (1997) for a discussion of different approaches to statistical inference based on likelihood.

28.2 LIKELIHOOD

Example 28.1

We will illustrate the idea of likelihood through an example, in which we are interested in estimating the risk of household transmission of tuberculosis (TB). We have tuberculin tested 12 household contacts of an index case of TB. Three of the twelve tested positive; the other nine tested negative. Using the notation introduced in Part C for binary outcomes, we have $d = 3$ and $h = 9$. The sample proportion, p equals 3/12 or 0.25. As always, we are not interested in this sample result in its own right but rather in what it tells us more generally about the risk of household transmission (π). Putting this another way, given that the sample proportion was 0.25, what can we deduce from this concerning the most likely value for π? Intuitively we would answer this question with $\pi = 0.25$, and we would be correct. We will now explain the mathematical basis for this, which can be extended to deriving estimates in more complicated situations.

The approach we use is to calculate the probability, or **likelihood**, of our observed result for different values of π: the likelihood gives a comparative measure of how compatible our data are with each particular value of π. We then find the value of π that corresponds to the largest possible likelihood. This value is called the **maximum-likelihood estimate (MLE)** of the parameter π.

$$\text{MLE} = \begin{array}{l}\text{the value of the parameter that } \textit{maximizes} \\ \text{the likelihood of the observed result}\end{array}$$

In this case, the likelihoods are calculated using the formula for the binomial distribution, described in Chapter 14. Figure 28.1 shows how the value of the likelihood varies with different values of π, and Table 28.1 shows the details of

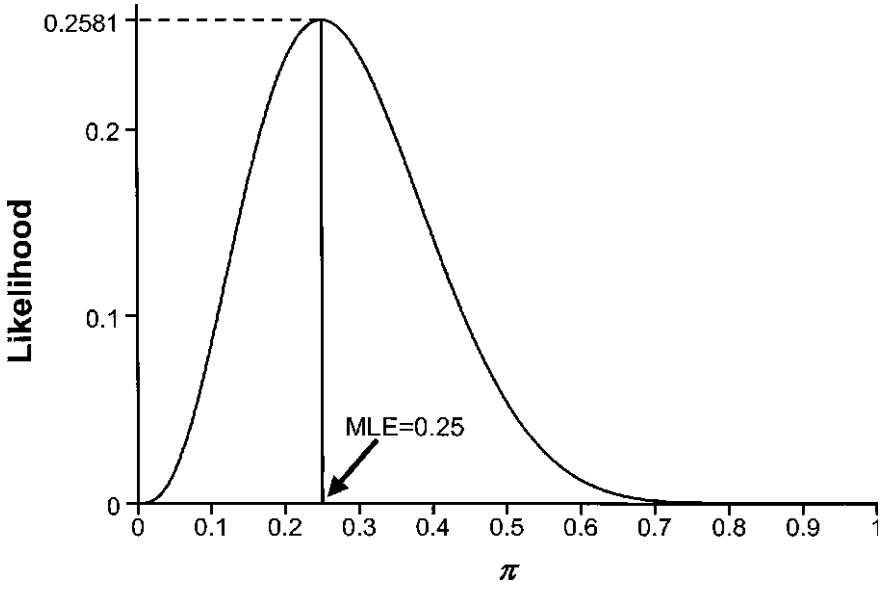

Fig. 28.1 Values of the likelihood for different values of π, if $d = 3$ and $h = 9$, showing that the maximum likelihood estimate is 0.25.

Table 28.1 Values of the likelihood of observing $d = 3$, $h = 9$ for different values of π.

Value of π	Likelihood of observed result $= \frac{12!}{3!9!}\pi^3 \times (1-\pi)^9$
0.1	$220 \times 0.1^3 \times 0.9^9 = 0.0852$
0.2	$220 \times 0.2^3 \times 0.8^9 = 0.2362$
0.25	$220 \times 0.25^3 \times 0.75^9 = 0.2581$
0.3	$220 \times 0.3^3 \times 0.7^9 = 0.2397$
0.4	$220 \times 0.4^3 \times 0.6^9 = 0.1419$
0.6	$220 \times 0.6^3 \times 0.4^9 = 0.0125$

the calculations for a few selected values. It can be seen that the likelihood increases as π increases, reaches a maximum when $\pi = 0.25$, and then decreases. Thus, our maximum likelihood estimate is MLE = 0.25, agreeing with our original guess.

This result can be confirmed mathematically. The MLE can be derived by differentiating the binomial likelihood $\pi^d \times (1-\pi)^h$ to find the value of π that maximizes it. The result is $d/(d+h)$ *or* d/n, which in this example equals 3/12 or 0.25.

In simple situations, such as the estimation of a single mean, proportion or rate, or the comparison of two means, proportions or rates, the MLE is given by the sample value for the parameter of interest (in other words the usual estimate). This is the case here; the MLE for the within-household risk of TB transmission equals

the proportion who tested tuberculin positive in the sample of 12 household contacts of the index case.

28.3 LIKELIHOOD RATIOS AND SUPPORTED RANGES

As well as concluding that 0.25 is the most likely value for the true probability π of the risk of household transmission of TB in our example, it is useful to know what other values of π are compatible with the data. We now describe how to use **likelihood ratios**, or more specifically their *logarithmic* equivalent, to give us a range of likely values for the population parameter (in this case π), which we wish to estimate.

In our example, the maximum likelihood equals 0.2581, and the corresponding maximum likelihood estimate is $\pi = 0.25$. The likelihood for any other value of π will be less than this. How much less likely is assessed using the **likelihood ratio (LR)**:

$$\text{Likelihood ratio (LR)} = \frac{\text{Likelihood for } \pi}{\text{Likelihood at the MLE}}$$

Figure 28.2 shows how the likelihood ratio varies across the range of possible values and Table 28.2 shows the details of the calculation for a few selected values

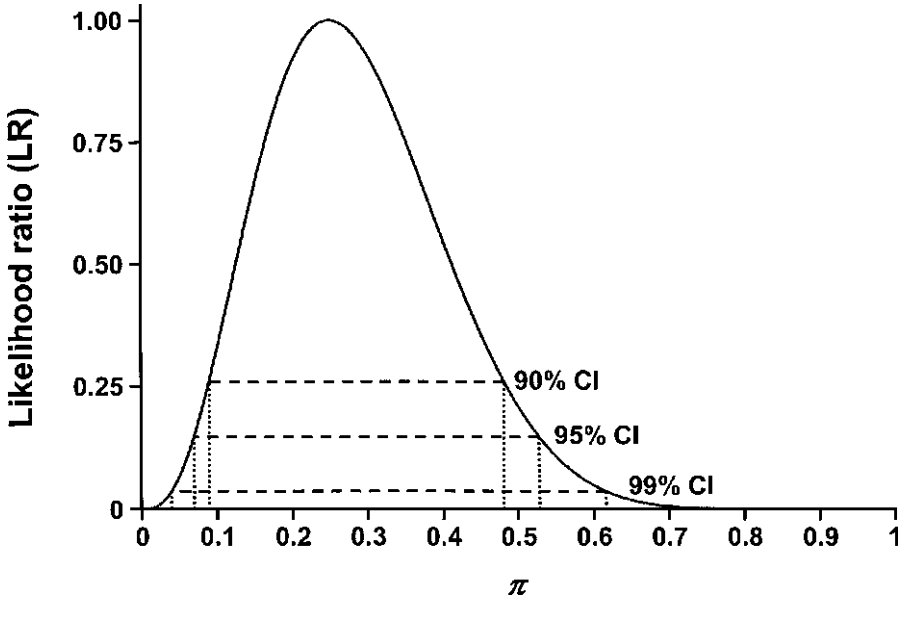

Fig. 28.2 Values of the likelihood ratio for different values of π, if $d = 3$ and $h = 9$. The horizontal dashed lines show the supported ranges corresponding to 90%, 95% and 99% confidence intervals (see Table 28.3), and the dotted vertical lines show the corresponding confidence limits.

Table 28.2 Values of the likelihood of observing $d = 3$, $h = 9$, and corresponding likelihood ratio, for different values of π.

Value of π	Likelihood	Likelihood ratio
0.1	0.0852	$0.0852/0.2581 = 0.3302$
0.2	0.2362	$0.2362/0.2581 = 0.9151$
0.25 (MLE)	0.2581	$0.2581/0.2581 = 1$
0.3	0.2397	$0.2397/0.2581 = 0.9287$
0.4	0.1419	$0.1419/0.2581 = 0.5498$
0.6	0.0125	$0.0125/0.2581 = 0.0484$

of π. By definition, the likelihood ratio equals 1 for the MLE (in this case for $\pi = 0.25$) and less than one for all other values. The shape of the curve of the likelihood ratio is *exactly* the same as that of the likelihood in Figure 28.1, since we have simply divided the likelihood by a constant amount, namely the maximum likelihood, which in this case equals 0.2581.

The likelihood ratio provides a convenient measure of the amount of support for a particular value(s) of π. The likelihood ratios for π equal to 0.2 or 0.3 are close to 1, suggesting that these values are almost as compatible with the observed data as the MLE. In contrast, the likelihood ratio for π equal to 0.6 is very small; it is therefore much less likely that the within-household transmission rate for TB is as high as 0.6. The conclusion is less immediately clear for likelihood ratios in between, such as a ratio of 0.3302 for π equal to 0.1 or 0.5498 for π equal to 0.4.

By choosing a cut-off value for the likelihood ratio, we can derive a **supported range** of parameter values. We classify values of π with likelihood ratios *above* the cut-off as supported by the data, and those with likelihood ratios *below* the cut-off as not supported by the data. This concept of a **supported range** of values is intuitively simple; the choice of the cut-off value is the critical issue. Although supported ranges arise from a different philosophical basis to confidence intervals, the two turn out to be closely linked. We will show below that, providing the sample size is sufficiently large, different choices of cut-off for the likelihood ratio correspond to different choices of confidence level, as illustrated in Figure 28.2. For example, a likelihood ratio of 0.1465 gives a supported range that approximately coincides with the 95% confidence interval for π, calculated in the usual way (see Table 28.3).

28.4 CONFIDENCE INTERVALS BASED ON THE LOG LIKELIHOOD RATIO AND ITS QUADRATIC APPROXIMATION

We work with the *logarithm* of the likelihood ratio to derive confidence intervals, rather than the likelihood ratio itself because, provided the sample size is sufficiently large, the log LR can be approximated by a quadratic equation, which is easier to handle mathematically than the likelihood ratio. Using the rules of logarithms (see the box on p. 156):

$$\log(\text{LR}) = \log(\text{likelihood for } \pi) - \log(\text{likelihood at the MLE})$$

Abbreviating this formula by using the letter L to denote log likelihood gives:

$$\log(\text{LR}) = \text{L}(\pi) - \text{L}(\text{MLE})$$

Note that, as in earlier parts of this book, we use logarithms to the base e (natural logarithms); see Section 13.2 for an explanation of logarithms and the exponential function.

The log(LR) corresponds to a *difference* in log likelihoods. Its maximum occurs at the MLE and equals zero. Figure 28.3(a) shows the log(LR) for the data in Example 28.1 on within-household transmission of TB. Figure 28.3(b) shows how the shape of the curve would change for a larger sample size (120 instead of 12), but with the same MLE of 0.25. The dashed lines in Figure 28.3 show the best quadratic approximations to these particular log likelihoods. For the small sample size in Figure 28.3(a) the quadratic approximation has a relatively poor fit, while for the larger sample size in Figure 28.3(b) there is a close fit between the log likelihood and the quadratic approximation.

The **quadratic approximation** is chosen to meet the log(LR) at the MLE and to have the *same curvature* as the log(LR) at this point. It is symmetrical about this point and its maximum value is zero. It can be shown that its equation can be written in the following way:

$$\text{Log(LR)} = -\frac{1}{2}\left(\frac{\text{MLE} - \theta}{S}\right)^2$$

where θ represents the parameter that we wish to estimate and $-1/S^2$ is the curvature at the maximum. In our example θ would be π, the within-household risk of transmission of TB. In Example 6.1, θ would be μ, the mean sprayable surface area of houses that we wished to estimate in order to be able to calculate how much insecticide would be needed to spray the whole area as part of the malaria control programme. In this case, we had a quantitative outcome which we assumed was normally distributed.

The quadratic approximation plays a key role in parameter estimation because:

1 In simple situations, such as the estimation of a single mean, proportion or rate, or the comparison of two means, proportions or rates:
   - the MLE equals the sample value for the parameter of interest (see Section 28.2);
   - the denominator S equals the usual estimate of the **standard error**.

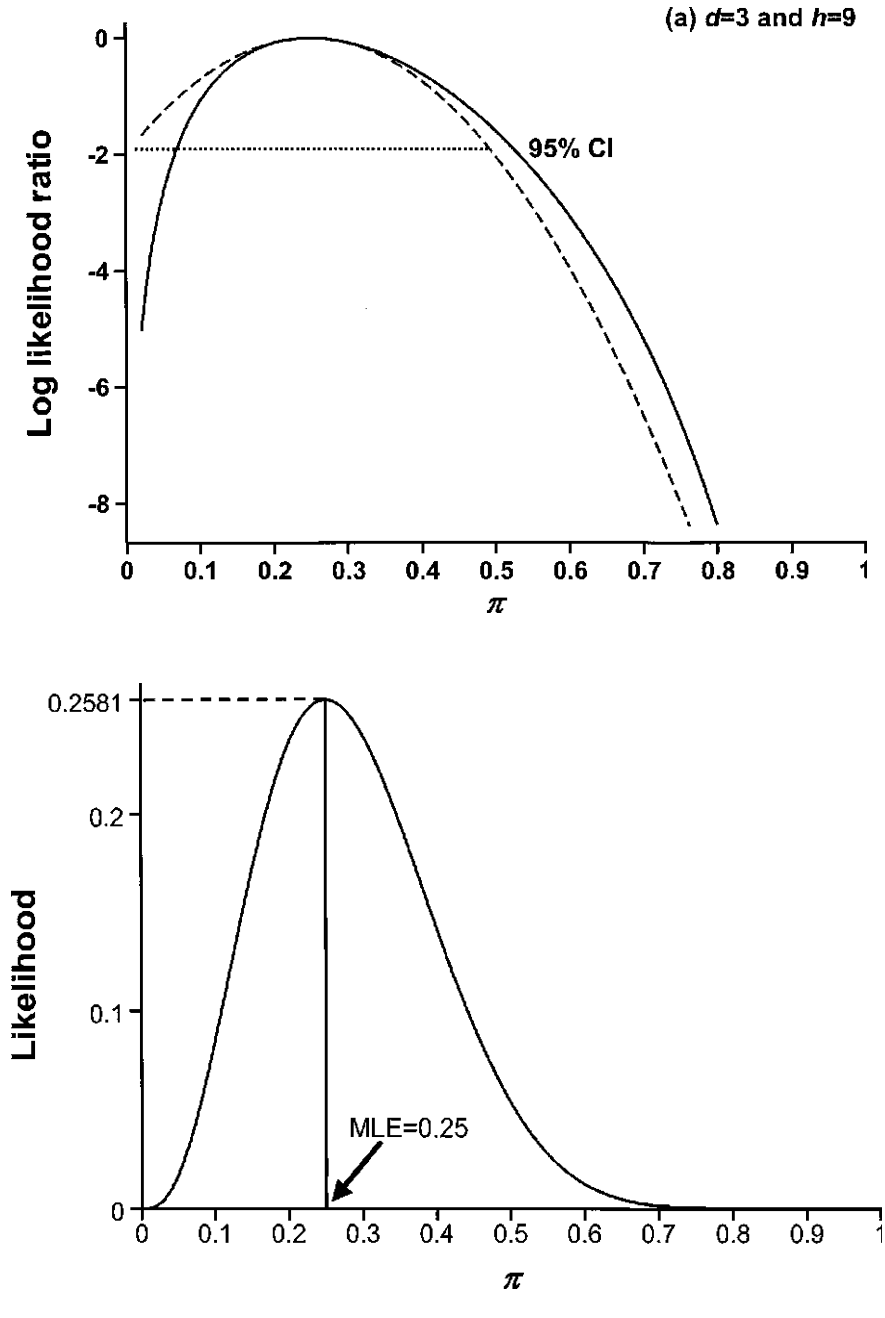

Fig. 28.3 Values of the likelihood ratio for different values of π, if (a) $d = 3$ and $h = 9$, or (b) $d = 30$ and $h = 90$. The dashed lines show the best quadratic approximations to the log likelihood ratio curves, fitted at the MLE ($\pi = 0.25$) and the dotted lines show the 95% confidence intervals based on the quadratic approximations.

2 When the underlying distribution is **normal**, the quadratic equation gives an *exact* fit to the log(LR).
3 When the sample size is sufficiently large, then the quadratic equation gives a *close* fit to the log(LR), regardless of the underlying probability distribution. This arises from the **Central Limit Theorem** (see Section 5.2), which shows that the normal distribution provides a good approximation to the *sampling distribution* of the parameter of interest, whatever its underlying distribution, provided that the sample size is sufficiently large.
4 The closest quadratic approximation to the log(LR) can be found using a process known as iteration, as explained in Section 28.6. This involves calculating the likelihood ratio, and its log, only at selected points of the curve. It avoids the need to calculate the whole of the curve.

These facts together mean that the quadratic approximation provides a method to derive MLEs and corresponding confidence intervals that avoids the need for complicated mathematics, and that works in situations with complex underlying distributions, as well as giving the same results as standard methods in simple situations.

Since fitting a quadratic approximation to the log(LR) is equivalent to using a normal approximation for the sampling distribution for the parameter θ that we wish to estimate, the **95% confidence interval** based on the quadratic approximation must be:

$$95\%\ \text{CI} = \text{MLE} - 1.96 \times S \text{ to } \text{MLE} + 1.96 \times S$$

Link between confidence intervals and supported ranges

At the end of Section 28.3, we noted that a likelihood ratio of 0.1465 gives a supported range that approximately coincides with the 95% confidence interval. We will now derive this link.

Since the quadratic approximation for log(LR) is:

$$\text{Log(LR)} = -\frac{1}{2}\left(\frac{\text{MLE} - \theta}{S}\right)^2$$

And since,

$$\text{MLE} - \text{lower } 95\%\ \text{CL} = \text{MLE} - (\text{MLE} - 1.96 \times S) = 1.96S$$

and

$$\text{MLE} - \text{upper } 95\%\ \text{CL} = \text{MLE} - (\text{MLE} + 1.96 \times S) = -1.96S$$

the **values of the log(LR) curve at the 95 % confidence limits** (CL) are both:

$$\text{Log(LR) for 95\% CI} = -\frac{1.96^2}{2} = -1.9208$$

since the S's in the numerator and denominator cancel out, and since $(-1.96)^2 = 1.96^2$. Antilogging this gives the cut-off value of the **likelihood ratio corresponding to the 95 % confidence interval**:

$$\text{LR for 95\% CI} = e^{-1.9208} = 0.1465$$

Table 28.3 summarizes the cut-off values of the likelihood ratio and its logarithm corresponding to 90 %, 95 % and 99 % confidence intervals. Note that there is only a close agreement between standard confidence intervals and supported ranges based on these cut-offs when the quadratic approximation gives a close fit to the log(LR).

Table 28.3 Cut-off values for the likelihood ratio, and its logarithm, corresponding to 90%, 95% and 99% confidence intervals, assuming that the underlying distribution is normal or approximately normal.

	90% CI	95% CI	99% CI
% point of normal distribution	1.6449	1.96	2.5763
Cut-off value for log(LR)	−1.3529	−1.9208	−3.3187
Cut-off value for LR	0.2585	0.1465	0.0362

Information and standard error

The quantity $1/S^2$ (the multiplier of $\frac{1}{2}$ $(\text{MLE} - \theta)^2$ in the quadratic approximation) is known as the **information** in the data. The larger the value for the information, the more sharply curved are the log(LR), its quadratic approximation, the likelihood ratio and the likelihood curves. The more information that the data contain about the parameter, the smaller is its standard error, the more precise is our estimate, and the narrower is the confidence interval.

28.5 LIKELIHOOD IN THE COMPARISON OF TWO GROUPS

Example 28.2

So far we have described the principles of likelihood in the simplest context of a single sample and a single parameter to be estimated. We will now illustrate its

extension to the comparison of two exposure groups, using the data from the Guatemala morbidity study presented in Table 23.1. This table compared the incidence rate, $\lambda_1 = 33/355$, of lower respiratory infections among children aged less than 5 years living in poor housing conditions, to the rate, $\lambda_0 = 24/518$ among those living in good housing. The rate ratio was:

$$\text{rate ratio}\,(\theta) = \lambda_1/\lambda_0 = \frac{33/355}{24/518} = 2.01$$

As explained in Chapter 24 on Poisson regression, we can re-express this as:

$$\text{rate in exposed group} = \text{rate in unexposed group} \times \text{exposure rate ratio}$$

giving us the basis for a model which expresses the rate in each group in terms of two *model parameters*. These are:

- the *baseline rate*, λ_0, in the unexposed group;
- the *exposure rate ratio*, θ.

Applying the likelihood approach means that we want to find the most likely values of these two parameters given the observed data. In other words we want to find their *maximum likelihood estimates* (MLEs). It can be shown that:

1 Using the distribution of the numbers of infections in each of the two groups, we can derive a formula for the log likelihood (L) of the observed data for various combinations of the two parameters. This is:

$$\text{L} = (d_0 + d_1)\log(\lambda_0) + d_1\log(\theta) - \lambda_0 T_0 - \theta\lambda_0 T_1 + \text{constant}$$

where d_1 and d_0 are the number of observed infections and T_1 and T_0 are the child-years of follow up in the exposed (poor housing) and unexposed (good housing) groups respectively.

2 As we have two parameters we have a log likelihood surface rather than a curve. This can be thought of as like the map of a hill; the two parameters correspond to the two axes of the map, and contours on the hill correspond to values of the log likelihood ratio. We want to find the MLEs (equivalent to finding the peak of the hill) and the curvature at this point in order to fit a three-dimensional quadratic approximation to the surface (of the hill).

3 In this case it is possible to show that the value of λ_0 that maximizes the log likelihood is:

$$\lambda_0 = (d_0 + d_1)/(T_0 + \theta T_1)$$

and that substituting this formula for λ_0 into the equation for log likelihood and rearranging it gives:

$$\text{L} = d_1\log\left(\frac{\theta T_1}{T_0}\right) - (d_0 + d_1)\log\left(1 + \frac{\theta T_1}{T_0}\right) + \text{constant}$$

This is called the **profile log likelihood** for θ. In our hill analogy, it is equivalent to slicing through the hill at its peak and working with the resulting cross-section.

4 Figure 28.4 shows the profile log likelihood ratio for various values of the rate ratio using this re-expression. Note that the rate ratio is plotted on a **log scale**, and that doing this makes the log likelihood ratio curve close to a quadratic.

5 The log likelihood (and corresponding likelihood) is maximized when

$$\lambda_0 = 24/518, \text{ the observed rate in the unexposed group;}$$
$$\theta = \lambda_1/\lambda_0 = 2.01, \text{ the observed rate ratio}$$

These MLEs are the same as the estimates obtained directly from the data in Example 23.1.

6 Because the rate ratio is plotted on a log scale, the equation of the quadratic approximation is:

$$\text{Log(LR)} = -\frac{1}{2}\left(\frac{\log(\text{MLE}) - \log(\theta)}{S}\right)^2, \text{ where } S = \text{s.e. of the } \textit{log} \text{ rate ratio}$$

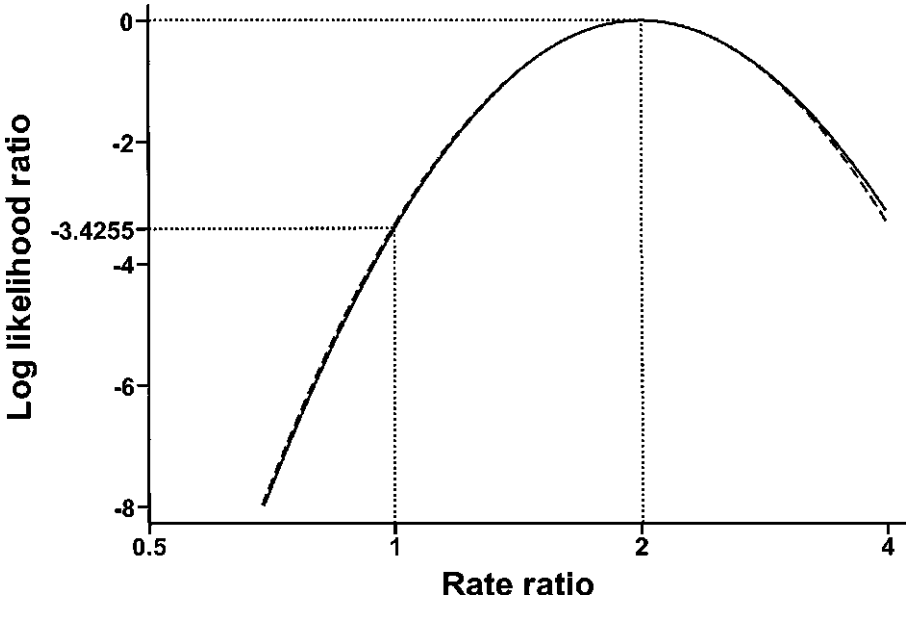

Fig. 28.4 Profile log likelihood ratios for the rate ratio (plotted on a log scale), for the data on respiratory infections in Guatemalan children. The dashed line shows the best quadratic approximation to the log likelihood ratio at the maximum, and the dotted lines show the values of the log likelihood ratio corresponding to the null value (1) and the maximum-likelihood estimate (2.01) of the rate ratio.

7 The 95% confidence interval is calculated from the MLE and the standard error of the log(rate ratio), using an *error factor*, as explained in Chapter 23. In this example,

$$S = \sqrt{(1/d_0 + 1/d_1)} = \sqrt{(1/33 + 1/24)} = 0.2683, \text{ giving}$$

$$\text{EF} = \exp(1.96 \times 0.2683) = 1.69$$

$$\text{Thus, } 95\% \text{ CI} = 2.01/1.69 \text{ to } 2.01 \times 1.69 = 1.19 \text{ to } 3.39$$

With 95% confidence, the rate of acute lower respiratory infections among children living in poor housing is between 1.19 and 3.39 times the rate among children living in good housing.

28.6 LIKELIHOOD IN MORE COMPLICATED MODELS

In most of this chapter, we show how likelihood theory can be used to reproduce results that we derived earlier in the book using properties of the normal distribution, and approximations to the normal distribution. The strength of the likelihood approach, however, lies in the way it can be generalized to any statistical model, for any number of parameters.

Thus the likelihood approach is used to derive maximum likelihood estimates (MLEs) and standard errors of the parameters in a regression model. Since the MLE for any one parameter will depend on the values of the other parameters, it is usually not possible to write down equations for what each of the MLEs will be. Instead, they are fitted by a computer program using a process known as **iteration**:

1 This starts with a guess for the MLEs of the parameters; for example, some programs use the null values corresponding to no effects of the exposure parameters on the outcome as the starting point.
2 Next, the value of the log likelihood is calculated using these 'guesstimates'.
3 The value of each of the parameters is then perturbed in both directions, and the values of the log likelihood calculated to obtain the gradient and curvature of the log likelihood curve at this point.
4 The gradient and curvature are then used to fit the best (multi-dimensional) quadratic approximation to the log likelihood curve at this particular point.
5 The maximum of the fitted quadratic is then located.
6 The whole process is then repeated using this maximum as the best guess for the MLEs.
7 The iteration stops when subsequent steps yield the same values for the guess for the MLEs. The fit is said to have **converged**. Some programs will record the number of iteration steps it required to obtain this convergence.
8 Occasionally the program fails to achieve convergence. The main causes of this are:
   - insufficient data to support the estimation of the number of parameters there are in the model;

- profile log likelihood(s) that are very non-quadratic.

Logistic, Poisson and Cox regression all use logarithmic transformations of the parameters in order to make the profile log likelihoods approximately quadratic in form. The likelihood for simple and multiple regression is based on the normal distribution, and has an exact quadratic form; the maximum likelihood estimates obtained are equivalent to those obtained using the least squares approach (see Chapters 10 and 11).

28.7 USING LIKELIHOOD FOR HYPOTHESIS TESTING

We will now describe how the likelihood approach can be used to provide a general means of hypothesis testing. As explained in Chapter 8, a hypothesis test is based on calculating a **test statistic** and its corresponding ***P*-value** (also known as a **significance level**), in order to assess the strength of the evidence against the **null hypothesis** (of *no* association between exposure and outcome in the population). The *smaller* the *P*-value, the *stronger* is the evidence against the null hypothesis.

There are three different types of tests based on the log likelihood:

1 The **likelihood ratio test**, based on the value of the log likelihood ratio at the null value of the parameter.
2 The **Wald test**, which is similar but uses the value of the *fitted quadratic approximation* to the log likelihood ratio at the null, rather than the actual value of the log likelihood ratio at this point.
3 The **score test**, based on fitting an alternative quadratic approximation to the log likelihood ratio, which has the same gradient and curvature at the *null* value of the parameter, rather than at the MLE.

Likelihood ratio tests

The likelihood ratio test is based on the value of the log likelihood ratio at the null value of the parameter, using the fact that it can be shown that *providing the log likelihood ratio curve is close to a quadratic*:

$$-2 \times \log(\text{likelihood ratio}) \text{ has a } \chi^2 \text{ distribution with 1 d.f.}$$

We therefore work with minus twice the log(likelihood ratio); this is called the **likelihood ratio statistic (LRS)**:

$$\text{LRS} = -2 \times \log(\text{LR}) = -2 \times (\text{L}_{\text{null}} - \text{L}_{\text{MLE}}) \text{ is } \chi^2 \text{ with 1 d.f.}$$

In Example 28.2, based on the data from the Guatemalan morbidity study presented in Table 23.1, we found that the MLE for the rate ratio of the incidence

of lower respiratory infections among children living in poor compared to good housing conditions was 2.01. We noted that the formula for the profile log likelihood shown in Figure 28.4 is:

$$\mathrm{L} = d_1 \log\left(\frac{\theta T_1}{T_0}\right) - d \log\left(1 + \frac{\theta T_1}{T_0}\right) + \text{constant}$$

Calculating this for $\theta = 1$ (null) and $\theta = 2.01$ (MLE) gives:

$$\begin{aligned}
\mathrm{L}_{\text{null}} &= 33 \times \log\left(\frac{1 \times 355}{518}\right) - 57 \times \log\left(1 + \frac{1 \times 355}{518}\right) + \text{constant} \\
&= (33 \times -0.37786) - (57 \times 0.52196) + \text{constant} = -42.2211 + \text{constant} \\
\mathrm{L}_{\text{MLE}} &= 33 \times \log\left(\frac{2.01 \times 355}{518}\right) - 57 \times \log\left(1 + \frac{2.01 \times 355}{518}\right) + \text{constant} \\
&= (33 \times 0.32028) - (57 \times 0.86605) + \text{constant} = -38.7956 + \text{constant}
\end{aligned}$$

The difference between these is the log(LR):

$$\mathrm{L}_{\text{null}} - \mathrm{L}_{\text{max}} = -42.2211 + 38.7956 = -3.4255$$

This is shown in Figure 28.4, in which the values of the log likelihood ratio at the null value ($\theta = 1$) and the MLE ($\theta = 2.01$) are depicted by the horizontal dotted lines.

The likelihood ratio statistic is:

$$\mathrm{LRS} = -2 \times (\mathrm{L}_{\text{null}} - \mathrm{L}_{\text{max}}) = -2 \times -3.4255 = 6.8510$$

The corresponding *P*-value, derived from the χ^2 distribution with 1 d.f., is $P = 0.0089$. There is therefore good evidence against the null hypothesis, suggesting that poor housing conditions did increase the rate of respiratory infections among the Guatamalan children.

Wald tests

The Wald test is similar to the likelihood ratio test, but is based on the value of the *fitted quadratic approximation* to the log likelihood ratio at the null value of the parameter of interest, rather than the actual value of the log likelihood ratio at this point. Recall from Section 28.4 that the quadratic approximation to the log likelihood ratio is of the form:

$$\mathrm{Log(LR)}_{\text{quad}} = -\frac{1}{2}\left(\frac{\mathrm{MLE} - \theta}{S}\right)^2$$

The **Wald test likelihood ratio statistic** based on the quadratic approximation is therefore:

$$\text{LRS}_{\text{Wald}} = -2 \times \log(\text{LR})_{\text{quad}} = \left(\frac{\text{MLE} - \theta_{\text{null}}}{S}\right)^2 = \left(\frac{\text{MLE}}{S}\right)^2, \text{ if } \theta_{\text{null}} = 0$$

For the data in Example 28.2 (and 23.1), the quadratic approximation to the log(LR) has been fitted using the log(rate ratio). Therefore:

$\theta = \log(\text{rate ratio})$
$\text{MLE} = \log(2.01) = 0.6963$
$S = 0.2683$ (see Section 28.5 above)

$$\text{LRS}_{\text{Wald}} = \left(\frac{0.6963}{0.2683}\right)^2 = 6.7352$$

$P = 0.0094$ (derived from χ^2 with 1 d.f.)

In this example, the Wald and likelihood ratio tests yield very similar results, as the quadratic approximation gives a close fit to the log likelihood ratio curve.

More commonly, the Wald test is carried out as a z-test, using the *square root* of the likelihood ratio statistic. This has a particularly convenient form:

$$\textbf{Wald statistic}, \; z = \frac{\text{MLE}}{S}, \text{ if } \theta_{\text{null}} = 0$$

and follows a standard normal distribution, since a χ^2 distribution with 1 d.f. is equivalent to the square of a standard normal distribution. This is the basis for the Wald tests described for logistic regression (Chapter 19), Poisson regression (Chapter 24) and Cox regression (Chapter 27).

For the data in Example 28.2, this formulation gives:

$$z = \frac{0.6963}{0.2683} = 2.5952 \text{ (equivalent to } \sqrt{6.7352})$$

As before, $P = 0.0094$.

Score tests

Much of the reasoning in this chapter has derived from fitting a quadratic approximation to the log likelihood ratio, chosen to have the same value and curvature at the MLE. The **score test** uses an alternative quadratic approximation, chosen to have the same value, gradient and curvature as the log likelihood ratio

at the *null* value of the parameter rather than at its MLE. Its form is similar to that of the log likelihood ratio and Wald tests:

$$\text{Score test} = -2 \times \log(\text{LR})_{\text{quad fitted at null}} = \frac{U^2}{V}$$
$$\text{where } U = \text{gradient and } V = -\text{curvature of the fitted } \log(\text{LR}) \text{ at } \theta_{\text{null}}$$

The Mantel–Haenszel statistics derived in Chapters 18 and 23 are of this form:

$$\chi^2_{\text{MH}} = \frac{U^2}{V}$$

and are score tests. U, the gradient of the log likelihood at the null value of the parameter, is also known as the **score**, and V (minus the curvature) is also known as the **score variance**. The standard chi-squared statistic (see Chapter 17)

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

can also be shown to be a special form of the score test.

Choice of method

All three methods described in this section for calculating a *P*-value are *approximate*. The exception is the special (and unusual) case when the parameter of interest is the mean, μ, for a normal distribution, for which we *know* the standard deviation, σ. In this instance, the three methods coincide, as the log likelihood ratio is *exactly* quadratic, and yield an *exact P*-value.

The three methods will give quite different answers unless the quadratic approximations provide a good fit to the log likelihood ratio curve over the region of the curve between the MLE and the null value. In general it is possible to get a reasonably close fit provided the sample size is sufficiently large, and provided an appropriate scale is used for the parameter(s) of interest. In particular, for odds, rates, odds ratios and rate ratios, it is generally preferable to use a **logarithmic transformation**, as was done in Example 28.2.

The values of the Wald and score tests are both derived from the quadratic approximation, which is influenced by the particular scale used for the parameter. Their values will therefore depend on what, if any, transformation is used. In contrast, the likelihood ratio test yields the same results whatever scale is used for the parameter of interest, since a change of scale simply changes the shape of the log(LR) curve in the horizontal direction, but does not affect the height of the

curve, or the relative heights between two values of the parameter. This is a considerable advantage.

However, if the three methods yield very different results, even after using an appropriate scale for the parameter(s), then it is usual to advise the use of exact *P*-values (see Clayton & Hills, 1993, for details), although these are not without their own difficulties.

Note that when the MLE and the null values are far apart, all three methods will always yield very small *P*-values. Thus, although it may not prove possible to obtain good quadratic approximations, and although the *P*-values may therefore differ numerically, this is unlikely to substantially affect the conclusions.

28.8 LIKELIHOOD RATIO TESTS IN REGRESSION MODELS

Hypothesis testing in regression models can be carried out using either **Wald tests** or **likelihood ratio tests**. We favour **likelihood ratio tests** for all but the simplest of cases, for the following reasons:

- the lack of dependence of the likelihood ratio statistic on the scale used for the parameter(s) of interest;
- the ease with which the calculation and interpretation of likelihood ratio statistics can be carried out in more complex situations, as described below;
- in contrast, although Wald tests are directly interpretable for exposure variables which are represented by a *single* parameter in the regression model (see Examples 19.1 and 24.1), they are less useful for a categorical variable, which is represented by a series of indicator variables in the regression model (see Section 29.4).

The likelihood ratio test described above for a single exposure is a special case of a more general likelihood ratio test that applies to more complex situations involving several model parameters. An example is in regression modelling where we have estimated the effect of a categorical exposure variable using k indicator variables and wish to test the null hypothesis that the exposure has no association with the outcome. In such situations we wish to test the joint null hypothesis that k parameters equal their null values. The likelihood ratio test is based on comparing the log likelihoods obtained from fitting the following two models:

1 L_{exc}, the log likelihood of the model *excluding* the parameter(s) to be tested;

2 L_{inc}, the log likelihood of the model *including* the parameter(s) to be tested.

Then the **likelihood ratio statistic (LRS)** has a χ^2 distribution with degrees of freedom equal to the number of parameters *omitted* from the model:

$$\text{LRS} = -2 \times \log(\text{LR}) = -2 \times (L_{exc} - L_{inc}) \text{ is } \chi^2 \text{ with } k \text{ d.f.}$$

Thus L_{inc} is the value of the log likelihood when all parameters equal their MLEs, and L_{exc} the value of the log likelihood when the k chosen parameters equal their

null values and the other parameters equal their MLEs for the *restricted* model, excluding these parameters.

The likelihood ratio can be used to compare any two models where one is a restricted form of the other. Its use in regression modelling will be described in detail in Chapter 29.

CHAPTER 29

Regression modelling

29.1 INTRODUCTION

In previous chapters we have described simple and multiple linear regression for the analysis of numerical outcome variables, logistic regression for the analysis of binary outcome variables, and Poisson and Cox regression for the analysis of rates and survival data from longitudinal studies, as summarized in Table 29.1. We have shown how all these types of regression modelling can be used to examine the effect of a particular exposure (or treatment) on an outcome variable, including:

- Comparing the levels of an outcome variable in two exposure (or treatment) groups.
- Comparing more than two exposure groups, through the use of indicator variables to estimate the effect of different levels of a categorical variable, compared to a baseline level (see Section 19.4).

- Estimating a linear (or dose–response) effect on an outcome of a continuous or ordered categorical exposure variable.
- Controlling for the confounding effect of a variable by including it together with the exposure variable in a regression model. We explained that this assumed that there was no interaction (effect modification) between the exposure and confounding variables. That is, we assumed that the effect of each variable on the outcome was the same regardless of the level of the other.

In this chapter, we focus on general issues in the choice of an appropriate regression model for a particular analysis. These are:

- Understanding the similarities and differences between the different types of regression models.
- Deciding between different expressions of the outcome variable, and their implication for the type of regression model.
- *Hypothesis testing* in regression models.
- Investigating *interaction* (*effect modification*) between two or more exposure variables, and understanding its implications.
- Investigating whether an exposure has a *linear* (*dose–response*) effect on the outcome variable.
- Understanding the problems caused when exposure and/or confounding variables are highly correlated. This is known as *collinearity*.
- Making the final choice of exposure/confounding variables for inclusion in the regression model.

29.2 TYPES OF REGRESSION MODEL

The different types of regression models described in this book are summarized in Table 29.1. It is useful to distinguish between:

- *Simple and multiple linear regression models*, in which the outcome variable is numerical, and whose general form for the effects of p exposure variables is:

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

  These are known as **general linear models**. The quantity 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 β's are the **regression coefficients** associated with the p exposure variables.
- *All other types of regression models*, including logistic, Poisson and Cox regression, in which we model a *transformation* of the outcome variable rather than the outcome itself. For example, in logistic regression we model the *log of the odds of the outcome*. Apart from this transformation, the general form of the model is similar to that for multiple regression:

$$\log \text{odds of outcome} = \log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

Table 29.1 Summary of the main regression models described in Parts B to D of this book.

	Regression model				
Type of outcome variable	Type	Chapter	Link function	Measure of exposure effect	Effects
Numerical	Linear (Simple/Multiple)	10/11	Identity	Mean difference	Additive
Binary	Logistic	19	Logit	Odds ratio	Multiplicative
Matched binary	Conditional logistic	21	Logit	Odds ratio	Multiplicative
Time to binary event	Poisson	24	Log	Rate ratio	Multiplicative
Time to binary event	Cox	27	Log	Hazard ratio	Multiplicative

These regression models are known as **generalized linear models**. The linear model for the exposure variables is said to be related to the outcome via a **link function**. For logistic regression, the link function is the logit (log odds) function, and for Poisson and Cox regressions, it is the logarithmic function.

Note that multiple regression is a special case of a generalized linear model in which the link function is the *identity* function $f(y) = y$.

- *Conditional regression models*, such as conditional logistic regression and Cox regression. These are special cases of generalized linear models in which estimation is based on the distribution of exposures within case–control strata or within risk sets. Likelihoods (see Chapter 28) for these models are known as *conditional likelihoods*.

All regression models are fitted using the maximum likelihood approach described in Chapter 28. The estimates obtained for the regression coefficients are called **maximum-likelihood estimates**. There are two important differences worth noting between multiple regression and the other types of generalized linear models:

1 Multiple regression models assume that the effect of exposures combine in an *additive* manner. In all the other generalized linear models discussed in this book it is a *log* transformation of the outcome (odds, rate or hazard) that is related to the linear predictor. This means that exposure effects are *multiplicative* (see the detailed explanation for logistic regression in Section 20.2) and that results of these models are most easily interpreted on the ratio scale.

2 Since multiple linear regression is based on the normal distribution, its log likelihood has an *exact* quadratic form (see Section 28.4). This means that Wald tests and likelihood ratio tests give identical results (see Sections 28.7 and 29.4). It also means that estimates obtained using maximum-likelihood are identical to those obtained using least-squares as described in Chapters 10 and 11.

29.3 DECIDING HOW TO EXPRESS THE OUTCOME VARIABLE

It is often the case that we have a choice of which regression model to use, depending on how the outcome variable is expressed. For example, blood pressure may be expressed as a continuous, ordered categorical or binary variable, in which case we would use linear, ordinal logistic or logistic regression respectively. Similarly, a study of factors influencing the duration of breastfeeding could be analysed using Poisson or Cox regression, or using logistic regression by defining the outcome as breastfed or not breastfed at, say, age 6 months.

In making such choices we need to balance two (sometimes opposing) considerations:

1 It is desirable to *choose the regression model that uses as much of the information in the data as possible*. In the blood pressure example, this would favour using linear regression with blood pressure as a continuous variable, since categorizing or dichotomizing it would discard some of the information collected (through using groups rather than the precise measurements). In the breastfeeding example, Cox or Poisson regression would be the preferred regression models, since the logistic regression analysis would discard important information on the precise time at which breastfeeding stopped.
2 It is often sensible to *use simpler models before proceeding to more complex ones*. For example, in examining the effect of exposures on an ordered categorical variable we might start by collapsing the variable into two categories and using logistic regression, before proceeding to use ordinal logistic regression to analyse the original outcome variable. We could then check whether the results of the two models are consistent, and assess whether the gain in precision of exposure effect estimates obtained using the original outcome variable justifies the extra complexity.

29.4 HYPOTHESIS TESTING IN REGRESSION MODELS

Hypothesis testing is used in regression models both to test the null hypothesis that there is no association between an exposure variable and the outcome, and in order to refine the model, for example by:

- Examining the assumption of no interaction (effect modification) between two or more exposure variables (see Section 29.5).
- Deciding between the different forms in which an exposure/confounder variable might be included, such as deciding between modelling the effect of a categorical exposure variable using indicator variables or including it as a *linear* (*dose–response*) effect (see Section 29.6).
- Deciding whether a variable needs to be included in the final regression model (see Section 29.8).

Hypothesis testing can be carried out using either **Wald tests** or **likelihood ratio tests**, as described in Section 28.7. The *P*-values corresponding to the different parameter estimates in computer outputs are based on Wald tests. These are

directly interpretable for exposure effects that are represented by a single parameter in the regression model. Examples have been given in Example 19.1 for the logistic regression of microfilarial infection with the binary exposure area (1 = rainforest/0 = savannah) and in Example 24.1 for the Poisson regression of myocardial infarction with the binary exposure 'cursmoke' (1 = men who were current smokers at the start of the study/0 = men who were never or ex-smokers at the start of the study). When an exposure effect is assumed to be linear (see Sections 19.3 and 29.6) it is also represented by a single parameter of the regression model.

Single parameter Wald tests are, however, less useful for a categorical variable, which is represented by a series of indicator variables in the regression model. Thus in Example 24.2, the Poisson regression output (Table 24.6) for the effect of social class on the rate of myocardial infarction has six parameter estimates, the rate in the baseline group and the five rate ratios comparing the other social class groups with the baseline. Wald z statistics and P-values are given for each of these five social class groups, enabling each of them to be compared with the baseline. What is needed, however, is a combined test of the null hypothesis that social class has no influence on the rate of myocardial infarction. Some computer packages have an option for a multi-parameter Wald test to do this.

We prefer instead to use **likelihood ratio tests** for all but the very simplest of cases, both for the reasons given in Chapter 28, and for the ease with which they can be calculated in all situations. As explained in Chapter 28, the **likelihood ratio statistic (LRS)** is calculated as minus twice the difference between the log likelihoods obtained from fitting the following two models:

1 L_{exc}, the log likelihood of the model *excluding* the variable(s) to be tested;
2 L_{inc}, the log likelihood of the model *including* the variable(s) to be tested.

This follows a χ^2 distribution with degrees of freedom equal to the number of parameters omitted from the model. For a simple binary exposure the degrees of freedom will equal one, and for a categorical exposure the degrees of freedom will equal the number of groups minus one.

$$\text{LRS} = -2 \times (L_{exc} - L_{inc})$$

is χ^2 with d.f. = number of additional parameters in the model including the exposure variable(s) to be tested

Note that the value of the log likelihood is a standard part of the computer output for a regression model.

Example 29.1

We will illustrate the use of the likelihood ratio test in the context of the Caerphilly cohort study, which was introduced in Chapter 24. We will base this on the

following three different Poisson regression models for rates of myocardial infarction fitted in that chapter:

1 Table 24.4: Cursmoke comparing smokers at recruitment with never/ex-smokers.
2 Table 24.6: Socclass comparing six social class groups.
3 Table 24.9: Model including both Cursmoke and Socclass.

The values of the log likelihoods for these three models, together with the model including no exposure variables (the 'constant-only model') are summarized in Table 29.2. We will refer to them as $L_{model\,1}$ to $L_{model\,4}$. The constant-only model, which has a single parameter corresponding to the constant term, is fitted by specifying the type of regression and the outcome, and nothing else.

Note that the parameter estimate corresponding to the 'constant' term is different for each of the four models. It represents the rate in the baseline group (those non-exposed to all of the exposure variables included in the model) against which all other comparisons are made. Its value therefore depends on which exposure variables are included in the model.

Hypothesis test for a single parameter

Cursmoke is a binary exposure variable. Model 2 therefore has two parameters:

1 Constant: the rate of myocardial infarction in the baseline group (never/ex-smokers), and
2 Cursmoke: the rate ratio comparing current smokers with never or ex-smokers.

The likelihood ratio statistic to test the null hypothesis that myocardial infarction rates are not related to smoking status at recruitment (Cursmoke) is based on a comparison of models 1 and 2. Note that as the value of L_{inc} ($L_{model\,2}$) is negative, minus becomes a plus in the calculation.

$$\begin{aligned} \text{LRS} &= -2 \times (L_{exc} - L_{inc}) = -2 \times (L_{model\,1} - L_{model\,2}) \\ &= -2 \times (-1206.985 + 1195.513) = 22.944 \end{aligned}$$

This is χ^2 with d.f. = number of additional parameters in the inclusive model $= 2 - 1 = 1$.

The corresponding *P*-value, derived from the χ^2 distribution on 1 degree of freedom, equals 0.0000017. There is thus strong evidence of an association between

Table 29.2 Log likelihood values obtained from different Poisson regression models fitted to data from the Caerphilly cohort study, as described in Chapter 24.

Model	Exposure(s) in model	No. of parameters	Log likelihood
1	None (Constant only model)	1	$L_{model\,1} = -1206.985$
2	Cursmoke (Yes/No)	2	$L_{model\,2} = -1195.513$
3	Socclass (6 groups)	6	$L_{model\,3} = -1201.002$
4	Cursmoke & Socclass	7	$L_{model\,4} = -1191.119$

current smoking and rate of myocardial infarction. The equivalent z statistic is $z = \sqrt{22.944} = 4.790$. This is similar to the corresponding Wald z statistic value of 4.680, given in the output in Table 24.5.

Hypothesis test for a categorical exposure with more than one parameter

When an exposure variable has more than two categories, its effect is modelled by introducing indicator variables corresponding to each of the non-baseline categories (as explained in Section 19.4). This is the case for Socclass, the men's social class at the start of the study. It has six categories, I = 1 (most affluent), II = 2, III non-manual = 3, III manual = 4, IV = 5, V = 6 (most deprived). Model 3 therefore has six parameters:

1 Constant: the rate of myocardial infarction in the baseline group, chosen to be III non-manual as more than half the men were in this group, and

2 Socclass: five rate ratios comparing each of the other social class groups with the baseline group.

To test the null hypothesis that social class has no effect on the rate of myocardial infarction, we compare the log likelihoods obtained in models 1 and 3. The likelihood ratio test statistic is

$$\begin{aligned} \text{LRS} &= -2 \times (L_{\text{exc}} - L_{\text{inc}}) = -2 \times (L_{\text{model 1}} - L_{\text{model 3}}) \\ &= -2 \times (-1206.985 + 1201.002) = 11.966 \\ \text{d.f.} &= \text{number of additional parameters in the inclusive model} = 6 - 1 = 5 \\ P &= 0.035 \end{aligned}$$

Because the effect of social class was modelled with five parameters, the P-value corresponding to this LRS is derived from the χ^2 distribution with 5 degrees of freedom. It equals 0.035. There is thus some evidence of an association between social class and rates of myocardial infarction. An alternative way to examine the effect of social class would be to carry out a test for linear trend, as was done in Example 24.2. Investigation of linear effects is discussed in detail in Section 29.6.

As mentioned above, it is also possible to derive a P-value from a multi-parameter version of the Wald test. This multi-parameter version is a χ^2 test with the same number of degrees of freedom as the likelihood ratio test. In this example the Wald statistic is $\chi^2 = 10.25$ with d.f. = 5. The corresponding P-value is 0.069, higher than that obtained from the likelihood ratio test.

Hypothesis tests in multivariable models

Models 2 and 3 in this example are univariable models, in which we examined the **crude** or **unadjusted effects** of a single exposure variable, namely the effects of smoking and of social class. We now consider the **multivariable model** including both smoking and social class. This is number 4 in Table 29.2. As previously

explained in Chapter 24, the effects in this model should be interpreted as the effect of smoking controlled for social class and the effect of social class controlled for smoking. To test the null hypothesis that there is no effect of social class after controlling for smoking, we compare:

1 the log likelihood of model 2, which includes only smoking, with

2 the log likelihood of model 4 which also includes social class, with the addition corresponding to the effect of social class controlled for smoking.

The likelihood ratio test statistic is:

$$\begin{aligned} \text{LRS} &= -2 \times (L_{exc} - L_{inc}) = -2 \times (L_{\text{model 2}} - L_{\text{model 4}}) \\ &= -2 \times (-1195.513 + 1191.119) = 8.788 \\ \text{d.f.} &= \text{number of additional parameters in the inclusive model} = 7 - 2 = 5 \\ P &= 0.118 \end{aligned}$$

There is therefore no good evidence for an association between social class and rates of myocardial infarction, other than that which acts through smoking. However, we should be aware that for an ordered categorical variable such as social class a more powerful approach may be to derive a *test for trend* by including social class as a linear effect in the model, rather than as a categorical variable. Modelling linear effects is discussed in detail in Section 29.6.

29.5 INVESTIGATING INTERACTION (EFFECT MODIFICATION) IN REGRESSION MODELS

Interaction was introduced in Section 18.5, where we explained that there is an *interaction* between the effects of two exposures if the effect of one exposure varies according to the level of the other exposure. For example, the protective effect of breastfeeding against infectious diseases in early infancy is more pronounced among infants living in poor environmental conditions than among those living in areas with adequate water supply and sanitation facilities. We also explained that an alternative term for interaction is **effect modification**. In this example, we can think of this as the quality of environmental conditions *modifying* the effect of breastfeeding. Finally, we noted that the most flexible approach to examine interaction is to use regression models, but that when we are using Mantel–Haenszel methods to control for confounding an alternative is to use a χ^2 test for effect modification, commonly called a **χ^2 test of heterogeneity**. Interaction, effect modification and heterogeneity are three different ways of describing exactly the same thing.

We have also seen that regression models including the effect of two or more exposures make the assumption that there is *no interaction* between the exposures. We now describe how to test this assumption by introducing **interaction terms** into the regression model.

Example 29.2

We will explain this in the context of the onchocerciasis dataset used throughout Chapters 19 and 20, where logistic regression was used to examine the effects of area of residence (forest or savannah) and of age group on the odds of microfilarial (*mf*) infection. We found strong associations of both area of residence and of age group with the odds of *mf* infection. We will do three things:

1 Remind ourselves of the results of the standard logistic regression model including both area and age group, which assumes that is there is *no interaction* between the two. In other words, it assumes that the effect of area is the same in each of the four age groups, and (correspondingly) that the effect of age is the same in the each of the two areas, and that any observed differences are due to sampling variation. Unless you are already familiar with how such models work, we strongly suggest that you read Section 20.2 where this is explained in detail, before continuing with this section.

2 We will then describe how to specify a regression model incorporating an interaction between the effects of area and age group, and how to interpret the regression output from such a model.

3 We will then calculate a likelihood ratio statistic using the log likelihoods of these two models to test the null hypothesis that there is no interaction between the effects of area and age group.

Model with two exposures and no interaction

Table 29.3 summarizes the results from the logistic regression model for *mf* infection including both area and age group, described in Section 20.2. Part (a) of the table shows the set of equations for the eight subgroups of the data that define the model in terms of its parameters. Note that the exposure effects represent odds ratios, and that they are *multiplicative*, since logistic regression models the *log odds*. The eight subgroups can be divided into four different types:

1 The *baseline* subgroup, consisting of those in the baseline groups of both area and age, namely those aged 5–9 years living in a savannah area. This is represented by the Baseline parameter in the model.

2 One subgroup consisting of those in the baseline group for age, but *not* for area, namely those aged 5–9 years living in a rainforest area. This subgroup is '*exposed to area but not to age*'. Its relative odds of *mf* infection compared to the baseline is modelled by the Area parameter.

3 Three subgroups corresponding to those in each of the three non-baseline age groups, but who are in the baseline group for area, namely those living in savannah areas aged 10–19 years, 20–39 years, or 40 years or more. These subgroups are '*exposed to age but not area*'. Their relative odds of *mf* infection compared to the baseline are modelled by the three age group parameters, Agegrp(1), Agegrp(2) and Agegrp(3), respectively.

4 Three subgroups corresponding to those in each of the three non-baseline age groups who are also in the non-baseline group for area, namely those living in rainforest areas aged 10–19 years, 20–39 years, or 40 years or more. These subgroups are '*exposed to both area and age*'. If we assume that there is *no interaction* between the two exposures, the relative odds of *mf* infection in these three subgroups compared to the baseline are modelled by multiplying together the Area parameter and the relevant age group parameter. This gives Area $\times$ Agegrp(1), Area $\times$ Agegrp(2) and Area $\times$ Agegrp(3), respectively.

The model for the odds of *mf* infection in the eight subgroups therefore contains just five parameters. This is made possible by the assumption of *no interaction*. The parameter estimates are shown in part (b) of Table 29.3. Part (c) shows the values obtained when these estimates are inserted into the equations in part (a) to give estimated values of the odds of *mf* infection according to area and age group. The observed odds of *mf* infection in each group are also shown.

Model incorporating an interaction between the two exposures

We now describe how to specify an alternative regression model incorporating an interaction between the effects of the two exposures. We no longer assume that the

Table 29.3 Results from the logistic regression model for *mf* infection, including both area of residence and age group, assuming *no interaction* between the effects of area and age group.

(a) Odds of *mf* infection by area and age group, expressed in terms of the parameters of the logistic regression model: Odds = Baseline $\times$ Area $\times$ Age group.

	Odds of *mf* infection	
Age group	Savannah areas (Unexposed)	Rainforest areas (Exposed)
0 (5–9 years)	Baseline	Baseline $\times$ Area
1 (10–19 years)	Baseline $\times$ Agegrp(1)	Baseline $\times$ Area $\times$ Agegrp(1)
2 (20–39 years)	Baseline $\times$ Agegrp(2)	Baseline $\times$ Area $\times$ Agegrp(2)
3 ($\geq$ 40 years)	Baseline $\times$ Agegrp(3)	Baseline $\times$ Area $\times$ Agegrp(3)

(b) Parameter estimates obtained by fitting the model.

	Baseline	Area	Agegrp(1)	Agegrp(2)	Agegrp(3)
Odds ratio	0.147	3.083	2.599	9.765	17.64

(c) Odds of *mf* infection by area and age group, as estimated from the logistic regression model, and as observed.

	Savannah areas: odds of *mf* infection		Rainforest areas: odds of *mf* infection	
Age group	Estimated	Observed	Estimated	Observed
0 (5–9 years)	0.147	0.208	$0.147 \times 3.083 = 0.453$	0.380
1 (10–19 years)	$0.147 \times 2.599 = 0.382$	0.440	$0.147 \times 3.083 \times 2.599 = 1.178$	1.116
2 (20–39 years)	$0.147 \times 9.765 = 1.435$	1.447	$0.147 \times 3.083 \times 9.765 = 4.426$	4.400
3 ($\geq$ 40 years)	$0.147 \times 17.64 = 2.593$	2.182	$0.147 \times 3.083 \times 17.64 = 7.993$	10.32

relative odds of *mf* infection in the subgroups '*exposed to both age and area*' can be modelled by multiplying the area and age effects together. Instead we introduce extra parameters, called **interaction parameters**, as shown in Table 29.4(a). These allow the effect of area to be different in the four age groups and, correspondingly, the effects of age to be different in the two areas. An interaction parameter is denoted by the exposure parameters for the subgroup written with a full stop between them. The three interaction parameters in this example are denoted Area.Agegrp(1), Area.Agegrp(2) and Area.Agegrp(3).

This new model is fitted using seven indicator variables as shown in Box 29.1. The parameter estimates for this model are shown in Table 29.4(b). Table 29.4(c) shows the values obtained when these are inserted into the equations in part (a). Note that:

1 Since this model has *eight* parameters, the same as the number of area × age subgroups, there is an exact agreement between the estimated odds of *mf* infection in each subgroup and the observed odds, as shown in Tables 29.3(c) and 20.3.

2 Including interaction terms leads to different estimates of the baseline, area and age group parameters than those obtained in the model assuming no interaction. It is important to realize that the interpretation of the area and age group parameters is also different.
- The Area parameter estimate (1.8275) is the odds ratio for area *in the baseline age group*. In the model assuming no interaction, the Area parameter estimate (3.083) is a weighted average of the odds ratios for area in the four age groups and is interpreted as the odds ratio for area after controlling for age group.
- Similarly, the age group parameter estimates represent the effect of age in the *baseline area group*, in other words the effect among those living in savannah areas.

3 The estimates for the interaction parameters are all greater than one. This corresponds to a synergistic effect between area and each of the age groups, with the combined effect more than simply the combination of the separate effects. A value of one for an interaction term is equivalent to no interaction effect. A value less than one would mean that the combined effect of both exposures is less than the combination of their separate effects.

4 The interaction parameters allow the area effect to be different in the four age groups. They can be used to calculate age-specific area odds ratios as follows:
- The Area parameter estimate equals 1.8275, and is the area odds ratio (comparing those living in rainforest areas with those living in savannah areas) in the *baseline* age group (5–9 years).
- Multiplying the Area parameter estimate by the interaction parameter estimate Area.Agegrp(1) gives the odds ratio for area in age group 1 (10–19 years):

$$\begin{aligned}\text{OR for area in age group 1} &= \text{Area} \times \text{Area.Agegrp(1)} \\ &= 1.8275 \times 1.3878 = 2.5362\end{aligned}$$

Table 29.4 Logistic regression model for *mf* infection, including both area of residence and age group, and *incorporating an interaction* between their effects.

(a) Odds of *mf* infection by area and age group, expressed in terms of the parameters of the logistic regression model, with the interaction parameters shown in bold: Odds = Baseline × Area × Agegroup × Area.Agegroup

	Odds of *mf* infection	
Age group	Savannah areas (Unexposed)	Rainforest areas (Exposed)
0 (5–9 years)	Baseline	Baseline × Area
1 (10–19 years)	Baseline × Agegrp(1)	Baseline × Area × Agegrp(1) × **Area.Agegrp(1)**
2 (20–39 years)	Baseline × Agegrp(2)	Baseline × Area × Agegrp(2) × **Area.Agegrp(2)**
3 (≥40 years)	Baseline × Agegrp(3)	Baseline × Area × Agegrp(3) × **Area.Agegrp(3)**

(b) Computer output showing the results from fitting the model (interaction parameters shown in bold).

	Odds ratio	*z*	$P > \|z\|$	95 % CI
Area.Agegrp(1)	**1.3878**	**0.708**	**0.479**	**0.560 to 3.439**
Area.Agegrp(2)	**1.6638**	**1.227**	**0.220**	**0.738 to 3.751**
Area.Agegrp(3)	**2.5881**	**2.171**	**0.030**	**1.097 to 6.107**
Area	1.8275	1.730	0.084	0.923 to 3.619
Agegrp(1)	2.1175	1.998	0.046	1.015 to 4.420
Agegrp(2)	6.9639	6.284	0.000	3.802 to 12.76
Agegrp(3)	10.500	7.362	0.000	5.614 to 19.64
Constant (Baseline)	0.2078	−5.72	0.000	0.121 to 0.356

(c) Odds of *mf* infection by area and age group, as estimated from the logistic regression model, with the interaction parameters shown in bold.

	Odds of *mf* infection	
Age group	Savannah areas	Rainforest areas
0 (5–9 years)	0.2078	0.2078 × 1.8275 = 0.380
1 (10–19 years)	0.2078 × 2.1175 = 0.440	0.2078 × 1.8275 × 2.1175 × **1.3878** = 1.116
2 (20–39 years)	0.2078 × 6.9639 = 1.447	0.2078 × 1.8275 × 6.9639 × **1.6638** = 4.400
3 (≥40 years)	0.2078 × 10.500 = 2.182	0.2078 × 1.8275 × 10.500 × **2.5881** = 10.32

Similarly,

$$\text{OR for area in age group 2} = \text{Area} \times \text{Area.Agegrp(2)} = 1.8275 \times 1.6638 = 3.0406$$

and

$$\text{OR for area in age group 3} = \text{Area} \times \text{Area.Agegrp(3)} = 1.8275 \times 2.5881 = 4.7300$$

These four age-group-specific area odds ratios are the same as those shown in Tables 20.3 and 20.4.

5 In exactly the same way, the interaction parameters can be used to calculate area-specific age group odds ratios. For example:

$$\begin{aligned}\text{OR for age group 1 in rainforest areas} &= \text{Agegrp(1)} \times \text{Area.Agegrp(1)} \\ &= 2.1175 \times 1.3878 = 2.9386\end{aligned}$$

6 An alternative expression of these same relationships is to note that the interaction parameter Area.Agegrp(1) is equal to the *ratio* of the odds ratios for area in age group 1 and age group 0, presented in Tables 20.3 and 20.4. For example:

$$\text{Area.Agegrp(1)} = \frac{\text{OR for area in age group 1}}{\text{OR for area in age group 0}} = \frac{2.5362}{1.8275} = 1.3878$$

If there is no interaction then the area odds ratios are the same in each age group and the interaction parameter equals 1.

7 Alternatively, we can express the interaction parameter Area.Agegrp(1) as the ratio of the odds ratios for age group 1 (compared to age group 0), in area 1 and area 0:

$$\text{Area.Agegrp(1)} = \frac{\text{OR for age group 1 in area 1}}{\text{OR for age group 1 in area 0}} = \frac{2.9386}{2.1175} = 1.3878$$

(The odds ratios for age group 1 were calculated using the raw data presented in Table 20.3).

8 The other interaction parameter estimates all have similar interpretations: for example the estimate for Area.Agegrp(2) equals the ratio of the area odds ratios in age group 2 and age group 0, and equivalently it equals the ratio of the odds ratios for age group 2 (compared to age group 0) in area 1 and area 0.

9 For a model allowing for interaction between two binary exposure variables, the *P*-value corresponding to the interaction parameter estimate corresponds to a Wald test of the null hypothesis that there is no interaction. When, as in this example, there is more than one interaction parameter, the individual *P*-values corresponding to the interaction parameters are not useful in assessing the evidence for interaction: we describe how to derive the appropriate likelihood ratio test later in this section.

Table 29.5 summarizes the interpretation of the interaction parameters for different types of regression models.

Table 29.5 Interpretation of interaction parameters.

Type of regression model	Interpretation of interaction parameters
Linear	Difference between mean differences
Logistic	Ratio of odds ratios
Poisson	Ratio of rate ratios

BOX 29.1 USING INDICATOR VARIABLES TO INVESTIGATE INTERACTION IN REGRESSION MODELS

Values of the seven indicator variables used in a model to examine the interaction between area (binary variable) and age group (4 groups):

Age group	Area	Area	Age(1)	Age(2)	Age(3)	Area.Age(1)	Area.Age(2)	Area.Age(3)
5–9 years (0)	Savannah	0	0	0	0	0	0	0
	Forest	1	0	0	0	0	0	0
10–19 (1)	Savannah	0	1	0	0	0	0	0
	Forest	1	1	0	0	1	0	0
20–39 years (2)	Savannah	0	0	1	0	0	0	0
	Forest	1	0	1	0	0	1	0
≥40 years (3)	Savannah	0	0	0	1	0	0	0
	Forest	1	0	0	1	0	0	1

Likelihood ratio test for interaction

To test the null hypothesis that there is no interaction between area and age group, we need to compare the log likelihoods obtained in the two models excluding and including the interaction parameters. These are shown in Table 29.6. The likelihood ratio test statistic is:

$$\text{LRS} = -2 \times (\text{L}_{\text{exc}} - \text{L}_{\text{inc}}) = -2 \times (-692.407 + 689.773) = 5.268$$

$$\text{d.f.} = \text{number of additional parameters in the inclusive model} = 8 - 5 = 3$$

$$P = 0.153$$

Therefore this analysis provides little evidence of interaction between the effects of area and age on the odds of microfilarial infection

Table 29.6 Log likelihood values obtained from the logistic regression models for *mf* infection by area of residence and age group, (a) assuming *no* interaction, and (b) incorporating an interaction between the effects of area and age group.

Model	Exposure(s) in model	No. of parameters	Log likelihood
(a) exc	Area and Agegrp	5	−692.407
(b) inc	Area, Agegrp and Area.Agegrp	8	−689.773

Interactions with continuous variables

It is straightforward to incorporate an interaction between the effects of a continuous exposure variable (x) and a binary exposure variable (b, coded as 0 for

unexposed and 1 for exposed individuals) in a regression model, by multiplying the values of the two exposures together to create a new variable ($x.b$) representing the interaction, as shown in Table 29.7. This new variable equals 0 for those unexposed to exposure b, and the value of exposure x for those exposed to b. The regression coefficient for $x.b$ then corresponds to the difference between the slope in individuals exposed to b and the slope in individuals not exposed to b, and the evidence for an interaction may be assessed either using the Wald P-value for $x.b$, or by omitting $x.b$ from the model and performing a likelihood ratio test.

To examine interactions between two continuous exposure variables w and x, it is usual to create a new variable $w.x$ by multiplying w by x. If the regression coefficient for $w.x$ is 0 (1 for models with exposure effects reported as ratios) then there is no evidence of interaction.

Table 29.7 Creating a variable to represent an interaction between a continuous and a binary exposure variable.

Continuous exposure (x)	Binary exposure (b)	Interaction variable ($x.b$)
x	0 (unexposed)	0
x	1 (exposed)	x

Confounding and interaction

Note that confounding and interaction may coexist. If there is clear evidence of an interaction between the exposure and the confounder, it is no longer adequate to report the effect of the exposure controlled for the confounder, since this assumes the effect of the exposure to be the same at each level of the confounder. This is not the case when interaction is present. Instead, we should report *separate* exposure effects for each *stratum* of the confounder. We can derive these by performing a separate regression to examine the association between the exposure and outcome variables, for each level of the confounding variable.

It is possible to derive stratum-specific effects in regression models by including appropriate indicator variables, or combining regression coefficients as was done in Table 29.4(c). This has the advantage of allowing estimation of such effects, controlled for the effects of other exposure variables. Confidence intervals for such combinations of regression coefficients need to take into account the covariance (a measure of the association) between the individual regression coefficients: some statistical packages provide commands to combine regression coefficients and derive corresponding confidence intervals.

An advantage of Mantel–Haenszel methods is that because the stratum-specific exposure effects tend to be presented in computer output, we are encouraged to look for evidence of interaction. In regression models we have to fit interaction terms explicitly to do this.

Regression models with more than two variables

The power of regression models is that, providing we make the simplifying assumption of no interactions, they allow us to examine the joint (simultaneous) effects of a number of exposure variables. For example, suppose we had four exposure variables, with 2, 3, 4 and 5 levels respectively. The number of subgroups defined by different combinations of these exposure groups would be $2 \times 3 \times 4 \times 5 = 120$. Mantel–Haenszel methods to adjust for confounding would need to stratify by all these 120 subgroups. Similarly, a regression model that included all the interactions between these four exposures would also have 120 parameters. However, a regression model that assumes no interaction between any of the exposures would contain only 11 parameters, one for the baseline (constant) term plus 1, 2, 3 and 4 parameters for each of the four exposure variables, since $(k - 1)$ parameters are needed for an exposure with k levels. Interactions between confounding variables are often omitted from regression models: this is discussed in more detail in Chapter 38.

Increasing power in tests for interaction

The interpretation of tests for interaction is difficult. As discussed in more detail in Sections 35.4 and 38.6, tests for interaction usually have *low power*, so that the absence of strong evidence that interaction is present does not imply that interaction is absent.

A further problem, in addition to that of low power, occurs in regression models with binary or time-to-event outcomes, when some subgroups contain no individuals who experienced the outcome event. If this is the case, then interaction parameters for that subgroup cannot be estimated, and statistical computer packages may then drop all individuals in such subgroups from the analysis. This means that the model including the interactions is not directly comparable with the one assuming no interaction.

A solution to both of these problems is to *combine exposure groups*, so that the interaction introduces only a small number of extra parameters. For example, to investigate possible interactions between area and age we might first combine age groups to create a binary age group variable, separating those aged 0 to 19 years from those aged 20 years or more. Note that it is perfectly permissible to examine interactions using indicator variables based on binary variables, while controlling for the exposure effects based on the original (ungrouped) variables.

Further advice on examining interactions is provided in Box 18.1 on page 188 and in Chapter 38.

29.6 INVESTIGATING LINEAR EFFECTS (DOSE–RESPONSE RELATIONSHIPS) IN REGRESSION MODELS

Exposure effects may be modelled as linear if the exposure is either a numerical or an ordered categorical variable. In modelling exposure effects as linear, we assume

that the outcome increases or decreases systematically with the exposure effect, as depicted in Figure 29.1, panels (a) and (b). If the observed association is as depicted in panel (c) of Figure 29.1, then it is appropriate to conclude that there is no linear effect. However, it is essential to be aware of the possibility that there is **extra-linear variation** in the exposure–outcome relationship. An example is depicted in panel (d) of Figure 29.1. Here, a regression model assuming a linear effect would conclude that there was no association between the exposure and the outcome. This would be incorrect, because there is in fact a non-linear association: the outcome level first increases and then decreases with increasing exposure.

The interpretation of linear effects, and the methods available to examine them, depends on the type of outcome and regression model:

- in linear or multiple regression models, the linear effect corresponds to a constant increase in the *mean* of the outcome per unit increase in the exposure variable;
- in logistic regression or conditional logistic regression models, it corresponds to a constant increase in the *log odds* per unit increase in the exposure variable;
- in Poisson regression models, it corresponds to a constant increase in the *log rate* per unit increase in the exposure variable; and
- in Cox regression models, it corresponds to a constant increase in the *log hazard* per unit increase in the exposure variable.

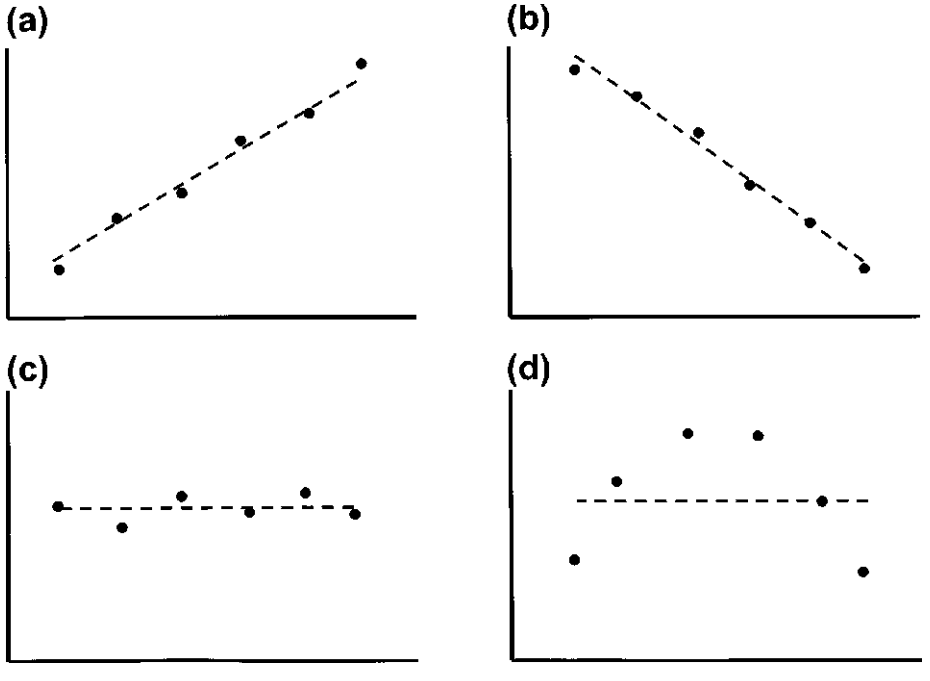

Fig. 29.1 Four possibilities for the association between outcome and exposure. Panels (a) and (b) show, respectively, positive and negative linear associations between the outcome and exposure. In panel (c) there is no association between the outcome and exposure; the estimated linear effect is zero. In panel (d) the linear effect is also zero, but there is a non-linear association between the outcome and the exposure.

When exposure effects are expressed as ratio measures, the linear effect corresponds to the amount by which the outcome is *multiplied* per unit increase in the exposure variable. For example, Table 19.11 shows that for the onchocerciasis data the log odds ratio for the linear association between microfilarial infection and age group was 0.930, corresponding to an odds ratio of 2.534 per unit increase in age group. The odds ratio comparing age group 2 with age group 0 is therefore $2.534^2 = 6.421$, and in general the odds ratio for an increase of k age groups is 2.534^k.

We saw in Chapter 10 on *linear regression* that the first step in examining the association between a numerical outcome and a numerical exposure is to draw a scatter plot. Such plots should protect us from making errors such as that depicted in panel (d) of Figure 29.1, where an assumption of a linear effect would lead to the incorrect conclusion that there is no association between the exposure and outcome.

For *logistic* and *Poisson regression*, such plots cannot be drawn without first grouping the exposure variable and then graphing the outcome (e.g. log odds, log rate) in each group. For example, the odds of a binary outcome for an individual are either $0/1 = 0$, or $1/0 =$ infinity. We cannot therefore graph the log odds for individuals, but we can calculate the log odds in groups (e.g. age groups) provided that there is at least one individual with and one without the disease outcome in each group. Therefore it is sensible to group numerical exposure variables into ordered categories in early analyses, in order to check for linearity in the measure of effect. If the exposure–outcome association appears approximately linear then the original continuous variable may be used in subsequent models. For example, Figure 19.2 shows that there is an approximately linear association between the log odds of microfilarial infection and age group in the onchocerciasis data.

In *conditional logistic regression* and *Cox regression*, in which exposure effects are calculated by comparing exposures within case–control strata or risk sets, it is not possible to draw such graphs of outcome against exposure, and it is essential to examine linearity assumptions within regression models.

Testing for a linear effect

We test the null hypothesis that there is no linear effect in the usual way using a likelihood ratio test, by comparing L_{inc}, the log likelihood from the model including the linear effect (and other exposure effects of interest), with L_{exc}, the log likelihood from the model excluding the linear effect. Standard regression output for the linear exposure effect reports the *P*-value corresponding to the Wald test of this null hypothesis.

Testing for departure from linearity

We test the *null hypothesis that the exposure effect is linear* by comparing the model assuming a linear effect with a *more general* model in which the exposure effect is not assumed to be linear. We will describe two ways of doing this:

1 for ordered categorical exposure variables, this comparison may be with a model including the exposure as a categorical variable, where indicator variables are used to estimate the difference in outcome, comparing each non-baseline category with the baseline;
2 for any ordered categorical or numerical exposure variable, we may examine the linearity assumption by introducing **quadratic** terms into the model.

Testing linearity for ordered categorical variables

The null hypothesis is that the exposure effect *is* linear. To test this, we derive a likelihood ratio statistic by comparing:

(a) L_{exc}, the log likelihood when the exposure effect is assumed to be linear (the null hypothesis);
(b) L_{inc}, the log likelihood of the model when we allow the exposure effect to be non-linear, and which therefore includes additional parameters.

Example 29.2 (continued)

We will illustrate this approach by examining the linear effect of age group in the onchocerciasis data. The two models are:

(a) A logistic regression model of the odds of *mf* infection with age group as a linear effect. This includes just two parameters, the baseline (constant) plus a linear effect for age group. Their estimates were given in Table 19.13.
(b) A logistic regression model of the odds of *mf* infection with age group as a categorical variable. This model makes no assumption about the shape of the relationship between age group and *mf* infection. It includes four parameters, the baseline and three indicator variables for comparing each of the other three age groups with the baseline group. The parameter estimates were given in Table 19.11.

Note that model (a) is a special case of the more general model (b). The log likelihood values obtained in these two models are shown in Table 29.8. The likelihood ratio test statistic is:

$$\begin{aligned} \text{LRS} &= -2 \times (L_{exc} - L_{inc}) = -2 \times (-729.240 + 727.831) = 2.818 \\ \text{d.f.} &= \text{number of additional parameters in the inclusive model} = 4 - 2 = 2 \\ P &= 0.24 \end{aligned}$$

Table 29.8 Log likelihood values obtained from the logistic regression models for *mf* infection by area of residence and age group, (a) assuming a linear effect of age group, and (b) allowing for a non-linear effect of age group, by including indicator variables.

Model	Exposure(s) in model	No. of parameters	Log likelihood
(a) exc	Age group (linear, see Table 19.13)	2	−729.240
(b) inc	Agegrp (categorical, see Table 19.11)	4	−727.831

There is no evidence against the null hypothesis that the effect of age group is linear. The likelihood ratio statistic has two degrees of freedom, corresponding to the extra number of parameters needed to include age group as a categorical variable compared to including it as a linear effect.

If the likelihood ratio test does provide evidence of non-linearity, then the exposure effect should be modelled using separate indicator variables for each non-baseline exposure level, as in model (b).

Testing linearity using quadratic exposure effects

We will illustrate the second approach to testing linearity in the context of the Caerphilly study by examining the effect of fibrinogen (a numerical exposure) on the rate of myocardial infarction (MI).

Example 29.3

Fibrinogen, a factor involved in blood coagulation that has been shown to be associated with rates of cardiovascular disease in a number of studies, was measured at the baseline examination in the Caerphilly study. Its distribution is shown by the histogram in Figure 29.2(a). An initial examination of the association between fibrinogen and rates of myocardial infarction was done by:

- dividing the distribution into deciles (the lowest 10% of fibrinogen measurements, the second 10% and so on; see Section 3.3);
- calculating the median fibrinogen level in each of these deciles. These were 2.63, 3, 3.22, 3.4, 3.6, 3.8, 4, 4.25, 4.59 and 5.23;
- graphing the rate of myocardial infarction (per 1000 person-years, log scale) in each decile against median fibrinogen in each decile.

The results are shown in Figure 29.2(b). There appears to be an approximately linear association between fibrinogen and the log rate of MI.

A Poisson regression model was then fitted for the linear effect of fibrinogen (using the original, ungrouped measurement) on rates of MI. The results are

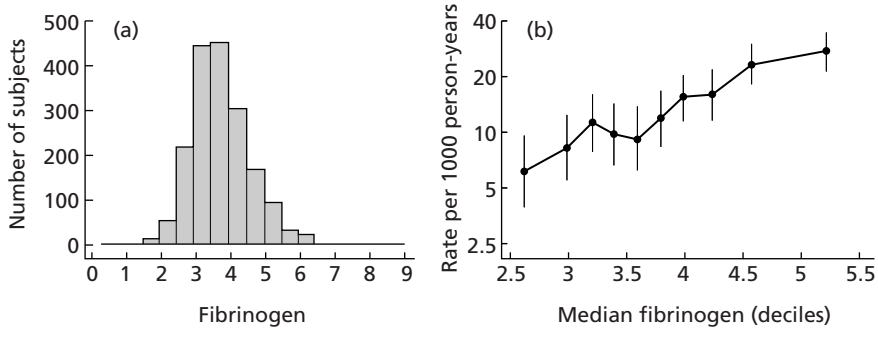

Fig. 29.2 (a) Histogram showing the distribution of fibrinogen (100g/dL) at the baseline examination of the Caerphilly study, and (b) MI rates (per 1000 person-years, log scale, with 95% confidence intervals for the rate in each group) for the median fibrinogen in each decile.

Table 29.9 Output from a Poisson regression model for the linear effect of fibrinogen on log rates of myocardial infarction in the Caerphilly study.

(a) Output on log scale

	Coefficient	s.e.	z	$P > \|z\|$	95% CI
Fibrinogen	0.467	0.054	8.645	0.000	0.361 to 0.573
Constant	−6.140	0.228	−26.973	0.000	−6.587 to −5.694

(b) Output on rate ratio scale

	Rate ratio	z	$P > \|z\|$	95% CI
Fibrinogen	1.595	8.645	0.000	1.435 to 1.773

shown in Table 29.9. The regression coefficient for fibrinogen is 0.467, corresponding to a rate ratio per unit increase of 1.595. This implies that the rate ratio for a three-unit increase in fibrinogen (from 2.5 to 5.5) is $1.595^3 = 4.057$. This is consistent with the increase seen over this range in Figure 29.2(b).

Although there is clear evidence of a linear (dose–response) association between fibrinogen and log rates of myocardial infarction, we may still wish to derive a formal test for extra-linear variation. Mathematically, the simplest departure from a linear relationship between the outcome and an exposure (x) is a **quadratic relationship**. The algebraic form of such a relationship is:

$$\text{outcome} = \beta_0 + \beta_1 x + \beta_2 x^2$$

To examine the evidence for a quadratic exposure effect, we create a new variable whose values are the squares of the exposure being examined. We then fit a regression model including *both* the exposure and the new variable (exposure squared).

Table 29.10 shows the Poisson regression output for the model including the linear effect of fibrinogen, and fibrinogen^2. There is only weak evidence (Wald P-value = 0.091) for a quadratic effect, so it would be reasonable to conclude that the effect of fibrinogen on log rates of MI is approximately linear. The fact that the regression coefficient for fibrinogen^2 is less than 0 (rate ratio < 1) implies that the effect of fibrinogen decreases as fibrinogen increases.

Because the linear and quadratic effects are sometimes *collinear* (see Section 29.7), it is preferable to examine the evidence for non-linearity using a likelihood ratio test comparing the models including and excluding the quadratic effect. When quadratic exposure effects are included in a model, we should not attempt to interpret the linear effect alone. In particular, the Wald P-value of 0.002 for the linear effect in Table 29.10 should *not* be interpreted as testing the null hypothesis that there is no linear effect of fibrinogen.

Table 29.10 Output from a Poisson regression model for the quadratic effect of fibrinogen on log rates of myocardial infarction in the Caerphilly study.

(a) Output on log scale

	Coefficient	s.e.	z	$P > \|z\|$	95% CI
Fibrinogen	1.038	0.338	3.073	0.002	0.376 to 1.700
Fibrinogen^2	−0.062	0.037	−1.688	0.091	−0.134 to 0.010
Constant	−7.383	0.757	−9.750	0.000	−8.868 to −5.899

(b) Output on rate ratio scale

	Rate ratio	z	$P > \|z\|$	95% CI
Fibrinogen	2.824	3.073	0.002	1.457 to 5.475
Fibrinogen^2	0.940	−1.688	0.091	0.874 to 1.010

Dose–response and unexposed groups

When examining dose–response relationships we should distinguish between the exposed group with the minimum exposure, and the unexposed group. For example, it may be that smokers in general have a higher risk of some disease than non-smokers. In addition, there may be an increasing risk of disease with increasing tobacco consumption. However, including the non-smokers with the smokers may bias our estimate of the dose–response relationship (linear effect) among smokers. This is illustrated in Figure 29.3. There are two possible ways to restrict estimation of the linear effect to exposed individuals:

1 Exclude the unexposed group, then estimate the linear effect among the exposed.
2 Include an indicator variable for exposed/unexposed together with linear effect of the exposure variable. The regression coefficient for the exposure will then estimate the linear effect among the exposed, while the regression coefficient for the indicator variable will estimate the difference between the outcome in the unexposed group and that projected by the linear effect in the exposed (dotted line in Figure 29.3).

Remarks on linear effects

1 It makes sense to model an exposure effect as linear if it is plausible that the outcome will increase (or decrease) systematically with the level of exposure. Such an exposure effect is known as a **dose–response** relationship, or **trend**.
2 A test for trend (see Section 17.5) is an approximation (based on a score test) to a likelihood ratio test of the null hypothesis that the regression coefficient for a linear effect is zero.
3 The existence of a dose–response relationship may provide more convincing evidence of a causal effect of exposure than a simple comparison of exposed with unexposed subjects.

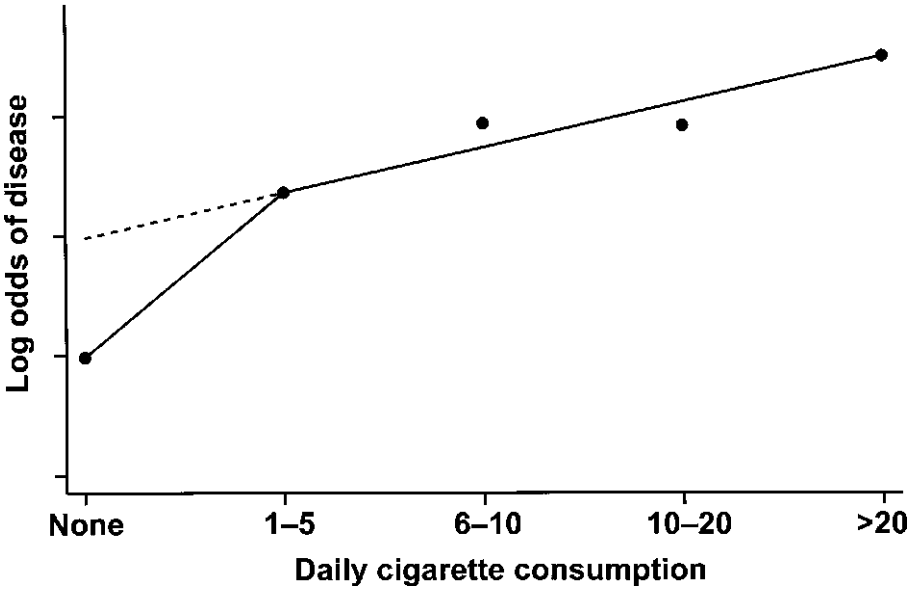

Fig. 29.3 Possible association between cigarette consumption and the log odds of a disease outcome. There is a larger difference between exposed (smokers) and unexposed (non-smokers) than would be expected given the magnitude of the dose–response relationship among the smokers.

4 Estimating a linear effect will often be the *most powerful* way to detect an association with an ordered exposure variable. This is because we only estimate one parameter, rather than a parameter for each non-baseline level. However, it is essential that this simplifying assumption, that an exposure effect may be modelled as a linear effect, be checked.
5 Modelling an exposure effect as linear will only be valid if the exposure is ordered categorical or numerical. Ideally, the category values should reflect the degree of exposure. For example, if the exposure was level of blood pressure and the four categories of exposure were obtained by grouping the blood pressures, then the category values could be the midpoints or mean of blood pressure in each of the four groups. In the absence of any genuine measurement (for instance when we model the effects of social class) it is usual to assign scores 0,1,2,3,4 . . . to the various exposure levels.

29.7 COLLINEARITY

When two exposure variables are highly correlated we say that they are **collinear**. Collinearity can cause problems in fitting and interpreting regression models, because inclusion of two highly correlated exposure variables in a regression model can give the impression that neither is associated with the outcome, even when each exposure is strongly associated (individually) with the outcome.

We will illustrate this by examining the regression of *height* (the outcome variable) and *age* (the exposure variable) from the study of 636 children living in Lima, Peru (see Chapter 11), in the presence of an artificially constructed variable *newage*. *Newage* has been computer-generated to be *collinear* with age, by adding a random 'error' to each age, with the standard deviation of this random error made equal to 1 year. This has led to a correlation of 0.57 between *age* and *newage*, which is high but not very high.

The correlation between *height* and *age* was 0.59, and the regression coefficient was 5.15 cm/year (s.e. = 0.28). A regression of *height* on *newage* alone gives a regression coefficient of 1.61. This is much smaller than the regression coefficient for *age* (5.15) because the addition of a random error component tends to reduce the regression coefficient (see Chapter 36 for a more detailed discussion of this issue).

When both *age* and *newage* are included in the model, the regression coefficient for *age* is slightly increased (5.31) compared to the value for the model with *age* alone (5.15), while the regression coefficient for *newage* is slightly less than zero. These results are shown in the first row of Table 29.11. Thus, the joint regression has correctly identified strong evidence of an association between *height* and *age*, taking *newage* into account, and no evidence of an association between *height* and *newage*, taking *age* into account. In this artificially created example, the regression has correctly identified the joint information of *age* and *newage* being contained in *age*, since in essence *newage* is a less accurate measure of age. This level of collinearity in this particular example has not caused a problem.

We will now demonstrate how problems can occur with increasing collinearity between *age* and *newage* by decreasing the standard deviation of the random error that is added to variable *age* to create variable *newage*. The second row of Table 29.11 shows that when this standard deviation is decreased to 0.1 the correlation between *age* and *newage* is very high (0.9904). The coefficient from the regression of *height* on *newage* alone is 5.06: close to the regression coefficient for *age* alone. When both *age* and *newage* are included in the model, there is a substantial increase in the regression coefficient for *age*, while the regression coefficient for

Table 29.11 Demonstration of the effect of collinearity, using data from the study of lung disease in children in Lima, Peru. Variable *newage* is variable *age* plus a random error whose standard deviation is given in the first column in the table.

s.d. of random error	Correlation between *age* and *newage*	Regression of height on *newage*	Regression of height on *age* and *newage*		
		Coefficient (s.e.) for *newage*	Coefficient (s.e.) for *age*	Coefficient (s.e.) for *newage*	Sum of coefficients
1	0.57	1.61 (0.20)	5.31 (0.33)	− 0.17 (0.20)	5.16
0.1	0.9904	5.06 (0.28)	6.81 (2.00)	− 1.66 (1.99)	5.15
0.01	0.9999	5.16 (0.28)	21.76 (19.94)	−16.62 (19.94)	5.14

newage is clearly negative. The important thing to notice is that there is an even more dramatic increase in the standard errors of both regression coefficients.

When the standard deviation of the random error is reduced to 0.01, the correlation between *age* and *newage* is extremely high (0.9999, third row of Table 29.11). The regression coefficient for *newage* alone is almost identical to that for *age*, as would be expected because the error now contained in *newage* as a measure of age is very small. Inclusion of both variables in the model has a dramatic effect: the regression coefficient for *age* is greatly increased to 21.76, while the regression coefficient for *newage* is reduced to −16.62. The standard error of each regression coefficient is large (19.94). This joint model could lead to the erroneous conclusion that neither *age* nor *newage* is associated with the outcome variable, *height*.

The final column of the table shows that although the regression coefficients for *age* and *newage* change dramatically as the collinearity between them increases, the sum of the two coefficients remains approximately constant, and is the same as the regression coefficient for *age* alone. This suggests a solution to the problem. It is not possible simultaneously to estimate the effects of both *age* and *newage*, because each has the same association with height. However we can estimate the association of the outcome with the sum (or, equivalently, the average) of the two variables. Alternatively, we can simply choose one of the variables for inclusion in our model and exclude the other one.

In conclusion, this example demonstrates that including two strongly collinear exposure variables in a regression model has the potential to lead to the erroneous conclusion that neither is associated with the outcome variable. This occurs when collinearity is high enough to lead to dramatic increases in the standard errors of the regression coefficients. Comparing the standard errors from the single exposure models with the joint exposure model can identify whether this problem is occurring. When it does occur, it is not possible to estimate the effect of each exposure controlling for the other in a regression model.

29.8 DECIDING WHICH EXPOSURE VARIABLES TO INCLUDE IN A REGRESSION MODEL

A key challenge in analysing studies that have data on a large number of exposure variables is how to decide which of these variables to include and which to exclude from a particular regression model, since it is usually unwise or impossible to include all of them in the same model. A rough guide is that there should be at least ten times as many observations (individuals) as exposure variables in a regression model: for example, a model which includes ten variables should be based on data from at least 100 individuals. Note that each separate indicator variable counts as a separate variable.

Two important considerations will influence how the choice of exposure variables is made:

1 Are you using multiple linear regression, or a different generalized linear model?

2 Is the main aim of the model to estimate the effect of a particular exposure as accurately as possible, to predict the outcome based on the values of a number of exposures, or to develop an explanatory model of those exposures that have an influence on the outcome?

Implication of type of regression model

For **multiple linear regression**, you should aim to include *all* exposure variables that are clearly associated with the outcome when estimating the effect of a particular exposure, *whether or not they are confounders* (with the exception that variables on the causal pathway between the exposure of interest and the outcome should *not* be included; see Section 18.2). Doing this will reduce the residual sum of squares (see Chapter 10) and so will increase the precision of the estimated effect of the main exposure, and the power of the associated hypothesis test. However, this is not the case with other **generalized linear models**. For example, inclusion of additional variables in logistic regression models will tend to *increase* the standard error of the exposure effect estimate.

Estimating the effect of a particular exposure

When estimating the effect of a particular exposure, we have seen that it is important to include potential confounding variables in the regression model, and that failure to do so will lead to a biased estimate of the effect. In considering which potential confounders should be included, it is essential that careful consideration be given to hierarchical relationships between exposures and confounders, as well as to statistical associations in the data. This is explained in detail in Chapter 38 on strategies for data analysis.

Deriving a regression model to predict the outcome

Different considerations apply when the main purpose of the analysis is to derive a regression model that can be used to predict future values of the outcome variable. For example, this approach has been used in developing countries to attempt to identify whether a pregnant woman may be at risk of obstetric difficulties, based on factors such as social class, previous pregnancy outcomes, and pre-pregnancy weight and height.

The aim in developing a predictive model is to identify a set of exposure variables that give a good prediction of the outcome. The emphasis is no longer on assessing the importance of a particular exposure or on understanding the aetiology of the outcome. However, a good starting point is to include those exposure variables that are known from other studies to be strongly associated with the outcome. In addition, it may be helpful to use an automated procedure to identify which (of what are often a large number of additional variables) might be included in the model. Such procedures are usually based on the magnitude of the

P-value for each variable and are known as **stepwise selection procedures**. For example, a typical stepwise procedure might be:

1 Fit a model including all exposure variables. Now omit each variable in turn, and record the *P*-value for each likelihood ratio test. The variable with the highest *P*-value is omitted from the next step of the procedure.
2 Fit the model including all variables except that omitted in step (1). Now proceed as in step (1) to select the next variable to be omitted.
3 Continue until the *P*-value for omission of each remaining variable is less than a chosen threshold (e.g. 0.2).
4 Now consider adding, in turn, each of the variables omitted in steps (1) to (3). Add the variable with the smallest *P*-value, providing this is less than 0.2.
5 Continue until no more variables with a *P*-value of < 0.2 can be added. The resulting model is the final model to be used for prediction.

Of course, different versions of such stepwise procedures can be chosen. Such procedures may appear attractive, because they seem to provide an objective way of choosing the best possible model. However *they have serious disadvantages*, which are summarized in Box 29.2. If it is necessary to use a stepwise selection procedure, then it is advisable to use a higher *P*-value threshold, such as 0.2 rather than 0.05 (the traditional threshold for statistical significance).

BOX 29.2 PROBLEMS WITH STEPWISE VARIABLE SELECTION IN REGRESSION MODELS

1 The major problem with stepwise regression is that the derived model will give an over-optimistic impression. The *P*-values for the selected variables will be too small, confidence intervals will be too narrow and, in the case of multiple regression, the proportion of variance explained (R^2) will be too high. This is because they do not reflect the fact that the model was selected using a stepwise procedure. The higher the original number of exposure variables from which the final model was selected, the higher the chance of selecting variables with chance associations with the outcome and thus the worse this problem will be.
2 The regression coefficients will be too large (too far away from their null values). This means that the performance of the model in predicting future values of the outcome will be less good than we might expect.
3 Computer simulations have shown minor changes in the data may lead to important changes in the variables selected for the final model.
4 Stepwise procedures should never be used as a substitute for thinking about the problem. For example, are there variables that should be included because they are known from previous work to be associated with the outcome? Are there variables for which an association with the outcome is implausible?

The quality of predictions from models that have been derived using stepwise procedures should be evaluated using a separate dataset (the **test dataset**) to that which was used to derive the model (the **development dataset**). This is for two reasons:

- as explained in Box 29.2, the regression coefficients in the model will tend to be too large;
- the individuals for whom we wish to predict the outcome may differ, in a manner not captured by the variables measured, from those in the development dataset.

Developing an explanatory model for the outcome

Sometimes the focus of a study is to understand the aetiology of the outcome, and to identify those exposures or risk factors that are important influences on it. The purpose of the regression model here is halfway between that of the other two situations just described. Thus the focus is neither on identifying which confounders to include for a particular risk factor, nor is it on identifying any combination of exposures that works, as in the prediction scenario. Instead it is intended to attach meaning to the variables chosen for inclusion in the final model. For this reason, we strongly recommend that the selection procedure is based on an underlying conceptual framework (see Chapter 38 for more detail), and that formal stepwise methods are avoided because of the problems with them described in Box 29.2.

CHAPTER 30

Relaxing model assumptions

30.1 INTRODUCTION

All the statistical methods presented so far have been based on assuming a specific probability distribution for the outcome, or for a transformation of the outcome. Thus we have assumed a normal distribution for numerical outcomes, a binomial distribution for binary outcomes and a Poisson distribution for rates. In this chapter we describe three types of methods that can be used when these assumptions are violated. These are:

- *non-parametric methods based on ranks*, which are used when we have a numerical outcome variable but wish to avoid specific assumptions about its distribution, or cannot find a transformation under which the outcome is approximately normal;
- *bootstrapping*, a very general technique that allows us to derive confidence intervals making only very limited assumptions about the probability distribution of the outcome;
- *robust standard errors*, which allow derivation of confidence intervals and standard errors based on the actual distribution of the outcome variable in the dataset rather than on an assumed underlying probability distribution.

30.2 NON-PARAMETRIC METHODS BASED ON RANKS

Non-parametric methods based on ranks are used to analyse a numerical outcome variable without assuming that it is approximately normally distributed. The key feature of these methods is that each value of the outcome variable is replaced by its rank after the variable has been sorted into ascending order of magnitude. For example, if the outcome values were 453, 1, 5 and 39 then analyses would be based on the corresponding ranks of 4, 1, 2 and 3.

As explained in Chapter 5, the central limit theorem tells us that as the sample size increases the sampling distribution of a mean will tend to be *normally* distributed even if the underlying distribution is non-normal. Rank methods are therefore particularly useful in a small data set when there is obvious non-normality that cannot be corrected by a suitable transformation, or when we do not wish

to transform the variable because transforming would make interpretation of the results harder. They are less powerful (efficient in detecting genuine differences) than parametric methods, but may be more *robust*, in the sense that they are less affected by extreme observations. Rank methods have three main disadvantages:

1 Their primary concern has traditionally been significance testing, since associated methods for deriving confidence limits have been developed only recently. This conflicts with the emphasis in modern medical statistics on estimation of the size of differences, and the interpretation of *P*-values in the context of confidence intervals (see Chapter 8). In particular, large *P*-values from rank order tests comparing two small samples have often been misinterpreted, in the absence of confidence intervals, as showing that there is no difference between two groups, when in fact the data are consistent either with no difference or with a substantial difference. Bootstrapping, described in Section 30.3, provides a general means of deriving confidence intervals and so overcomes this difficulty.
2 When sample sizes are extremely small, such as in comparing two groups with three persons in each group, rank tests can *never* produce small *P*-values, even when the values of the outcomes in the two groups are very different from each other, such as 1, 2 and 3 compared with 21, 22 and 23. In contrast, the *t-test* based on the normal distribution is able to detect such a clear difference between groups. It will, of course, never be possible to verify the assumption of normality in such small samples.
3 Non-parametric methods are less easily extended to situations where we wish to take into account the effect of more than one exposure on the outcome. For these reasons the emphasis in this book is on the use of parametric methods, providing these are valid.

The main rank-order methods are listed in Table 30.1 together with their parametric counterparts. The most common ones, the Wilcoxon signed rank test, the Wilcoxon rank sum test, Spearman's rank correlation and Kendall's tau, will be described using examples previously analysed using parametric methods. For a detailed account of non-parametric methods the reader is referred to Conover (1999), Siegel and Castellan (1988) or Sprent and Smeeton (2000). Details of methods to derive confidence intervals are given by Altman *et al.* (2000).

Wilcoxon signed rank test

This is the non-parametric counterpart of the paired *t*-test, and corresponds to a test of whether the median of the differences between paired observations is zero in the population from which the sample is drawn.

Example 30.1

We will show how to derive the Wilcoxon signed rank test using the data in Table 30.2, which shows the number of hours of sleep obtained by 10 patients when they

Table 30.1 Summary of the main rank order methods. Those described in more detail in this section are shown in italics.

Purpose of test	Method	Parametric counterpart
Examine the difference between paired observations	*Wilcoxon signed rank test*	Paired *t*-test
Simplified form of Wilcoxon signed rank test	Sign test	
Examine the difference between two groups	*Wilcoxon rank sum test*	Two-sample *t*-test
Alternatives to Wilcoxon rank sum test that give identical results	Mann–Whitney *U*-test Kendall's *S*-test	Two-sample *t*-test
Examine the difference between two or more groups. Gives identical results to Wilcoxon rank sum test when there are two groups	Kruskal–Wallis one-way analysis of variance	One-way analysis of variance
Measure of the strength of association between two variables	*Kendall's rank correlation (Kendall's tau)*	Correlation coefficient
Alternative to Kendall's rank correlation that is easier to calculate.	*Spearman's rank correlation*	Correlation coefficient

Table 30.2 Results of a placebo-controlled clinical trial to test the effectiveness of a sleeping drug (reproduced from Table 7.3), with ranks for use in the Wilcoxon signed rank test.

	Hours of sleep			
Patient	Drug	Placebo	Difference	Rank (ignoring sign)
1	6.1	5.2	0.9	2
2	6.0	7.9	−1.9	5
3	8.2	3.9	4.3	10
4	7.6	4.7	2.9	8
5	6.5	5.3	1.2	3
6	5.4	7.4	−2.0	6
7	6.9	4.2	2.7	7
8	6.7	6.1	0.6	1
9	7.4	3.8	3.6	9
10	5.8	7.3	−1.5	4

took a sleeping drug and when they took a placebo, and the differences between them. The test consists of five steps:

1 Exclude any differences that are zero. Put the remaining differences in ascending order of magnitude, *ignoring* their signs and give them **ranks** 1, 2, 3, etc., as shown in Table 30.2. If any differences are equal then average their ranks.

2 Count up the ranks of the positive differences and of the negative differences and denote these sums by T_+ and T_- respectively.

$$T_+ = 2 + 10 + 8 + 3 + 7 + 1 + 9 = 40$$
$$T_- = 5 + 6 + 4 = 15$$

3 If there were no difference in effectiveness between the sleeping drug and the placebo then the sums T_+ and T_- would be similar. If there were a difference then one sum would be much smaller and one sum would be much larger than expected. Denote the smaller sum by T.

$$T = \text{smaller of } T_+ \text{ and } T_-$$

In this example, $T = 15$.

4 The Wilcoxon signed rank test is based on assessing whether T is smaller than would be expected by chance, under the null hypothesis that the median of the paired differences is zero. The P-value is derived from the sampling distribution of T under the null hypothesis. A range for the P-value can be found by comparing the value of T with the values for $P = 0.05$, $P = 0.02$ and $P = 0.01$ given in Table A7 in the Appendix. Note that the appropriate sample size, n, is the number of differences that were ranked rather than the total number of differences, and does not therefore include the zero differences.

$$n = \text{number of non-zero differences}$$

In contrast to the usual situation, the *smaller* the value of T the *smaller* is the P-value. This is because the null hypothesis is that T is equal to the sum of the ranks divided by 2, so that the smaller the value of T the more evidence there is against the null hypothesis. In this example, the sample size is 10 and the 5%, 2% and 1% percentage points are 8, 5 and 3 respectively. The P-value is therefore greater than 0.05, since 15 is greater than 8.

It is more usual to derive the P-value using a computer: in this example $P = 0.20$ so there is no evidence against the null hypothesis, and hence no evidence that the sleeping drug was more effective than the placebo.

5 To derive an approximate 95% confidence interval for the median difference, we consider the averages of the $n(n+1)/2$ possible pairs of differences. The resulting $10 \times 11/2 = 55$ possible averages for this example are shown in Table 30.3. The approximate 95% CI is given by:

95% CI (median difference) = T^{th} smallest average to T^{th} largest average of the $n(n+1)/2$ possible pairs of differences, where T is the value corresponding to the 2-sided P-value of 0.05 in Table A7

Table 30.3 Fifty-five possible averages of the ten differences between patients' hours of sleep after taking a sleeping drug and their hours of sleep after taking a placebo.

	−2.0	−1.9	−1.5	0.6	0.9	1.2	2.7	2.9	3.6	4.3
−2.0	−2	−1.95	−1.75	−0.7	−0.55	−0.4	0.35	0.45	0.8	1.15
−1.9		−1.9	−1.7	−0.65	−0.5	−0.35	0.4	0.5	0.85	1.2
−1.5			−1.5	−0.45	−0.3	−0.15	0.6	0.7	1.05	1.4
0.6				0.6	0.75	0.9	1.65	1.75	2.1	2.45
0.9					0.9	1.05	1.8	1.9	2.25	2.6
1.2						1.2	1.95	2.05	2.4	2.75
2.7							2.7	2.8	3.15	3.5
2.9								2.9	3.25	3.6
3.6									3.6	3.95
4.3										4.3

In this example, $T = 8$, and so the 95% confidence interval is from the 8th smallest average to the 8th largest average. These are found from Table 30.3 to be −0.65 and 2.9 respectively.

95% confidence interval for median difference = −0.65 to 2.9

Further details of the assumptions underlying the Wilcoxon signed rank test and the confidence interval for the median difference are given in Conover (1999).

Wilcoxon rank sum test

This is one of the non-parametric counterparts of the *t*-test, and is used to assess whether an outcome variable differs between two exposure groups. Specifically, it examines whether the median difference between pairs of observations from the two groups is equal to zero. If, in addition, we assume that the distributions of the outcome in the two groups are identical except that they differ by a constant amount (that is, they 'differ only in location') then the null hypothesis of the test is that the difference between the medians of the two distributions equals zero.

Example 30.2

The use of the Wilcoxon rank sum test will be described by considering the data in Table 30.4, which shows the birth weights of children born to 15 non-smokers and 14 heavy smokers. It consists of three steps:

1. Rank the values of the outcome from both groups together in *ascending* order of magnitude, as shown in the table. If any of the values are equal, average their ranks.
2. Add up the ranks in the group with the smaller sample size. If there were no difference between the groups then the ranks would on average be similar. In

Table 30.4 Comparison of birth weights of children born to 15 non-smokers with those of children born to 14 heavy smokers (reproduced from Table 7.1), with ranks for use in the Wilcoxon rank sum test.

Non-smokers ($n = 15$)		Heavy smokers ($n = 14$)	
Birth weight (kg)	Rank	Birth weight (kg)	Rank
3.99	27	3.18	7
3.89	26	2.74	4
3.6*	17.5	2.9	6
3.73	24	3.27	9
3.31	10	3.65†	20.5
3.7	23	3.42	13
4.08	28	3.23	8
3.61	19	2.86	5
3.83	25	3.6*	17.5
3.41	12	3.65†	20.5
4.13	29	3.69	22
3.36	11	3.53	15
3.54	16	2.38	2
3.51	14	2.34	1
2.71	3		
	Sum = 284.5		Sum = 150.5

*Tied 17th and 18th and so ranks averaged
†Tied 20th and 21st and so ranks averaged

this case the group with the smaller sample size is the heavy smokers, and their ranks sum to 150.5. If the two groups are of the same size either one may be picked.

$$T = \text{sum of ranks in group with smaller sample size}$$

3 Compare the value of T with the values in Table A8, which is arranged somewhat differently to the tables for the other tests. Look up the row corresponding to the sample sizes of the two groups, in this case row 14, 15. The range shown for $P = 0.01$ is 151 to 269: values inside this range (i.e. between 151 and 269) correspond to P-values greater than 0.01. Sums of 151 and below or 269 and above correspond to P-values less than 0.01. The sum of 150.5 in this example is just below the lower limit of 151, so the P-value is slightly less than 0.01.

As with the signed rank test, the P-value is usually derived using a computer. In this case $P = 0.0094$: there is good evidence against the null hypothesis that the median birth weight of children born to heavy smokers is the same as the median birth weight of children born to non-smokers.

Details of how to derive a confidence interval for the difference in medians (assuming that the two distributions differ only in location) are given by Conover

(1999) and in Altman *et al.* (2000). Such confidence intervals are known as **Hodges–Lehmann** estimates of shift. In this example, we find (using a computer) that the 95% CI is from -0.77 to -0.09. In Section 30.3 we see how bootstrap methods can also be used to provide a confidence interval for the difference between medians.

Rank correlations

We will now consider two rank order measures of the association between two numerical variables: Kendall's tau and Spearman's rank correlation. The parametric counterpart of these measures is the correlation coefficient, sometimes known as the **Pearson product moment correlation**, which was described in Chapter 10.

Example 30.3

We will explain these measures of association using the data in Table 30.5 on the relationship between plasma volume and body weight in eight healthy men. We will call these two quantitative variables Y and X. The Pearson correlation between these was shown in Section 10.3 to be 0.76.

To calculate **Spearman's rank correlation** coefficient r_s:

1 Independently rank the values of X and Y.

2 Calculate the Pearson correlation between the ranks, rather than between the original measurements. Other formulae for the Spearman correlation are often quoted; these give identical results. This gives a value of 0.81 in this example.

Kendall's tau (denoted by the Greek letter τ) is derived as follows:

1 Compare the ranks of X and Y between each pair of men. There are $n(n-1)/2$ possible pairs. The pairs of ranks for subjects i and j are said to be:

(a) *concordant* if they differ in the same directions, that is if both the X and Y ranks of subject i are lower than the corresponding ranks of subject j, or both are higher. For example, the ranks of subjects 1 and 2 are concordant

Table 30.5 Relationship between plasma volume and body weight in eight healthy men (reproduced from Table 10.1), with ranks used in calculating the Spearman rank correlation.

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

as subject 1 has a lower rank than subject 2 for both the variables. The pair 3 and 8 is also concordant: subject 3 has higher ranks than subject 8 on both variables.

(b) *discordant* if the comparison of the ranks of the two variables is in opposite directions. For example, the ranks of subjects 3 and 6 are discordant as subject 3 has a more highly ranked X value than subject 6 but a lower ranked Y value.

2 Count the number of concordant pairs (n_C) and the number of discordant pairs (n_D), and calculate τ as:

$$\tau = \frac{n_C - n_D}{n(n-1)/2}$$

In this example, Kendall's tau (derived using a computer) is 0.64. If all pairs are concordant then $\tau = 1$, while if all pairs are discordant then $\tau = -1$. More details, including an explanation of how to deal with ties, are given by Conover (1999).

All three measures of correlation have values between 1 and -1. Although Spearman's rank correlation is better known, its only advantage is that it is easier to calculate without a computer. Kendall's tau is the preferred rank measure, because its statistical properties are better and because it is easier to interpret. Given two pairs of observations (X_1, Y_1) and (X_2, Y_2) Kendall's tau is the difference between the probability that the bigger X is with the bigger Y, and the probability that the bigger X is with the smaller Y.

If X and Y are each normally distributed then there is a direct relationship between the Pearson correlation (r) and both Kendall's τ and Spearman's rank correlation (r_s):

$$r = \sin\left(\frac{\pi}{2}\tau\right) = 2\sin\left(\frac{\pi}{6}r_s\right)$$

This means that Pearson correlations of 0, $\pm 1/2$, ± 0.7071 and $\pm$ 1 correspond to Kendall τ values of 0, $\pm 1/3$, $\pm 1/2$ and $\pm$ 1 and to Spearman rank correlations of 0, $\pm$ 0.4826, $\pm$ 0.6902 and $\pm$ 1, respectively.

30.3 BOOTSTRAPPING

Bootstrapping is a way of deriving confidence intervals while making only very limited assumptions about the probability distribution that gave rise to the data. The name derives from the expression 'pull yourself up by your bootstraps', which means that you make progress through your own efforts; without external help. It

is based on a remarkably simple idea: that if we take repeated samples from the data themselves, mimicking the way that the data were sampled from the population, we can use these samples to derive standard errors and confidence intervals.

The new samples are drawn *with replacement* from the original data. That is, we pick an observation at random from the original data, note down its value, then pick another observation at random from the same original data, regardless of which observation was picked first. This continues until we have a new dataset of the same size as the original one. The samples differ from each other because some of the original observations are picked more than once, while others are not picked at all.

Example 30.4

We will illustrate this using the data on birth weight and smoking, shown in Table 30.4. The median birth weight among the children born to the non-smokers was 3.61 kg, while the median among children born to the smokers was 3.25 kg. The difference in medians comparing smokers with non-smokers was therefore −0.36 kg. The *P*-value for the null hypothesis that the median birth weight is the same in smokers and non-smokers (derived in Section 30.2 using the Wilcoxon rank sum test) was 0.0094. The non-smokers and heavy smokers were recruited separately in this study, and so the bootstrap sampling procedure mimics this by sampling separately from the non-smokers and from the heavy smokers. Therefore each bootstrap sample will have 15 non-smokers and 14 heavy smokers.

This process is illustrated, for two bootstrap samples, in Table 30.6. In the first bootstrap sample observations 1, 3, 4 and 5 were not picked, observation 2 was picked four times, observations 6 and 7 were picked once and so on. In this sample the difference in median birth weight was −0.48 kg, while in the second sample the difference was −0.26 kg.

We repeat this procedure a large number of times, and record the difference between the medians in each sample. To derive confidence intervals, a minimum of around 1000 bootstrap samples is needed. Figure 30.1 is a histogram of the differences in medians derived from 1000 bootstrap samples from the birth weight data.

The simplest way to derive a 95% confidence interval for the difference between medians is to use the *percentile method* and take the range within which 95% of these bootstrap differences lie, i.e. from the 2.5^{th} percentile to the 97.5^{th} percentile of this distribution. This gives a 95% CI of −0.87 to −0.01 kg.

Unfortunately the percentile method, though simple, is not the most accurate method for deriving bootstrap confidence intervals. This has led to the development of **bias corrected** (BC) and **bias corrected and accelerated** (BCa) **intervals**, of which BCa intervals have been shown to have the best properties. For the birth weight data, use of the BC method gives a 95% CI of −0.80 to 0.025 kg, while the BCa method gives a 95% CI of −0.71 to 0.12 kg. More information about the use of bootstrap methods can be found in Efron and Tibshirani (1993) and in Davison and Hinkley (1997).

Table 30.6 Two bootstrap samples, based on data on birth weights (kg) of children born to 15 non-smokers and of children born to 14 heavy smokers.

Original data			First bootstrap sample			Second bootstrap sample		
Obs. no.	Birth weight	Smoker	Original obs. no.	Birth weight	Smoker	Original obs. no.	Birth weight	Smoker
1	3.99	No	2	3.89	No	1	3.99	No
2	3.89	No	2	3.89	No	1	3.99	No
3	3.60	No	2	3.89	No	2	3.89	No
4	3.73	No	2	3.89	No	3	3.60	No
5	3.31	No	6	3.70	No	3	3.60	No
6	3.70	No	7	4.08	No	4	3.73	No
7	4.08	No	8	3.61	No	6	3.70	No
8	3.61	No	8	3.61	No	6	3.70	No
9	3.83	No	8	3.61	No	8	3.61	No
10	3.41	No	9	3.83	No	8	3.61	No
11	4.13	No	9	3.83	No	9	3.83	No
12	3.36	No	10	3.41	No	12	3.36	No
13	3.54	No	11	4.13	No	12	3.36	No
14	3.51	No	11	4.13	No	12	3.36	No
15	2.71	No	15	2.71	No	15	2.71	No
16	3.18	Yes	16	3.18	Yes	19	3.27	Yes
17	2.74	Yes	19	3.27	Yes	19	3.27	Yes
18	2.90	Yes	19	3.27	Yes	19	3.27	Yes
19	3.27	Yes	20	3.65	Yes	21	3.42	Yes
20	3.65	Yes	20	3.65	Yes	22	3.23	Yes
21	3.42	Yes	20	3.65	Yes	22	3.23	Yes
22	3.23	Yes	20	3.65	Yes	23	2.86	Yes
23	2.86	Yes	21	3.42	Yes	25	3.65	Yes
24	3.60	Yes	24	3.60	Yes	25	3.65	Yes
25	3.65	Yes	26	3.69	Yes	25	3.65	Yes
26	3.69	Yes	28	2.38	Yes	26	3.69	Yes
27	3.53	Yes	29	2.34	Yes	27	3.53	Yes
28	2.38	Yes	29	2.34	Yes	27	3.53	Yes
29	2.34	Yes	29	2.34	Yes	29	2.34	Yes
Median in non-smokers = 3.61			Median in non-smokers = 3.83			Median in non-smokers = 3.61		
Median in smokers = 3.25			Median in smokers = 3.35			Median in smokers = 3.35		
Difference in medians = −0.36			Difference in medians = −0.48			Difference in medians = −0.26		

We have illustrated the use of bootstrapping using a simple comparison of medians, but the method is quite general and can be used to derive confidence intervals for any parameter of a statistical model. For example, we might fit a regression model for the effect of smoking on birthweight, controlling for a number of other variables, then derive a bootstrap confidence interval by repeating this regression on 1000 different bootstrap samples and recording the value of the regression coefficient estimated in each. An example of the derivation of different types of bootstrap confidence interval for proportional hazards models is given by Carpenter and Bithell (2000). *If the model assumptions are not*

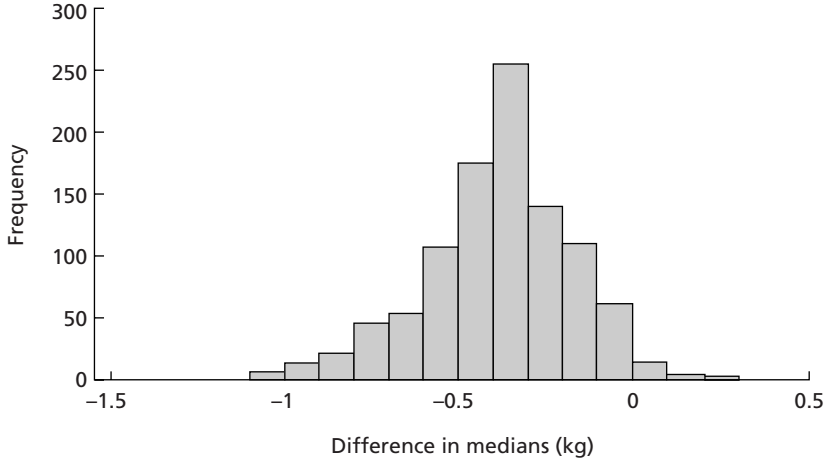

Fig. 30.1 Histogram of the differences in medians (kg) derived from 1000 bootstrap samples of the data on birth weight and smoking.

violated then the bootstrap confidence interval should be similar to the usual confidence interval reported in the regression output.

An example of the use of bootstrapping is provided by Thompson and Barber (2000), who consider the analysis of data on costs of treatment in clinical trials. Costs are often highly skewed, because a small minority of patients incur much higher costs of treatment than the rest. Because of this such data have often been analysed by log transforming the costs and performing a *t*-test. This is a valid approach, but it will lead to an estimate of the difference in mean log costs (which can be converted to a ratio of geometric mean costs, see Chapter 13). The problem is that health service planners are interested in a comparison of mean costs and not in a comparison of mean *log* costs, or in the difference in median costs that might be evaluated using non-parametric methods. Bootstrapping provides a way of deriving confidence intervals for the difference in mean costs between two groups, in circumstances when the non-normality of costs means that confidence intervals from standard methods (*t*-tests or regression) may not be valid.

30.4 ROBUST STANDARD ERRORS

It was explained in Chapter 28 that when we estimate parameters using the likelihood approach then the standard error of the parameter estimate is derived from the curvature of the likelihood at the maximum – the more information which the data provide about the parameter the more sharply curved is the likelihood and the smaller the standard error. Throughout this book we have used such *model-based* standard errors to derive confidence intervals and *P*-values.

Sometimes, we are not confident that the precise probability model underlying the likelihood is correct, and so we may not wish to rely on the likelihood to

provide standard errors for our parameter estimates. Examples of this situation are when the residuals in a multiple regression model are clearly non-normal (see Chapter 12) or when the data are clustered (as discussed in Chapter 31).

An alternative approach, suggested independently by Huber (1967) and White (1980) is to estimate standard errors using the variability in the data. The formula is based on the **residuals** (the difference between the outcome and its predicted value in the regression model, see Section 12.3). Standard errors estimated in this way are known as **robust standard errors** and the corresponding variance estimate is known as the **sandwich variance estimate**, because of the mathematical form of the formula used to estimate it. If the sample size is large enough then, *providing that our basic regression model for the mean of the outcome given the level of the exposure variables is correct*, robust standard errors will be correct, even if the probability model for the outcome variable is wrong. Robust standard errors thus provide a general means of checking how reasonable are the model-based standard errors (which are calcuated assuming that the probability model is correct).

Example 30.5

In Section 11.3 we fitted a multiple regression model of lung function (FEV_1, litres) on age, height and gender among 636 children aged 7 to 10 years living in a suburb of Lima, Peru. However, in Section 12.3 we saw that there may be an association between the residuals and predicted values in this regression model: if this association is real, it would violate an assumption underlying the regression model.

Table 30.7 shows the results of re-analysing these data specifying robust standard errors, compared to the results using model-based standard errors. Note that the regression coefficients are the same whichever we use. The effect of specifying robust standard errors varies for each of the exposure variables. For age and gender (variable 'male') the standard error is only slightly increased but for height the standard error is increased by about 17%, with a corresponding reduction in the *t*-statistic (from 14.04 to 11.51) and an increase in the width of the confidence intervals. In this example, our conclusions are broadly similar whether we use model-based or robust standard errors.

Table 30.7 Regression coefficients, model-based standard errors and robust standard errors, each with corresponding *t*-statistics from the linear regression model relating FEV_1 to age, height and gender of the child in the Peru study.

	Regression	Model-based standard error		Robust standard error	
FEV_1	coefficient	s.e.	*t*	s.e.	*t*
Age	0.0946	0.0152	6.23	0.0159	5.96
Height	0.0246	0.0018	14.04	0.0021	11.51
Male	0.1213	0.0176	6.90	0.0177	6.87
Constant	−2.360	0.1750	−13.49	0.208	−11.34

CHAPTER 31

Analysis of clustered data

31.1 INTRODUCTION

The statistical methods discussed so far in this book are based on the assumption that the observations in a sample are **independent** of each other, that is the value of one observation is not influenced by the value of another. This assumption of independence will be violated if the data are **clustered**, that is if observations in one cluster tend to be *more similar* to each other than to individuals in the rest of the sample. Clustered data arise in three main ways:

1 **Repeated measures in longitudinal studies**. In this case the clusters are the subjects; repeated observations on the same subject will be more similar to each other than to observations on other subjects. For example:
   - in studies of asthma or other chronic diseases, episodes of disease may occur on more than one occasion in the same subject;
   - in longitudinal studies of common childhood diseases in developing countries, children may experience several episodes of diarrhoea, malaria or acute respiratory infections during the course of the study;
   - in a study of cardiovascular disease and obesity, measurements of blood pressure, body mass index and cholesterol levels may be repeated every 3 months.
2 **Multiple measures on the same subject.** For example, in dental research observations are made on more than one tooth in the same subject. In this case the clusters are again subjects.
3 **Studies in which subjects are grouped.** This occurs for example in:
   - **cluster randomized trials** (see Chapter 34), in which groups rather than individuals are randomized to receive the different interventions under trial. For example, the unit of randomization might be general practices, with all patients registered in a practice receiving the same intervention. Since patients in a general practice may be more similar to each other than

to patients in other general practices, for example because some areas tend to be more deprived than others or because of exposure to a common environmental hazard, the data are clustered. In this case the cluster is the group of patients registered with a general practice;

- **family studies**, since individuals in the same family are likely to be more similar to each other than to individuals in different families, because they share similar genes and a similar environment. In this case the cluster is the family;
- surveys where **cluster sampling** is employed (see Chapter 34). For example, in order to estimate the percentage of 14-year-olds in London that work at weekends, we might select 1000 children by randomly sampling 20 schools from all the schools in London, then randomly sample 50 children from each of the selected schools. As the children within a school may be more similar to each other than to children in different schools, the data are clustered. In this case the clusters are the schools.

It is essential that the presence of clustering is allowed for in statistical analyses. The main reason for this, as we shall see, is that *standard errors may be too small* if they do not take account of clustering in the data. This will lead to confidence intervals that are too narrow, and *P*-values that are too small.

We will discuss four appropriate ways to analyse clustered data:

1 calculate **summary measures** for each cluster, and analyse these summary measures using standard methods;
2 use **robust standard errors** to correct standard errors for the clustering;
3 use **random effects** models which explicitly model the similarity between individuals in the same cluster;
4 use **generalized estimating equations (GEE)** which adjust both standard errors and parameter estimates to allow for the clustering.

We will illustrate the importance of taking clustering into account in the context of the following hypothetical example.

Example 31.1

In a study of the effect of 'compound X' in drinking water on rates of dental caries, 832 primary school children in eight different schools were monitored to ascertain the time until they first required dental treatment. Table 31.1 shows data for the first 20 children in the study (all of whom were in school 1). Since compound X is measured at the school level, it is constant for all children in the same school. The data are therefore clustered and the clusters are the eight schools.

Table 31.2 summarizes the data for each school by showing the number of children requiring dental treatment, the total child-years of follow-up, the treatment rate per 100 child-years and the level of compound X in the school's drinking water. Results from a Poisson regression analysis of these data are shown in Table 31.3. This shows strong evidence that increased levels of compound X were associated with decreased rates of dental treatment among the school children.

Table 31.1 Data on the first 20 children in a study of the relationship between rates of dental treatment and the level of compound X in drinking water.

Child's id	Years of follow up	Required dental treatment during follow up?	School number	Level of compound X in school's water supply (1000 × ppm)
1	4.62	No	1	7.1
2	3.00	No	1	7.1
3	4.44	No	1	7.1
4	3.89	No	1	7.1
5	3.08	No	1	7.1
6	2.45	Yes	1	7.1
7	2.64	Yes	1	7.1
8	4.16	No	1	7.1
9	4.25	No	1	7.1
10	2.02	Yes	1	7.1
11	3.13	No	1	7.1
12	3.49	No	1	7.1
13	4.75	No	1	7.1
14	2.39	Yes	1	7.1
15	3.66	No	1	7.1
16	3.43	No	1	7.1
17	2.63	Yes	1	7.1
18	4.21	No	1	7.1
19	2.63	Yes	1	7.1
20	2.74	No	1	7.1

Table 31.2 Total child-years of follow-up, treatment rate per 100 child-years and the level of compound X in each school's drinking water, from a study of the effect of compound X in drinking water on the 832 children attending eight primary schools.

School	Number of children requiring dental treatment	Child-years of follow-up	Rate per 100 child-years	Level of compound X (1000 × ppm)
1	46	456.3	10.08	7.1
2	19	215.1	8.83	7.6
3	17	487.8	3.49	8.2
4	46	459.9	10.00	5.4
5	15	201.2	7.46	8.4
6	20	187.7	10.66	6.8
7	58	399.1	14.53	6.2
8	20	212.5	9.41	8.9

However, treatment rates among different children in the same school may tend to be more similar than treatment rates in children in different schools for reasons unrelated to the levels of compound X in the water, for example because children in the same school are of similar social background. There would then be more observed *between-school variability* than would be expected in the absence of clustering, in which case the strength of the association between treatment rates

Table 31.3 Poisson regression of the effect of compound X in drinking water on rates of dental treatment among 832 children attending eight primary schools.

(a) Results on rate ratio scale

	Rate ratio	z	$P > \|z\|$	95% CI
Compound X	0.821	−3.47	0.001	0.734 to 0.918

(b) Results on log scale

	Coefficient	s.e.	z	$P > \|z\|$	95% CI
Compound X	−0.1976	0.0570	−3.47	0.001	−0.3094 to −0.0859
Constant	3.6041	0.3976	9.07	0.000	2.8248 to 4.3833

and levels of compound X may be exaggerated by the analysis in Table 31.3, which does not allow for such clustering.

31.2 ANALYSES USING SUMMARY MEASURES FOR EACH CLUSTER

The simplest way to analyse clustered data is to derive summary measures for each cluster. Providing that the outcomes in different clusters are independent, standard methods may then be used to compare these summary measures between clusters.

Example 31.1 (continued)

For example, we might analyse the compound X data by doing a linear regression of the log treatment rate in each school on levels of compound X in the school. Results of such a regression are shown in Table 31.4. The estimated increase in the log rate ratio per unit increase in level of compound X is −0.1866, similar to the value of −0.1976 estimated in the Poisson regression analysis in Table 31.3. However, the standard error is much larger and there is now no evidence of an association ($P = 0.177$). Note that the estimated rate ratio per unit increase in level of compound X is simply $\exp(-0.1866) = 0.830$, and that 95% confidence limits for the rate ratio may be derived in a similar way, from the 95% CI in the regression output.

The regression analysis in Table 31.4 is a valid way to take into account the clustering in the data. It suggests that the standard error for the compound X effect in the Poisson regression analysis in Table 31.3 was too small, and therefore that the assumption made in that analysis, that treatment rates among different children in the same school were statistically independent, was incorrect. Thus this analysis using summary measures has confirmed the presence of clustering within schools.

Table 31.4 Linear regression of the effect of compound X on the log of the treatment rate in each school.

	Coefficient	s.e.	t	$P > t$	95% CI
Compound X	−0.1866	0.1220	−1.53	0.177	−0.4850 to 0.1119
Constant	3.5334	0.9035	3.91	0.008	1.3227 to 5.7441

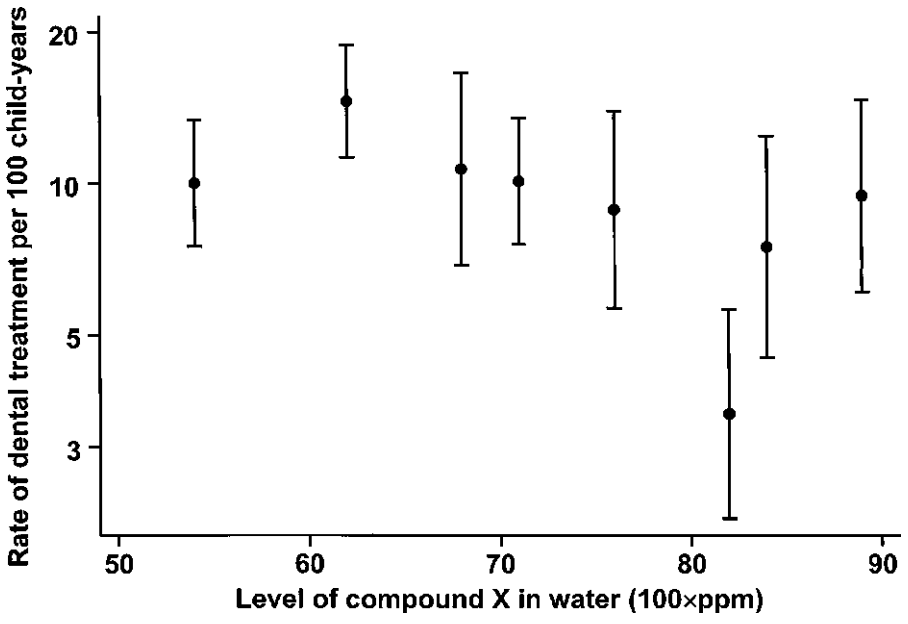

Fig. 31.1 Rate of dental treatment in each school (log scale), with corresponding 95% confidence intervals shown by the vertical lines.

Although analyses based on summary measures may be perfectly adequate in some circumstances, they can have disadvantages:

1 They do *not enable us to estimate the effect of characteristics of individuals within the cluster*. For example, rates of treatment might vary according to the age and gender of the children. Similarly, in a longitudinal study of factors associated with episodes of asthma, this approach would not allow us to examine whether subjects who had a viral infection were at increased risk of an episode of asthma during the subsequent week.
2 They take *no account of the precision with which each of the cluster measures is estimated*. In this example, the cluster measures are the rates in each school. The more events (children requiring treatment), the more precise is the estimated rate. For example, in school 5 only 15 children required treatment while in school 7, 58 children required treatment. The varying precision with which the treatment rate in each school is estimated is illustrated by the varying widths of the confidence intervals in Figure 31.1.

31.3 USE OF ROBUST STANDARD ERRORS TO ALLOW FOR CLUSTERING

As explained in the last section, the presence of clustering means that the standard errors obtained from the usual regression model will be too small. In Chapter

30 we introduced robust standard errors, which are estimated using the variability in the data (measured by the residuals) rather than the variability assumed by the statistical model. We can use a modified type of robust standard error as another approach to correct for clustering. To do this we add the residuals within each cluster together, and then use the resulting cluster-level residuals to derive standard errors that are valid in the presence of clustering.

Example 31.1 (continued)

Table 31.5 shows the results from a Poisson regression analysis based on robust standard errors that allow for within-school clustering. The rate ratio is identical to that from the standard Poisson regression analysis shown in Table 31.3, but the standard error of the log rate ratio has increased from 0.0570 to 0.1203. This analysis gives similar results to the linear regression analysis using summary measures shown in Table 31.4: there is at most weak evidence for an association between levels of compound X and treatment rates. However, *because the analysis is based on individual children we could now proceed to control for the effect of child characteristics*.

Important points to note in the use of robust standard errors to correct standard errors for clustering are:

- Robust standard errors use cluster-level residuals to take account of the similarity of individuals in the same cluster. In the presence of clustering, they will be larger than standard errors obtained from the usual regression model ignoring clustering.
- Use of robust standard errors does not affect the parameter estimate.
- Robust standard errors will be correct providing our model is correct and we have a reasonable number of clusters (≥ 30).
- The log likelihood is not affected when we specify robust standard errors, and so *likelihood ratio tests do not take account of the clustering*. Wald tests must therefore be used.

Table 31.5 Poisson regression of the effect of compound X levels in drinking water on rates of dental treatment in eight primary schools, using robust standard errors to allow for the clustering.

(a) Results on rate ratio scale

	Rate ratio	z	$P > \|z\|$	95% CI
Compound X	0.821	−1.643	0.100	0.648 to 1.039

(b) Results on log scale

	Coefficient	s.e.	z	$P > \|z\|$	95% CI
Compound X	−0.1976	0.1203	−1.643	0.100	−0.4333 to 0.0381
Constant	3.6041	0.8147	4.42	0.000	2.0073 to 5.2008

31.4 RANDOM EFFECTS (MULTILEVEL) MODELS

Arguably the most satisfactory approach to the analysis of clustered data is to use **random effects** models that *explicitly* allow for the clustering. The simplest such models allow the average response to vary between clusters. This is done by modifying the standard linear predictor (see Section 29.2) to include an amount that varies randomly between clusters:

$$\text{linear predictor for an individual in cluster } j = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + u_j$$

The **random effect** u_j is assumed to have mean zero, and to vary randomly between clusters. It is assumed that the set of random effects $\{u_j\}$ explain the clustering in the data so that, having allowed for the random effects, different observations in the same cluster are independent. Random effects models are also known as **multilevel models**, because of the hierarchical data structure in which observations at the first level (the individuals) are nested within observations at the second level (the cluster). Table 31.6 shows common assumptions made for the distribution of the random effects for different types of regression models.

For numerical outcomes it is usual to assume that both the outcome variable within clusters and the random effects are normally distributed; the resulting distribution is also normal. For Poisson regression models, it is commonly assumed that the random effects $\{u_j\}$ have a **gamma** distribution, which is a generalization of the χ^2 distribution. The combination of the Poisson distribution for the outcome within clusters and the gamma distribution of the random effects leads to a distribution called the negative binomial, so such random effects models are also called **negative binomial** regression models. For logistic regression models, there is no such mathematically well-defined 'composite' distribution, and estimation of these random-effects models has until recently been either unavailable or difficult and time-consuming.

Random-effects models are now available in a number of statistical computer packages, and are fairly straightforward to fit. In addition, specialist software packages are available. The relevant routines are referred to as **random effects models**, **mixed models**, **multilevel models**, **hierarchical models** and **cross-sectional time series** depending on the particular package. The latter name arises from the

Table 31.6 Distribution used for random effects in commonly used regression models.

Type of outcome	Type of standard regression	Distribution of random effects
Numerical	Linear	Normal
Binary	Logistic	Normal
Rate	Poisson	Gamma

use of this approach for repeated measures in longitudinal data (see Section 31.1), but is equally applicable to other types of clustered data.

Example 31.1 (continued)

Table 31.7 shows results from a random-effects Poisson regression analysis of the effect of levels of compound X in drinking water on rates of dental treatment. Compared to the standard Poisson regression model shown in Table 31.3 the log rate ratio is only slightly changed, but after allowing for the clustering the standard error is much larger than the model-based standard error, and there now appears to be at most weak evidence for an association. We can conclude there is more between-school variability than assumed by the Poisson model, because of the increase in the standard error. A **likelihood ratio test for clustering** can be derived by comparing the log-likelihood for this model with that from a standard Poisson regression model.

The standard error from the random effects model (0.1030) is similar to that in the Poisson regression model with robust standard errors (0.1203). Note, however, that in the random effects model both the parameter estimate (the log rate ratio) and its standard error are modified when we allow for clustering.

Table 31.7 Random-effects Poisson regression of the effect of compound X levels in drinking water on rates of dental treatment in eight primary schools, allowing for within-school clustering.

(a) Results on rate ratio scale

	Rate ratio	z	$P > \|z\|$	95% CI
Compound X	0.8333	−1.77	0.077	0.6809 to 1.0198

(b) Results on log scale

	Coefficient	s.e.	z	$P > \|z\|$	95% CI
Compound X	−0.1824	0.1030	−1.77	0.077	−0.3843 to 0.0196
Constant	3.5291	0.7459	4.73	0.000	2.0672 to 4.9909

Example 31.2

In a clinical trial to assess the efficacy and safety of budesonide for the treatment of patients with chronic asthma, 91 patients were treated with a daily dose of 200 μg of budesonide (treatment group) and 92 patients were treated with placebo (control group). The outcome variable was FEV_1 (the maximum volume of air that an individual can exhale in 1 second, see Section 11.2), and this was recorded at baseline (before the start of treatment) and at 2, 4, 8 and 12 weeks after the start of treatment. Figure 31.2 shows that the mean FEV_1 in the treatment and control groups were similar at the start of treatment (as would be expected in a randomized trial) but diverged subsequently: FEV_1 improved in the treatment group but not in the control group.

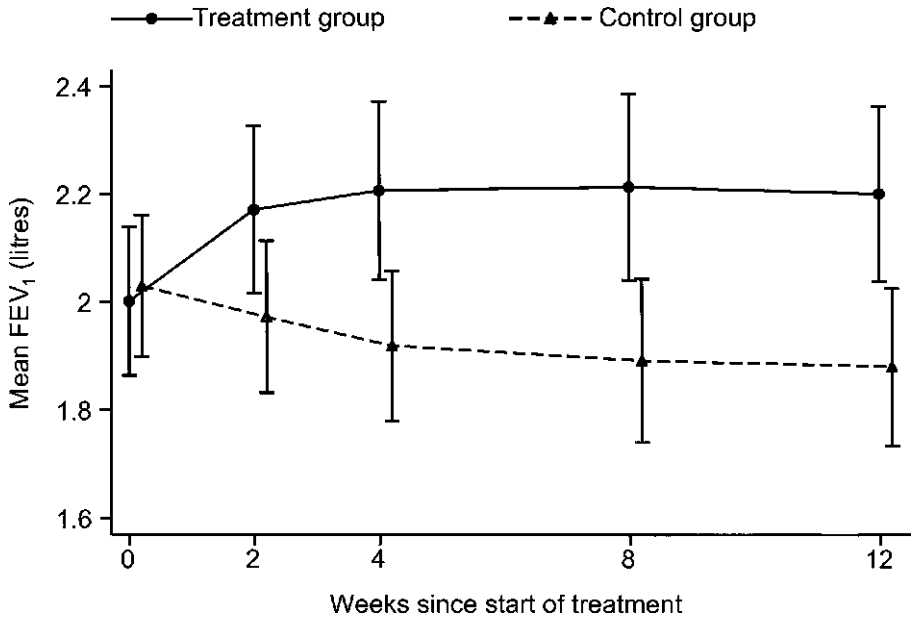

Fig. 31.2 Mean FEV_1 (with 95% CIs) in the treatment and control groups at baseline (0 weeks) and up to 12 weeks from the start of treatment, in a trial of 183 patients with chronic asthma.

Table 31.8 shows the results of three possible analyses of these data that take into account the fact that the means at different times are based on the same two groups of patients:

1. The first uses the average post-treatment FEV_1 for each patient, based on four time points for patients for whom there was complete follow-up, and on one, two or three time points for patients for whom some post-treatment measurements were missed. The linear regression of the mean post-treatment FEV_1 in each subject estimates that the average post-treatment FEV_1 is 0.2998 litres higher for those who received budesonide compared to those who received placebo. Note that this is equivalent to a *t*-test comparing the mean of the average post-treatment FEV_1 measurements between the treatment and control groups.
2. In the second analysis, the linear regression is based on the individual post-treatment measurements with robust standard errors used to allow for clustering of the measurements at different time points within subjects.
3. The third analysis is a random-effects linear regression of the post-treatment FEV_1 in each subject at each time.

The conclusions are similar in each case: treatment increased FEV_1 by a mean of approximately 0.3 litres. Standard errors, and hence confidence intervals and *P*-values, are also similar in the three models.

A random effects model explicitly includes both *between-cluster* and *within-cluster* variation. For a numerical outcome (as in Example 31.2) the model is:

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \ldots + \beta_p x_{pij} + e_{ij} + u_j, \text{ where}$$

y_{ij} is the outcome for individual i in cluster j

x_{1ij} *to* x_{pij} are the values of the p exposure variables for that individual

e_{ij} is *the individual-level* random error, and is normally distributed with mean 0 and variance σ_e^2

u_j is the *cluster-level* random error, and is normally distributed with mean 0 and variance σ_u^2

This model is the same as the multiple regression model described in Section 11.4, with the addition of the cluster-level random effect u_j. The regression output for the random-effects model in Table 31.8(c) shows the estimated between-patient standard deviation ($\sigma_u = 0.6828$) and within-patient standard deviation ($\sigma_e = 0.2464$).

Table 31.8 Regression models to investigate the effect of budesonide treatment on FEV_1 in a clinical trial of 183 patients with chronic asthma. Analyses by kind permission of Dr Carl-Johan Lamm and Dr James Carpenter.

(a) Standard linear regression using the mean post-treatment FEV_1 measurements in each subject

	Coefficient	s.e.	t	$P > \|t\|$	95% CI
Treatment	0.2998	0.1033	2.90	0.004	0.0960 to 0.5037
Constant	1.8972	0.0729	26.04	0.000	1.7534 to 2.0409

(b) Linear regression using the post-treatment FEV_1 measurements in each subject at each time, with robust standard errors allowing for clustering within subjects

	Coefficient	Robust s.e.	t	$P > \|t\|$	95% CI
Treatment	0.2812	0.1044	2.69	0.008	0.0753 to 0.4872
Constant	1.9157	0.0679	28.22	0.000	1.7818 to 2.0497

(c) Random-effects linear regression

	Coefficient	s.e.	z	$P > \|z\|$	95% CI
Treatment	0.2978	0.1028	2.90	0.004	0.0963 to 0.4993
Constant	1.8992	0.0727	26.13	0.000	1.7567 to 2.0416
σ_u	0.6828	0.0370	18.46	0.000	0.6103 to 0.7553
σ_e	0.2464	0.0076	32.23	0.000	0.2314 to 0.2614

Intraclass correlation coefficient

The amount of clustering can be measured using the **intraclass correlation coefficient** (ICC), which is defined as the ratio of the between-cluster variance to the

total variance, which is a combination of the between- and within-cluster variances.

$$\text{Intraclass correlation coefficient, ICC} = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2}$$

If all the variation is explained by differences between clusters, so that there is no variation within clusters and $\sigma_e^2 = 0$, then ICC = 1. If σ_u^2 is estimated to be zero then there is no evidence of clustering and ICC = 0. In Example 31.2,

$$\text{ICC} = 0.6828^2/(0.6828^2 + 0.2464^2) = 0.885$$

so nearly 90% of the variation in FEV_1, after accounting for the effect of treatment, was between patients rather than within patients.

Although the *P*-value for σ_u corresponds to a Wald test of the presence of clustering, it is preferable to test for clustering using a likelihood ratio test; by comparing the log likelihood from the random-effects model (L_{inc}) with the log likelihood from a standard regression model assuming no clustering (L_{exc}).

Including cluster-level and individual-level characteristics in random effects models

The effects on the outcome variable of both cluster characteristics and of characteristics of individual observations within clusters may be included in random-effects models. For the asthma trial data, this corresponds to including characteristics of patients and of observations at different times on the same patient.

Example 31.2 (continued)

In Table 31.9 the random-effects model shown in Table 31.8(c) has been extended to include patients' FEV_1 measurements before the start of treatment (a cluster

Table 31.9 Random-effects linear regression of the effect of budesonide treatment on FEV_1 in a clinical trial of 183 patients with chronic asthma, including baseline FEV_1 and a treatment-time interaction.

	Coefficient	s.e.	*z*	$P > \|z\|$	95% CI
treatment	0.2695	0.0772	3.49	0.000	0.1182 to 0.4207
weeks	−0.0104	0.0035	−2.96	0.003	−0.0173 to −0.0035
treat.weeks	0.0127	0.0049	2.62	0.009	0.0032 to 0.0222
fevbase	0.7562	0.0577	13.12	0.000	0.6432 to 0.8692
constant	0.4039	0.1293	3.12	0.002	0.1504 to 0.6574
σ_u	0.4834	0.0271	17.85	0.000	0.4303 to 0.5364
σ_e	0.2445	0.0076	32.17	0.000	0.2296 to 0.2594

characteristic) and the interaction between treatment and time (a covariate that varies within clusters).

The interpretation of regression coefficients in models including interaction was explained in detail in Section 29.5. Variable *weeks* was coded as time since the 2-week measurement, so the regression coefficient for variable *treatment* estimates the treatment effect (mean difference in FEV_1) at 2 weeks (the baseline of the post-treatment groups) while the regression coefficient for variable *weeks* estimates the mean increase in FEV_1 per week in the control group (the group corresponding to the baseline value of *treatment*). The regression coefficient for the interaction parameter (variable *treat.weeks*) estimates the mean *increase* in the effect of treatment per week: thus the effect of treatment is estimated to increase by 0.0127 litres per week, between week 2 and week 12. As might be expected, there is a strong association between baseline FEV_1 and post-treatment FEV_1 (regression coefficient 0.7562 for variable *fevbase*), and controlling for baseline FEV_1 has substantially reduced the estimated between-patient standard deviation ($\sigma_u = 0.4834$, compared to 0.6828 in the model including only the effect of treatment). The intraclass correlation coefficient from this model is 0.796.

31.5 GENERALIZED ESTIMATING EQUATIONS (GEE)

Estimation of generalized linear models incorporating random effects is difficult mathematically if the outcome is non-normal, except in the case of random-effects Poisson models which exploit a mathematical 'trick' where assuming a particular distribution for the random effect leads to a well-defined 'composite' distribution for the outcome (the negative binomial distribution). For other models, in particular logistic regression models, no such trick is available and estimation of random-effects models has until recently been either unavailable or difficult and time consuming.

Generalized estimating equations (GEE) were introduced by Liang and Zeger (1986) as a means of analysing longitudinal, non-normal data without resorting to fully specified random-effects models. They combine two approaches:

1 **Quasi-likelihood estimation**, where we specify only the mean and variance of the outcome, rather than a full probability model for its distribution. In GEE, the quasi-likelihood approach is generalized to allow a choice of structures for the correlation of outcomes within clusters; this is called a 'working' correlation structure. However, it is important to understand that these correlation structures need not (and often do not) correspond to a correlation structure derived from a full, random effects, probability model for the data.
2 **Robust standard errors** are used to take account of the clustering, and the fact that the parameter estimates are not based on a full probability model.

Note that for normally distributed outcomes, parameter estimates from GEE are identical to those from standard random-effects models.

The most common choice of correlation structure, and the only one that we shall consider here, is the 'exchangeable' correlation structure in which the correl-

ation between a pair of observations in the same cluster is assumed to be the same for all pairs in each cluster.

Example 31.3

The data set we shall use to compare GEE with other approaches to the analysis of clustered data comes from a study of the impact of HIV on the infectiousness of patients with pulmonary TB (Elliott *et al., AIDS* 1993, 7:981–987). This study was based on 70 pulmonary TB patients in Zambia, 42 of whom were infected with HIV and 28 of whom were uninfected. These patients are referred to as *index cases*. The aim of the study was to determine whether HIV-infected index cases were more or less likely than HIV-negative index cases to transmit *M. tuberculosis* infection to their household contacts.

Three hundred and seven household contacts were traced, of whom 181 were contacts of HIV-infected index cases. The mean number of contacts per HIV-infected index case was 4.3 (range 1 to 13), while the mean number of contacts per HIV-uninfected case was 4.5 (range 1 to 11). All these contacts underwent a Mantoux skin test for tuberculosis infection. An induration (skin reaction) of diameter $\geq$ 5 mm was considered to be a positive indication that the contact had tuberculosis infection. Information on a number of household level variables (e.g. HIV status of TB patient, crowding) and on a number of individual contact level variables (e.g. age of contact, degree of intimacy of contact) was recorded. If some index cases are more infectious than others, or household members share previous exposures to TB, then the outcome (result of the Mantoux test in household contacts) will be clustered within households.

Table 31.10 shows that, overall, 184/307 (59.9%) of household contacts had positive Mantoux tests, suggesting that they had tuberculosis infection. This proportion appeared lower among the contacts of HIV-infected index cases (51.9%) than among contacts of HIV-uninfected index cases (71.4%).

Table 31.11(a) shows the results from a standard logistic regression model, ignoring any clustering within households. The odds ratio comparing contacts of HIV-infected index cases with contacts of HIV-uninfected index cases was 0.432 (95% CI 0.266 to 0.701). However, as explained earlier in the chapter, ignoring within-household clustering may mean that this confidence interval is too narrow.

Table 31.10 2×2 table showing the association between Mantoux test status in household contacts of tuberculosis patients and the HIV status of the index case.

	HIV status of index case		
Mantoux test status	Positive	Negative	Total
Positive	94 (51.9%)	90 (71.4%)	184 (59.9%)
Negative	87 (48.1%)	36 (28.6%)	123 (40.1%)
Total	181	126	307

Table 31.11 Regression outputs (odds ratio scale) for the association between Mantoux test positivity in household contacts of tuberculosis patients, and the HIV-infection status of the index case.

(a) Standard logistic regression

	Odds ratio	z	$P > \|z\|$	95% CI
HIV-infected	0.432	−3.40	0.001	0.266 to 0.701

(b) Logistic regression, using robust standard errors to allow for within-household clustering

	Odds ratio	z	$P > \|z\|$	95% CI
HIV-infected	0.432	−2.52	0.012	0.225 to 0.829

(c) Generalized estimating equations (GEE) with robust standard errors to allow for within-household clustering

	Odds ratio	z	$P > \|z\|$	95% CI
HIV-infected	0.380	−2.96	0.003	0.200 to 0.721

We will now compare these results with those in parts (b) and (c) of Table 31.11 from two different methods that allow for clustering. First, part (b) shows the results specifying robust standard errors in the logistic regression model to allow for within-household clustering (see Section 31.3). This approach does not change the estimated odds ratio. However, the 95% confidence interval is now wider, and the *P*-value has increased to 0.012 from 0.001 in the standard logistic regression model.

Part (c) of Table 31.11 shows results from a GEE analysis assuming an 'exchangeable' correlation structure. As well as correcting the standard errors, confidence intervals and *P*-values to account for the clustering, the odds ratio has reduced from 0.43 to 0.38 after taking account of within-household clustering. This is because the GEE analysis gives relatively less weight to contacts in large households. Box 31.1 summarizes theoretical issues in the GEE approach to the analysis of clustered data.

31.6 SUMMARY OF APPROACHES TO THE ANALYSIS OF CLUSTERED DATA

1 If data are clustered, it is *essential* that the clustering should be allowed for in the analyses. In particular, failure to allow for clustering may mean that standard errors of parameter estimates are too small, so that confidence intervals are too narrow and *P*-values are too small.
2 It is always valid to derive summary measures for each cluster, then analyse these using standard methods. However, analyses based on such summary statistics cannot take account of exposure variables that vary between individuals in the same cluster.

BOX 31.1 THEORETICAL ISSUES IN USING GEE

- We do not need to assume that the correlation matrix in GEE is correct; hence it is known as a 'working' correlation matrix. The parameter estimates and standard errors will still be correct ('consistent') provided that the sample size is large enough.
- However, the choice of correlation matrix will affect the parameter estimates. If we assume independence, that is no clustering within groups, then the parameter estimates will be the same as for the corresponding generalized linear model. To derive parameter estimates adjusted as far as possible for the clustering, we need to specify the most realistic correlation matrix possible.
- The GEE approach treats the clustering as a nuisance of no intrinsic interest, but provides parameter estimates and standard errors corrected for the clustering. Unlike random effects models, GEE estimates are not based on a fully specified probability model for the data (except for models with an identity link function: see Section 29.2). GEE models are also known as **'population-averaged'** or **'marginal' models** because the parameter estimates refer to average effects for the population rather than to the effects for a particular individual within the population.
- The GEE approach allows flexibility in modelling correlations, but little flexibility in modelling variances. This can have serious limitations for modelling of grouped counts or proportions, such as in the compound X example above, or in a study of malaria risk if the outcome was the proportion of mosquitoes landing on a bednet that were found to be infective.
- Assumptions about the processes leading to missing data are stronger for GEE than for random-effects models. For example, consider a longitudinal study in which repeated examinations are scheduled every three months, but in which some individuals do not attend some examinations. In GEE, it is assumed that data from these examinations are *missing completely at random*, which means that the probability that an observation is missing is independent of all other observations. For random-effects models the assumption is that data are *missing at random*, which means that the probability that an observation is missing is independent of its true value at that time, but may depend on values at other times, or on the values of other variables in the dataset.

3 The likely effect of the clustering on standard errors may be assessed by specifying robust standard errors that allow for the clustering. Parameter estimates will not be affected. For such robust standard errors to be reliable we need a reasonable number of clusters (at least 30). Wald tests, rather than likelihood ratio tests, must be used.
4 Random-effects (multilevel) models allow for the presence of clustering by modifying the linear predictor by a constant amount u_j in cluster j. The random effects $\{u_j\}$ are assumed to vary randomly between clusters. Random-effects models work well for normally distributed outcomes and Poisson regression, but estimation of random-effects logistic models is difficult and computationally demanding.
5 Generalized estimating equations (GEE) modify both parameter estimates and standard errors to allow for the clustering. Again, there should be a reasonable number of clusters. The GEE approach is particularly useful in logistic regression analyses and when the focus of interest is on the estimated exposure effect and where the clustering is of no intrinsic interest.

In this chapter we have described only the simplest types of model for the analysis of clustered data. In particular the random effects models presented in Section 31.4 include a single random effect to allow for the clustering. Such models have a wealth of possible extensions: for example, we may investigate whether exposure effects, as well as cluster means, vary randomly between clusters. For more details on the analysis of clustered data and random-effects (multilevel) models, see Goldstein (1995), Donner and Klar (2000) or Bryk and Raudenbush (2001).

CHAPTER 32

Systematic reviews and meta-analysis

32.1 INTRODUCTION

There has been an explosion in research evidence in past decades; over half a million articles are published annually in the biomedical literature. It is common for important issues in medical research to be addressed in several studies. Indeed, we might be reluctant to introduce a new treatment based on the result of one trial alone. This chapter focuses on how the evidence relating to a particular research question can be summarized in order to make it accessible to medical practitioners and inform the practice of **evidence-based medicine**. In particular we discuss:

- systematic reviews of the medical literature;
- the statistical methods which are used to combine effect estimates from different studies (meta-analysis);
- sources of bias in meta-analysis and how these may be detected.

Because systematic reviews and meta-analyses of medical research are mainly (though not exclusively) used in combining evidence from randomized trials, we will refer throughout to *treatment* effects, rather than to *exposure* effects.

More detail on all the statistical methods presented in this chapter can be found in *Systematic Reviews in Health Care: Meta-Analysis in Context* edited by Egger, Davey Smith and Altman (2001); see www.systematicreviews.com.

32.2 SYSTEMATIC REVIEWS

The need to summarize evidence systematically was illustrated by Antman *et al.* (1992), who compared accumulating data from randomized controlled trials of treatments for myocardial infarction (heart attack) with the recommendations of clinical experts writing review articles and textbook chapters. By the mid-1970s, based on a meta-analysis of around ten trials in more than 2500 patients, there was good evidence of a protective effect of thrombolytic therapy after myocardial infarction against subsequent mortality. However, trials continued to be performed for the next 15 years (the cumulative total patients had reached more than 48 000 by 1990). It was not until the late 1980s that the majority of textbooks and review articles recommended the routine use of thrombolytic therapy after myocardial infarction.

It is now recognized that a conventional 'narrative' literature review – a 'summary of the information available to the author from the point of view of the author' – can be very misleading as a basis from which to draw conclusions on the overall evidence on a particular subject. Reliable reviews must be *systematic* if bias in the interpretation of findings is to be avoided.

Cook *et al.* (1995) defined a **systematic review** of the literature as 'the application of scientific strategies that limit bias by the systematic assembly, critical appraisal and synthesis of all relevant studies on a specific topic'. The main feature which distinguishes systematic from narrative reviews is that they have a methods section which clearly states the question being addressed, the subgroups of interest and the *methods and criteria employed for identifying and selecting relevant studies and extracting and analysing information*. Systematic reviews are a substantial undertaking and a team with expertise in both the content area and review methodology is usually needed.

Guidelines on the conduct of systematic reviews may be found in Egger, Davey Smith and Altman (2001) and in the Cochrane Collaboration handbook. The QUOROM statement (Moher *et al.* 1999) suggests guidelines for the reporting of systematic reviews.

32.3 THE COCHRANE AND CAMPBELL COLLABORATIONS

We have seen that:

- medical practice needs to be based on the results of systematic reviews, rather than (non-systematic) 'expert reviews' of the literature;
- to perform a systematic review is a substantial undertaking

The *Cochrane Collaboration* (www.cochrane.org), which started in 1993, is an attempt to address these issues. It aims to produce systematic, periodically updated reviews of medical and public health interventions. Cochrane reviews are available in electronic form (via CD-ROM and on the internet), which means that reviews can be updated as new evidence becomes available or if mistakes have been identified. Already, more than 1000 systematic reviews are available as

part of the Cochrane Collaboration, and some 150 000 studies are indexed in the database of randomized controlled trials.

The *Campbell Collaboration* (www.campbellcollaboration.org) is a similar initiative for systematic reviews of social and educational policies and practice, some of which include an impact on health-related outcomes.

32.4 META-ANALYSIS

The statistical methods for combining the results of a number of studies are known as **meta-analysis**. It should be emphasized that not all systematic reviews will contain a meta-analysis; this will depend on whether the systematic review has located studies that are sufficiently similar to make it reasonable to consider combining their results. The increase in interest in meta-analysis is illustrated by the fact that while in 1987 there were five MEDLINE citations using the term META-ANALYSIS, this had increased to 380 by 1991, and 580 by 2001.

We will illustrate methods for meta-analysis using studies with a binary outcome and measuring treatment effects using odds ratios. Corresponding methods exist for other treatment effect estimates such as risk ratios or risk differences, and for continuous outcome measures.

Example 32.1 Effect of diuretics on pre-eclampsia in pregnancy

In an early meta-analysis, Collins *et al*. (1985) examined the results of randomized controlled trials of diuretics in pregnancy. After excluding trials in which they considered that there was a possibility of severe bias, they found nine trials in which the effect of diuretics on pre-eclampsia (a rapid increase in blood pressure or proteinuria which may have severe sequelae) was reported. Table 32.1 summarizes the results of these trials.

Table 32.1 Results of nine randomized controlled trials of diuretics in pregnancy.

	Pre-eclampsia/total		
First author	Treated patients	Control patients	Odds ratio (95% CI)
Weseley	14/131	14/136	1.043 (0.477, 2.28)
Flowers	21/385	17/134	0.397 (0.203, 0.778)
Menzies	14/57	24/48	0.326 (0.142, 0.744)
Fallis	6/38	18/40	0.229 (0.078, 0.669)
Cuadros	12/1011	35/760	0.249 (0.128, 0.483)
Landesman	138/1370	175/1336	0.743 (0.586, 0.942)
Kraus	15/506	20/524	0.770 (0.390, 1.52)
Tervila	6/108	2/103	2.971 (0.586, 15.1)
Campbell	65/153	40/102	1.145 (0.687, 1.91)

In order to make an overall assessment of the effect of diuretics on pre-eclampsia, we would like to combine the results from these nine studies into a single summary estimate of the effect, together with a confidence interval. In doing this:

- Treated individuals should only be compared with control individuals from the same study, since the characteristics of patients in the different studies may differ in important respects, for example, because of different entry criteria, or because they come from different study populations which may have different underlying risks of pre-eclampsia. Thus simply combining patients across the studies would not be an appropriate way to estimate the overall treatment effect.
- Note that even if all the studies are broadly comparable, sampling error will inevitably mean that the observed treatment effects will vary. In this example the estimated odds ratios vary from 0.229 (Fallis) to 2.971 (Tervila).
- The relative sizes of the studies should be taken into account. Note that the most extreme results (odds ratios furthest away from 1) come from the smaller studies.

In the next two sections we describe fixed-effect and random-effects approaches to meta-analysis. A **fixed-effect meta-analysis** can be conducted if it is reasonable to assume that the underlying treatment effect is the same in all the studies, and that the observed variation is due entirely to sampling variation. The fixed-effect assumption can be examined using a **test of heterogeneity** between studies, as described at the end of Section 32.5. A **random-effects meta-analysis** aims to allow for such heterogeneity, and is described in Section 32.6.

32.5 FIXED-EFFECT META-ANALYSIS

In a fixed-effect meta-analysis, we assume that the observed variation in treatment effects in the different studies is due entirely to sampling variation, and that the underlying treatment effect is the same in all the study populations. Table 32.2 shows the notation we will use for the results from study i (when we have a binary outcome, as in Example 32.1). The estimate of the odds ratio for the treatment effect in study i is

$$\mathrm{OR}_i = \frac{d_{1i} \times h_{0i}}{d_{0i} \times h_{1i}}$$

In Example 32.1, we have nine such tables of the effects of treatment with diuretics on pre-eclampsia, one from each of the nine trials, and nine odds ratios. The

Table 32.2 Notation for the 2 × 2 table of results from study i.

	Outcome		
	Experienced event: D (Disease)	Did not experience event: H (Healthy)	Total
Group 1 (intervention)	d_{1i}	h_{1i}	n_{1i}
Group 0 (control)	d_{0i}	h_{0i}	n_{0i}
Total	d_i	h_i	n_i

summary estimate of the treatment effect is calculated as a **weighted average** (see Section 18.3) of the log odds ratios from the separate trials:

$$\log(\text{OR}_F) = \frac{\Sigma[w_i \times \log(\text{OR}_i)]}{\Sigma w_i}$$

The subscript F denotes the assumption that the effect of diuretics is the same, or *fixed*, in each study. Note that individuals are only compared with other individuals in the same study (via the study log odds ratio).

In the **inverse variance method**, the weight w_i for study i equals the inverse of the variance, v_i, of the estimated log odds ratio in that study (see Section 16.7):

$$\text{Inverse variance weights: } w_i = 1/v_i,$$
$$\text{where } v_i = 1/d_{1i} + 1/h_{1i} + 1/d_{0i} + 1/h_{0i}$$

This choice of weights minimizes the standard error of the summary log odds ratio, which is:

$$\text{s.e.}\,(\log(\text{OR}_F)) = \sqrt{\frac{1}{\Sigma w_i}}$$

This can be used to calculate confidence intervals, a z statistic and hence a P-value for the summary log odds ratio. An alternative weighting scheme is to use **Mantel–Haenszel weights** to combine the odds ratios from the individual studies. These are:

$$\text{Mantel–Haenszel weights: } w_i = d_{0i}h_{1i}/n_i$$

Example 32.1 (continued)

Results from a fixed-effect meta-analysis of the data on the effect of diuretics in pregnancy are shown in Table 32.3. This gives clear evidence that the odds of pre-eclampsia were reduced in mothers treated with diuretics. As usual, the estimated summary log odds ratio and its confidence interval have been converted to an odds ratio, for ease of interpretation.

Table 32.3 Results of a fixed-effect meta-analysis of results from nine randomized controlled trials of diuretics in pregnancy.

OR_F	z	P-value	95% CI
0.672	−4.455	< 0.001	0.564 to 0.800

Note on sparse data

If any of the cells in the 2×2 table for one (or more) of the contributing studies contains zero, then the formulae for the log OR_i and corresponding variance, v_i, in that table break down. When this happens, it is conventional to add 0.5 to all cells in the table, and it may be preferable to use Mantel–Haenszel weights. In other circumstances the inverse-variance and Mantel–Haenszel methods will give similar results.

Forest plots

Results of meta-analyses are displayed in a standard way known as a 'forest plot', and such a plot of the diuretics data is shown in Figure 32.1. The horizontal lines correspond to the 95% confidence intervals for each study, with the corresponding box area drawn proportional to the weight for that individual study in the meta-analysis. Hence the wider is the confidence interval the smaller is the box area. The

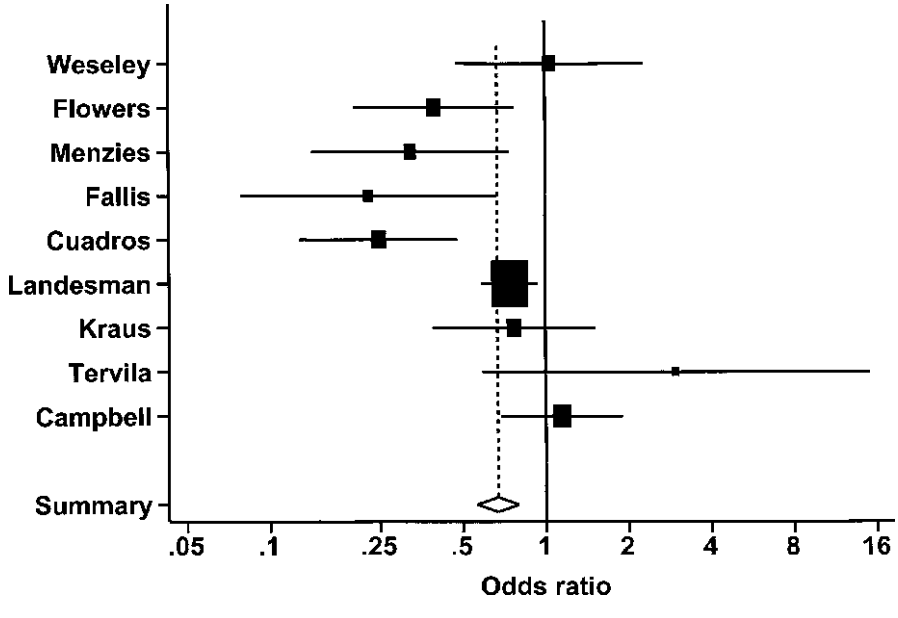

Fig. 32.1 Forest plot of the results of a fixed-effect meta-analysis of nine studies of the effect of diuretics in pregnancy.

diamond (and broken vertical line) represents the summary estimate, and the confidence interval for the summary estimate corresponds to the width of the diamond. The unbroken vertical line is at the null value (1) of the odds ratio, and is equivalent to no treatment effect. Note that the horizontal axis is plotted on a log scale, so that confidence intervals are symmetrical and an odds ratio of (e.g.) 2 is the same distance from 1 as $1/2 = 0.5$.

The exact origin of the name 'forest plot' is not clear. One possible derivation is that it allows one to avoid the pitfall of 'not being able to see the wood for the trees'.

Testing for heterogeneity between studies

The fixed-effect estimate is based on the assumption that the true effect does not differ between studies. This assumption should be checked. We can do this using a **χ^2 test of heterogeneity**, similar to that described for Mantel–Haenszel methods in Section 18.5. The greater the average distance between the log odds ratios estimated in the individual studies and the summary log odds ratio, the more evidence against the null hypothesis that the true log odds ratios are the same. The χ^2 test of heterogeneity (often denoted by Q) is based on a weighted sum of the squares of these differences:

$$\chi^2 = Q = \Sigma w_i \, [\log(\mathrm{OR}_i) - \log(\mathrm{OR}_F)]^2$$

$$\text{d.f.} = \text{number of studies} - 1$$

Example 32.1 (continued)
For the data on the effect of diuretics in pregnancy,

$$\chi^2 = 27.265, \ \text{d.f.} = 9{-}1 = 8, \ P = 0.001$$

There is therefore strong evidence (confirming the impression in the graph) that the effect of diuretics differs between studies.

32.6 RANDOM-EFFECTS META-ANALYSIS

If there is evidence of heterogeneity between studies, how should we proceed? Although it can be argued that it is inappropriate to calculate a summary measure (this is discussed further below), it is also possible to allow for the heterogeneity by incorporating a model for the heterogeneity between studies into the meta-analysis. This approach is called random-effects meta-analysis.

In random-effects meta-analysis, we assume that the 'true' log odds ratio in each study comes from a normal distribution:

$$\log(\text{OR}_i) \approx N(\log(\text{OR}_R),\ \tau^2)$$

whose mean equals the true 'overall' treatment effect and whose variance is usually denoted by τ^2 (τ is the Greek letter tau). We estimate this between-study variance, τ^2, from the observed data (see below) and use this to modify the weights used to calculate the **random-effects summary estimate**:

$$\log(\text{OR}_R) = \frac{\Sigma[w_i^* \times \log(\text{OR}_i)]}{\Sigma w_i^*}$$

$$w_i^* = \frac{1}{v_i + \tau^2},\ \text{where}\ v_i = 1/d_{1i} + 1/h_{1i} + 1/d_{0i} + 1/h_{0i}$$

The standard error of the random-effects summary estimate is calculated from the inverse of the sum of the adjusted weights:

$$\text{s.e.}\ (\log(\text{OR}_R)) = \sqrt{\frac{1}{\Sigma w_i^*}}$$

Estimating the between-study variance

The most commonly used formula for estimating the between-study variance, τ^2, from the observed data was put forward by DerSimonian and Laird (1986). It is based on the value of the χ^2 test of heterogeneity, represented by Q, the unadjusted weights, w_i, and the number of contributing studies, k:

$$\tau^2 = \max\left[0,\ \left(\frac{Q - (k-1)}{W}\right)\right],$$

$$\text{where}\ Q = \chi^2 = \Sigma w_i\ (\log(\text{OR}_i) - \log(\text{OR}_F))^2$$

$$\text{and}\ W = \Sigma w_i - \left(\frac{\Sigma w_i^2}{\Sigma w_i}\right)$$

The mathematical details are included here for completeness. In practice the computer would calculate this as part of the random-effects meta-analysis routine.

Table 32.4 Comparison of fixed-effects and random-effects meta-analysis results of nine randomized controlled trials of the impact of diuretics in pregnancy on pre-eclampsia.

Method	Summary OR	95% CI	z	P-value
Fixed-effects	0.672	0.564 to 0.800	−4.455	<0.001
Random-effects	0.596	0.400 to 0.889	−2.537	0.011

Example 32.1 (continued)

For the data on the effect of diuretics in pregnancy, the estimate of the between-study variance is $\tau^2 = 0.230$, and the summary OR is $\text{OR}_R = 0.596$, somewhat smaller than the fixed-effect estimate. The confidence interval is correspondingly much wider, as can be seen in Table 32.4, which presents the results from both the fixed-effect and random-effects meta-analyses.

Comparison of fixed-effect and random-effects meta-analysis

Because of the addition of τ^2 (the estimated between-study variance) to their denominators, random-effects weights are:

1 smaller, and

2 much more similar to each other

than their fixed-effect counterparts. Table 32.5 illustrates this for the diuretics trials of Example 32.1. This results in:

3 smaller studies being given greater relative weight,

4 a wider confidence interval for the summary estimate, and

5 a larger P-value

compared to the corresponding fixed-effect meta-analysis (see Table 32.4). Thus a random-effects meta-analysis will in general be *more conservative* than its fixed-effect counterpart. This reflects the greater uncertainty inherent in the random-effects approach, because it is assumed that, in addition to sampling variation, the true effect varies between studies.

Table 32.5 Comparison of the weights used in the fixed-effect and random-effects meta-analyses of the diuretics trial data, shown in Table 32.1.

Study	Odds ratio (95% CI)	Fixed-effects weight	Random-effects weight
Weseley	1.04 (0.48 to 2.28)	6.27	2.57
Flowers	0.40 (0.20 to 0.78)	8.49	2.88
Menzies	0.33 (0.14 to 0.74)	5.62	2.45
Fallis	0.23 (0.08 to 0.67)	3.35	1.89
Cuadros	0.25 (0.13 to 0.48)	8.75	2.91
Landesman	0.74 (0.59 to 0.94)	68.34	4.09
Kraus	0.77 (0.39 to 1.52)	8.29	2.85
Tervila	2.97 (0.59 to 15.1)	1.46	1.09
Campbell	1.14 (0.69 to 1.91)	14.73	3.36

Note that the greater the estimate of τ^2, the greater the difference between the fixed-effect and random-effects weights. If τ^2 (the between-study variance) is estimated to be zero, then the fixed-effect and random-effects estimates will be identical.

Interpretation of the summary estimate from a random-effects meta-analysis

The *interpretation* of the random-effects summary estimate is in fact very different to that of the fixed-effect one. In fixed-effect meta-analysis it is *assumed* that the true effect is the same in each study and that the only reason for variation in the estimates between studies is sampling error. In other words, it is assumed that the treatment effect is universal, and the meta-analysis provides the best available estimate of it.

In random-effects meta-analysis, the estimate is of a *mean* effect about which it is assumed that the true study effects vary. There is disagreement over whether it is appropriate to use random-effects models to combine study estimates in the presence of heterogeneity, and whether the resulting summary estimate is meaningful. This will be illustrated in Example 32.2.

Example 32.2 BCG vaccination

It has been recognized for many years that the protection given by BCG vaccination against tuberculosis varies between settings. For example, the risk ratio comparing vaccinated with unvaccinated individuals in the MRC trial in the UK (conducted during the 1960s and 1970s) was 0.24 (95% CI 0.18 to 0.31), while in the very large trial in Madras, south India, there appeared to be no protection (risk ratio 1.01, 95% CI 0.89 to 1.14).

In a meta-analysis published in 1994, Colditz *et al*. used all trials in which random or systematic allocation was used to decide vaccine or placebo, and in which both groups had equivalent surveillance procedures and similar lengths of follow-up. Using a random-effects meta-analysis (having noted the highly significant heterogeneity between trials) they concluded that the risk ratio was 0.49 (95% CI 0.34 to 0.70).

While Colditz *et al*. concluded that 'the results of this meta-analysis lend added weight and confidence to arguments favouring the use of BCG vaccine', Fine (1995) reached different conclusions. Noting, like Colditz *et al*., the strong association between latitude and estimated effect of the vaccine (BCG appeared to work better further away from the equator) he commented that 'it is invalid to combine existing data into a single overall estimate' and further that 'most of the studies of BCG have been at relatively high latitudes whereas their current use is mainly at lower latitudes'. Thus it can be argued that random-effects meta-analysis is simply a means of combining 'apples and pears': forming an average of estimates of quantities whose values we know to be different from each other.

We also saw earlier that in a random-effects meta-analysis studies are weighted more equally than in a fixed-effect meta-analysis. If a random-effects summary

estimate differs from the fixed-effect estimate, this is a sign that the average estimate from the smaller studies differs from the average of the large ones. Given that small studies are more subject to publication bias than large ones (see Section 32.7), this is clearly a disadvantage of random-effects meta-analyses. While *explanations* for heterogeneity may provide useful insights into differences between studies, and may have implications for clinical practice, we should be very cautious about an approach that *adjusts* for heterogeneity without *explaining* it.

Meta-regression

While there is disagreement over whether it is appropriate to use random-effects models to combine study estimates in the presence of heterogeneity, it is clear that the investigation of **sources of heterogeneity** (such as study latitude in the example above) may yield important insights. In the case of BCG vaccination, Fine discusses how the association with latitude may be because of differential exposure to environmental mycobacteria in different populations, which may in turn yield insights into mechanisms of immunity to mycobacterial diseases.

Meta-regression can be used to examine associations between study characteristics and treatment effects. In this approach, we postulate that the treatment effect (e.g. log odds ratio) is related in a linear manner to one or more study covariates.

Then, as with random-effects meta-analysis, we incorporate an additional variance component τ^2 that accounts for unexplained heterogeneity between studies. The meta-regression procedure iterates between (i) estimating τ^2, and (ii) using this estimate in a weighted regression to estimate the covariate effects. The estimated covariate effects lead to a new estimate of τ^2, and so on. The process stops when consecutive steps in the iteration yield almost identical values for τ^2 and for the covariate effects; the model is then said to have converged.

32.7 BIAS IN META-ANALYSIS

The emphasis on the importance of sound methodology for systematic reviews arises from the observation that severe bias may result if this methodology is not applied. Summarizing the results of five biased trials will give a precise but biased result!

Causes of bias: poor trial quality

Empirical evidence that methodological quality of studies was associated with estimates of treatment effect in clinical trials was first provided in an important study by Schulz *et al.* (1995), who assessed the methodological quality of 250 controlled trials from 33 meta-analyses of treatments in the area of pregnancy and childbirth. They found that trials in which treatment allocation was inadequately concealed (see Chapter 34) had odds ratios which were exaggerated (i.e. further

away from 1) by 41% compared to trials which reported 'adequate concealment'. Trials that were not double-blind yielded 17% larger estimates of effect.

An important consequence of the recognition that the quality of a trial may affect its results was to encourage improved standards of conduct and reporting of randomized trials. In particular the CONSORT statement (see Moher, Schulz and Altman (2001), www.consort-statement.org and Chapter 34), which was published in 1996 and updated in 2001, aims to standardize the reporting of trials in medical journals.

Causes of bias: publication bias

In general, a study showing a beneficial effect of a new treatment is more likely to be considered worthy of publication than one showing no effect. There is a considerable bias that operates at every stage of the process, with negative trials considered to contribute less to scientific knowledge than positive ones:

- those who conducted the study are more likely to submit the results to a peer-reviewed journal;
- editors of journals are more likely to consider the study potentially worth publishing and send it for peer review;
- referees are more likely to deem the study suitable for publication.

This situation has been accentuated by two factors: first that studies have often been too small to detect a beneficial effect even if one exists (see Chapter 35) and second that there has been too much emphasis on 'significant' results (i.e. $P < 0.05$ for the effect of interest).

A proposed solution to the problem of publication bias is to establish registers of all trials in a particular area, from when they are funded or established. It has also been proposed that journals consider studies for publication 'blind' of the actual results (i.e. based only on the literature review and methods). It is also clear that the active discouragement of studies that do not have power to detect a clinically important effect would alleviate the problem. Publication bias is a lesser problem for larger studies, for which there tends to be general agreement that the results are of interest, whatever they are.

Funnel plots to examine bias in meta-analysis

The existence of publication bias may be examined graphically by the use of 'funnel plots'. These are simple scatter plots of the treatment effects estimated from individual studies on the horizontal axis and the standard error of the treatment effect (reflecting the study size) on the vertical axis. The name 'funnel plot' is based on the fact that the precision in the estimation of the underlying treatment effect will increase as the sample size of component studies increases. Effect estimates from small studies will therefore scatter more widely at the bottom of the graph, with the spread narrowing among larger studies. In the absence of bias the plot will resemble a symmetrical inverted funnel, as shown in panel (a) of Figure 32.2.

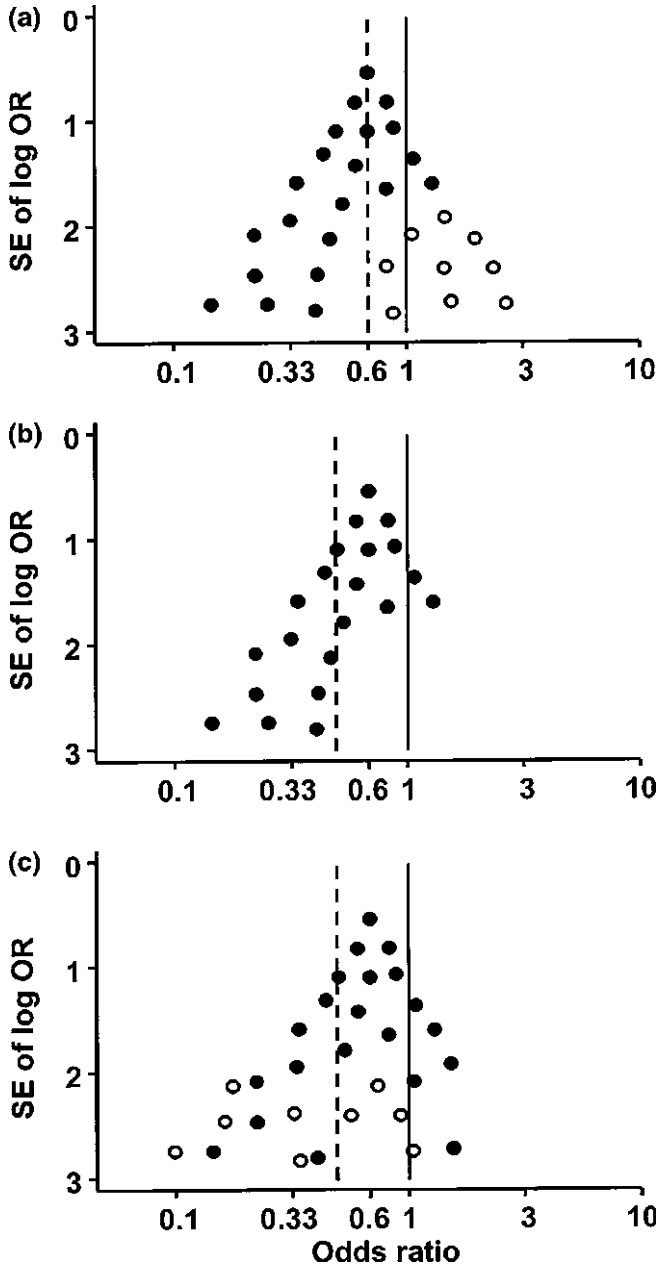

Fig. 32.2 Hypothetical funnel plots: (a) symmetrical plot in the absence of bias (open circles indicate smaller studies showing no beneficial effects); (b) asymmetrical plot in the presence of publication bias (smaller studies showing no beneficial effects are missing); (c) asymmetrical plot in the presence of bias due to low methodological quality of smaller studies (open circles indicate small studies of inadequate quality whose results are biased towards larger beneficial effects).

Relative measures of treatment effect (risk ratios or odds ratios) are plotted on a logarithmic scale. This is important to ensure that effects of the same magnitude but opposite directions, for example risk ratios of 0.5 and 2, are equidistant from 1 (corresponding to no effect). Treatment effects have generally been plotted against sample sizes. However, the statistical power of a trial is determined both by the total sample size and the number of participants developing the event of interest. For example, a study with 100 000 patients and 10 events is less likely to show a statistically significant effect of a treatment than a study with 1000 patients and 100 events. The standard error of the effect estimate, rather than total sample size, has therefore been increasingly used in funnel plots (Sterne and Egger 2001).

If there is bias, for example because smaller studies showing no statistically significant effects (open circles in the figure) remain unpublished, then such publication bias will lead to an asymmetrical appearance of the funnel plot with a gap in the right bottom side of the graph (panel (b) of Fig. 32.2). In this situation the combined effect from meta-analysis will overestimate the treatment's effect. The more pronounced the asymmetry, the more likely it is that the amount of bias will be substantial.

What factors can lead to asymmetry in funnel plots?

Publication bias has long been associated with funnel plot asymmetry, but it is important to realise that publication bias is not the only cause of funnel plot asymmetry. We have already seen that trials of lower quality may yield exaggerated estimates of treatment effects. Smaller studies are, on average, conducted and analysed with less methodological rigour than larger studies, so that asymmetry may also result from the over-estimation of treatment effects in smaller studies of lower methodological quality (panel (c) of Fig. 32.2).

Funnel plot asymmetry may have causes other than bias. Heterogeneity between trials can also lead to funnel plot asymmetry if the true treatment effect is larger (or smaller) in the smaller trials because these are conducted, for example, among high-risk patients. Such trials will tend to be smaller, because of the difficulty in recruiting such patients and because increased event rates mean that smaller sample sizes are required to detect a given effect. In addition, in some large trials, interventions may be implemented under routine conditions rather than in trial conditions where it is possible to invest heavily in assuring all aspects are perfect. This will result in relatively lower treatment effects. For example, an asymmetrical funnel plot was found in a meta-analysis of trials examining the effect of geriatric assessment programmes on mortality. An experienced consultant geriatrician was more likely to be actively involved in the smaller trials and this may explain the larger treatment effects observed in these trials.

Because publication bias is only one of the possible reasons for asymmetry, the funnel plot should be seen more as a means of examining '**small study effects**' (the tendency for the smaller studies in a meta-analysis to show larger treatment

effects). The presence of funnel plot asymmetry should lead to consideration of possible explanations, and may bring into question the interpretation of the overall estimate of treatment effect from a meta-analysis.

Statistical tests for funnel plot asymmetry

Symmetry or asymmetry is generally defined informally, through visual examination, but different observers may interpret funnel plots differently. More formal statistical methods to examine associations between the studies' effects and their sizes have been proposed. Begg and Mazumdar (1994) proposed an adjusted rank correlation test for publication bias which involves calculation of the rank correlation between the treatment effect and its estimated standard error (or, equivalently, variance) in each study. Egger *et al.* (1997a) proposed a linear regression test in which the standardized treatment effect from each study, that is the treatment effect divided by its standard error, is regressed against the precision of the treatment effect. For binary outcomes, the regression equation is:

$$y_i = \beta_0 + \beta_1 x_i, \text{ where}$$

$$y_i = \log(\text{OR}_i)/\text{s.e. } [\log(\text{OR}_i)] = \log(\text{OR}_i) \times \sqrt{w_i}$$

$$x_i = 1/\text{s.e. } [\log(\text{OR}_i)] = \sqrt{w_i}$$

and evidence for bias is found if the intercept β_0 differs from zero.

This test is equivalent to a regression of the log odds ratio against standard error (Sterne *et al.* 2000). This can be seen by multiplying the regression equation above by s.e. $[\log(\text{OR}_i)]$, which gives:

$$\log(\text{OR}_i) = \beta_0 \times \text{s.e. } [\log(\text{OR}_i)] + \beta_1$$

where the regression accounts for between-subject heterogeneity by weighting according to the inverse of the variance of $\log(\text{OR}_i)$. The greater the association between $\log(\text{OR}_i)$ and s.e. $[\log(\text{OR}_i)]$, measured by the size of the regression coefficient β_0, the greater the evidence for funnel plot asymmetry. The test is therefore very closely related to a **meta-regression** of $\log(\text{OR}_i)$ on s.e. $[\log(\text{OR}_i)]$. There is thus the potential to include s.e.$[\log(\text{OR}_i]$ together with other study characteristics (for example measures of study quality) in a multiple meta-regression to examine competing explanations for differences between studies.

The power and sensitivity of these tests is not well established. It appears that the regression method is more powerful than the rank correlation method, but

that power is low unless the amount of bias is substantial and the number of studies in the meta-analysis exceeds ten (Sterne *et al*. 2000).

32.8 META-ANALYSIS OF OBSERVATIONAL STUDIES

Although the emphasis in this chapter has been on the meta-analysis of data from randomized trials, there are many questions which can only be addressed in observational studies. These include:

- studies of the aetiology of disease (for example, does passive smoking cause lung cancer?);
- evaluations of the effectiveness of interventions that have already been introduced, such as BCG vaccination;
- evaluation of the effectiveness of an intervention on rare adverse outcomes, such as mortality, for which the sample size required for randomized controlled trials might be prohibitive;
- evaluation of the effectiveness of interventions that need to be applied on a widespread basis, such as a mass media campaign, and for which therefore it is not possible to have control groups;
- evaluation of the effectiveness of interventions in populations other than those in which they were first evaluated.

For this reason a substantial proportion of published meta-analyses are based on observational studies rather than on randomized trials.

However, the issues involved in meta-analysis of observational studies are very different, and more difficult, than for the meta-analysis of randomized trials. In particular, the appropriate control of confounding factors is of fundamental importance in the analysis and interpretation of observational studies while, in contrast, appropriate randomization should mean that confounding is not a problem in trials (providing that their size is large enough, see Chapters 34 and 35). Other types of bias, for example recall bias, may also be of greater concern in observational studies than in randomized trials.

A striking example of the potential for meta-analyses of observational studies to give misleading results was given by Egger *et al*. (1997b). They compared the results of six observational cohort studies of the association between intake of beta-carotene (a precursor of the antioxidant vitamin A) and cardiovascular mortality, with those from four randomized trials in which participants randomized to beta-carotene supplements were compared with participants randomized to placebo. As can be seen from Figure 32.3, the cohort studies indicated a strong protective effect of beta-carotene while the randomized trials suggest a moderate adverse effect of beta-carotene supplementation. An individual's diet is strongly associated with other characteristics associated with cardiovascular mortality (for example physical activity and social class) and these results suggest that failure to control for such factors, or other types of bias, led to the apparent protective effect of beta-carotene in the observational studies.

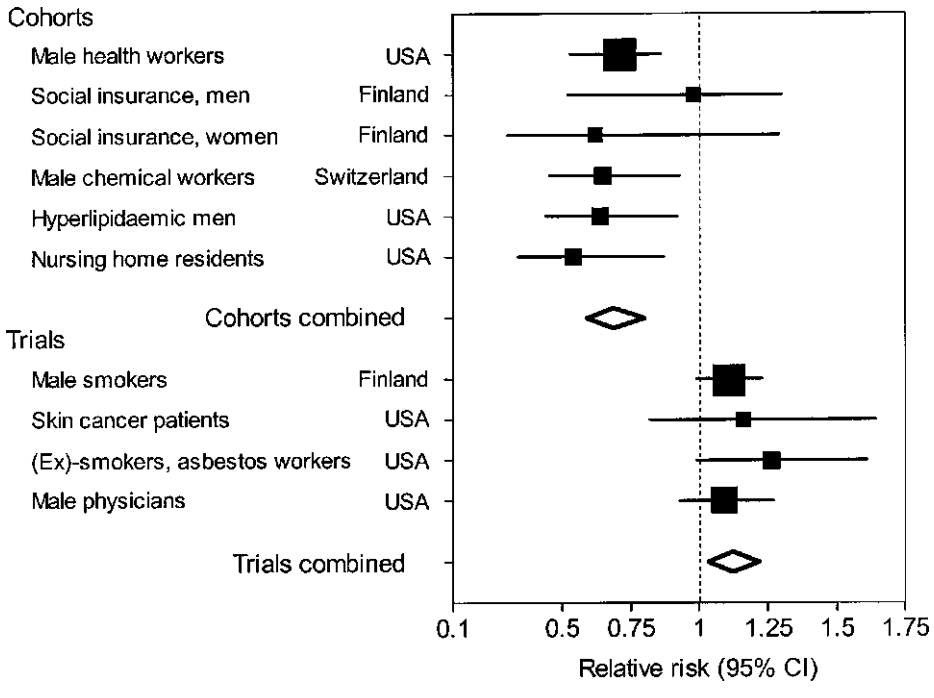

Fig. 32.3 Meta-analysis of the association between beta-carotene intake and cardiovascular mortality. Results from observational studies indicate considerable benefit whereas the findings from randomized controlled trials show an increase in the risk of death. We are grateful to Matthias Egger for permission to reproduce the figure.

This suggests that the statistical combination of studies should not, in general, be a prominent component of systematic reviews of observational studies, which should focus instead on possible sources of heterogeneity between studies and the reasons for these.

32.9 CONCLUSIONS

Systematic reviews and meta-analysis (the quantitative analysis of such reviews) are now accepted as an important part of medical research. While the analytical methods are relatively simple, there is still controversy over appropriate methods of analysis. Systematic reviews are substantial undertakings, and those conducting such reviews need to be aware of the potential biases which may affect their conclusions. However, the explosion in medical research information and the availability of reviews on-line mean that synthesis of research findings is likely to be of ever increasing importance to the practice of medicine.

CHAPTER 33

Bayesian statistics

33.1 INTRODUCTION: BAYESIAN INFERENCE

In this chapter we give a brief description of the Bayesian approach to statistical inference, and compare it to the frequentist approach which has been used in the rest of the book. The Bayesian approach is based on **Bayes' formula** for relating **conditional probabilities** (see Chapter 14):

$$\text{prob}\,(\text{B given A}) = \frac{\text{prob}\,(\text{A given B}) \times \text{prob}\,(\text{B})}{\text{prob}\,(\text{A})}$$

We have seen that a statistical model specifies how the probability distribution of an outcome variable (the data) depends on model parameters. For example, consider a trial of the effect of thrombolysis on the risk of death up to 1 year after a myocardial infarction. The data are the number of patients and number of deaths in each group, and the model parameters are the risk of death in the control group, and the risk ratio comparing the risk of death in patients given thrombolysis with the risk of death in the control group. In Chapter 28 we explained that the model parameters are fitted using the maximum likelihood approach. This is based on calculating the *conditional* probability of the observed data given model parameters.

The **Bayesian approach** to statistical inference starts with a *prior belief* about the likely values of the model parameters, and then uses the observed data to modify these. We will denote this prior belief by prob (parameters). Bayes' formula provides the mechanism to update this belief in the light of the data:

$$\text{prob}\,(\text{model parameters given data}) = \frac{\text{prob}\,(\text{data given model parameters}) \times \text{prob}\ (\text{parameters})}{\text{prob}\,(\text{data})}$$

The prior belief concerning the values of the parameters is often expressed in terms of a probability distribution, such as a normal or binomial distribution, represent-

ing a range of possible values, rather than as single values. This is called the **prior distribution**. The probability distribution of the model parameters given the data is known as the **posterior distribution**.

33.2 COMPARISON OF BAYESIAN AND FREQUENTIST STATISTICAL INFERENCE

In this book we have concentrated on the **frequentist** approach to statistical inference, in which we think of probability in terms of the proportion of times that an event would occur in a large number of similar repeated trials. In frequentist statistical inference, we think of model parameters (for instance the risk ratio for the effect of thrombolysis on the risk of death following heart attack, compared to placebo) as fixed. We use the data to make inferences about model parameters, via parameter estimates, confidence intervals and *P*-values.

In the Bayesian approach our inferences are based on the posterior probability distribution for the model parameters. For example, we might derive a **95 % credible interval**, based on the posterior distribution, within which there is 95 % probability that the parameter lies. Box 33.1 compares the Bayesian and frequentist approaches

BOX 33.1 COMPARISON OF FREQUENTIST AND BAYESIAN APPROACHES TO STATISTICAL INFERENCE

Frequentist statistics	Bayesian statistics
We use the data to make inferences about the true (but unknown) population value of the risk ratio.	We start with our *prior* opinion about the risk ratio, expressed as a probability distribution. We use the data to modify that opinion (we derive the *posterior* probability distribution for the risk ratio based on *both* the data and the prior distribution).
The *95 % confidence interval* gives us a range of values for the population risk ratio that is consistent with the data. 95 % of the times we derive such a range it will contain the true (but unknown) population value.	A *95 % credible interval* is one that has a 95 % chance of containing the population risk ratio.
The *P*-value is the probability of getting a risk ratio at least as far from the null value of 1 as the one found in our study.	The posterior distribution can be used to derive direct probability statements about the risk ratio, e.g. the probability that the drug *increases* the risk of death.

to statistical inference. See also the book by Royall (1997), which describes and compares different approaches to statistical inference.

If our prior opinion about the risk ratio is very vague (we consider a very wide range of values to be equally likely) then the results of a frequentist analysis are very similar to the results of a Bayesian analysis—both are based on the likelihood for the data. This is because a vague prior distribution will have little influence on the posterior probability, compared to the influence of the data:

- the 95% confidence interval is the same as the 95% credible interval, except that the latter has the interpretation often incorrectly ascribed to a confidence interval;
- the (1-sided) *P*-value is the same as the probability that the drug increases the risk of death (assuming that we found a protective effect of the drug).

However, the two approaches can give very different results if our prior opinion is not vague relative to the amount of information contained in the data. This issue is at the heart of a long-standing argument between proponents of the two schools of statistical inference. Bayesians may argue that it is appropriate to take external information into account by quantifying this as prior belief. Frequentists, on the other hand, may argue that our inferences should be made based only on the data. Further, prior belief can be difficult to quantify. For example, consider the hypothesis that a particular exposure is associated with the risk of a particular cancer. In quantifying our prior belief, how much weight should be given to evidence that there is a biologically plausible mechanism for the association, compared to evidence that international differences in disease rates show some association with differences in the level of the risk factor?

In some situations, Bayesian inference allows a more natural way to consider consequences of the data than does frequentist reasoning. For example:

- in a clinical trial in which an interim analysis reveals that the estimated risk of disease is identical in the treatment and control groups, Bayesian statistics could be used to ask the question 'What is the probability that there is a clinically important effect of treatment, given the data currently accrued?' This question has no meaning in frequentist statistics, since the effect of treatment is treated as a fixed but unknown quantity;
- in a trial whose aim is to examine whether a new treatment (B) is at least as clinically effective as an existing treatment (A), it is perfectly meaningful, in a Bayesian framework, to ask 'What is the probability that drug B is at least as good as drug A?' In contrast, frequentist statistics tends to focus on testing the evidence against the null hypothesis that the effect of drug B is *the same as* the effect of drug A.

33.3 MARKOV CHAIN MONTE-CARLO (MCMC) METHODS

In recent years there has been a resurgence of interest in Bayesian statistics. This has been based less on arguments about approaches to statistical inference than on a powerful means of estimating parameters in complex statistical models based on

the Bayesian approach. The idea is that if we know the values of all the parameters except for one, then we can derive the *conditional distribution* of the unknown parameter, conditional on the data and the other (known) parameter values. Such a conditional distribution can be derived for each parameter, assuming that the values of all the others are known.

The **Markov Chain Monte-Carlo (MCMC)** procedure is used to generate a value for each parameter, by sampling randomly from its conditional distribution. This then acts as the 'known' value for that parameter. This process is carried out iteratively. A new parameter value is sampled from the distribution of each parameter in turn, and is used to update the 'known' values for the conditional distribution of the next parameter. The phrase 'Markov Chain' refers to the fact that the procedure is based only on the last sampled values of each parameter, while 'Monte-Carlo' refers to the random sampling of the parameter values.

After a suitable 'burn in' period (e.g. 10 000 iterations), the dependence of the procedure on the initial chioce of the parameter values is lost. The parameter values generated over the next (say) 10 000 iterations are then recorded. These correspond to the posterior distribution of the parameters, based on the data and the prior probabilities. The high speeds of modern desktop computers mean that such computationally intensive procedures can be run in reasonable amounts of time, although they are not as quick as standard (maximum-likelihood) methods.

MCMC methods can thus be used as an alternative to maximum-likelihood estimation, for models such as random-effects logistic regression where maximum-likelihood estimation is computationally difficult. This can be carried out using specialised computer software such as BUGS (available at www.mrc-bsu.cam.ac.uk/bugs), which stands for Bayesian inference Using Gibbs Sampling and allows users to specify a wide range of statistical models which are then estimated using MCMC. Note, however, that both model specification and use of the MCMC estimation procedure currently require considerably more technical knowledge than is needed to use a standard statistical software package.

This page intentionally left blank