PART C

ANALYSIS OF BINARY OUTCOMES

In this part of the book we describe methods that are used when the outcome is a **binary variable**; a variable where for each individual in the sample the value is one of two alternatives. For example, at the end of the study a subject may have experienced the particular disease (or event) of interest, or remained healthy. Other examples are that a patient dies or survives, or that a specimen is positive or negative.

Of particular interest is the **proportion** (p) of individuals in our sample who experience the event of interest. We use this sample proportion to estimate the **probability** or **risk** of the event in the population as a whole. For example, we might be interested in:

- the risk of death in the five years following diagnosis of prostate cancer;
- the risk of vertical transmission of HIV during pregnancy or childbirth in HIV-infected mothers given antiretroviral therapy during pregnancy.

Probabilities, risks and the related concept of the **odds** of an event are described in Chapter 14, together with the rules for calculating and manipulating probabilities. This lays the foundations for the rest of this part of the book. In Chapter 15, we derive the sampling distribution of a proportion, which is known as the **binomial distribution**, and show how it can be approximated by the normal distribution to give a confidence interval and *z*-test for a single proportion. In Chapter 16 we describe different ways to compare the occurrence of a binary outcome in two exposure groups; by examining the difference between the proportions, the ratio of the risks, or the ratio of the odds. In Chapter 17, we cover the use of chi-squared tests to examine associations between categorical exposure and outcome variables.

Confounding, which was briefly introduced in Chapter 11, is explained in detail in Chapter 18. It arises when there are differences between the exposure groups, in addition to the exposure itself, which are related to the outcome variable. We show how Mantel–Haenszel methods may be used to control for **confounding** using **stratification**; failure to do this would **bias** the interpretation of the comparison of the exposure groups.

In Chapter 19 we introduce **logistic regression** for the analysis of binary outcome variables, and describe how it can be used to compare two or more exposure groups. We extend this in Chapter 20, by explaining the control of confounding using logistic regression, and briefly describing other regression models for binary and categorical outcome variables. Finally, Chapter 21 introduces the special methods needed for **matched data**, in particular matched case–control studies.

This page intentionally left blank

CHAPTER 14

Probability, risk and odds (of disease)

14.1 INTRODUCTION

Probability has already been used several times in preceding chapters, its meaning being clear from the context. We now need to introduce it more formally and to give rules for manipulating it, before we can introduce methods for the analysis of binary outcome variables. We need to do this for two reasons:

1. There is a close link between the proportion of individuals in the sample who experience the event of interest defined by the binary outcome variable, and the definition of the **probability** or **risk** that an individual in the population as a whole will experience the outcome event (see Section 14.2).
2. We need to be able to carry out calculations involving probabilities in order to be able to derive the **binomial distribution** that describes the sampling distribution of a proportion. This is done in the next chapter.

14.2 DEFINING PROBABILITY

Frequentist definition: probability and risk

Although probability is a concept used in everyday life, and one with which we have an intuitive familiarity, it is difficult to define exactly. The **frequentist definition** is usually used in statistics. This states that the **probability** of the occurrence of a particular event equals the proportion of times that the event would (or does) occur in a large number of similar repeated trials. It has a value between 0 and 1, equalling 0 if the event can never occur and 1 if it is certain to occur. A probability may also be expressed as a percentage, taking a value between 0% and 100%. For example, suppose a coin is tossed thousands of times and in half the tosses it lands head up and in half it lands tail up. The probability of getting a head at any one toss would be defined as one-half, or 50%.

Similarly the probability of death in the five years following diagnosis of prostate cancer would be defined as the proportion of times that this would occur among a large number of men diagnosed with prostate cancer. This probability is then said to be the **risk** of death in the five years following diagnosis of prostate cancer.

Subjective (or Bayesian) definition

An alternative approach is to use a **subjective definition**, where the size of the probability simply represents one's degree of belief in the occurrence of an event, or in an hypothesis. This definition corresponds more closely with everyday usage and is the foundation of the **Bayesian** approach to statistics. In this approach, the investigator assigns a **prior probability** to the event (or hypothesis) under investigation. The study is then carried out, the data collected and the probability modified in the light of the results obtained, using Bayes' rule (see Section 14.4). The revised probability is called the **posterior probability**. The Bayesian approach to statistical inference is described in Chapter 33.

14.3 PROBABILITY CALCULATIONS

There are just two rules underlying the calculation of all probabilities. These are:

1 the **multiplicative rule** for the probability of the occurrence of *both* of two events, A and B, and;

2 the **additive rule** for the occurrence of *at least one of* event A or event B. This is equivalent to the occurrence of *either* event A *or* event B (or both).

We will illustrate these two rules in the context of the following example.

Example 14.1

Consider a couple who plan to have two children. There are four possible combinations for the sexes of these children, as shown in Table 14.1. Each combination is equally likely and so has a probability of 1/4.

Table 14.1 Possible combinations for the sexes of two children, with their probabilities.

	Second child	
First child	Boy 1/2	Girl 1/2
Boy 1/2	1/4 (boy, boy)	1/4 (boy, girl)
Girl 1/2	1/4 (girl, boy)	1/4 (girl, girl)

Multiplicative rule

In fact each of these probabilities of 1/4 derives from the individual probabilities of the sexes of each of the children. Consider in more detail the probability that *both children are girls*. The probability that the first child is a girl is 1/2. There is then a probability of 1/2 of this (i.e. 1/2 of $1/2 = 1/4$) that the second child will also be a girl. Thus:

$$\begin{aligned}\text{Prob (both children are girls)} &= \text{prob (first child is a girl)} \times \\ & \text{prob (second child is a girl)} \\ &= 1/2 \times 1/2 = 1/4\end{aligned}$$

The general rule for the probability of *both* of two events is:

$$\text{Prob (A } \textit{and} \text{ B)} = \text{prob (A)} \times \text{prob (B given that A has occurred)}$$

Prob (B given that A has occurred) is called a **conditional probability**, as it is the probability of the occurrence of event B conditional upon the occurrence of event A. If the likelihood of event B is unaffected by the occurrence or non-occurrence of event A, and *vice versa*, events A and B are said to be **independent** and the rule simplifies to:

$$\text{Prob (A } \textit{and} \text{ B)} = \text{prob (A)} \times \text{prob (B), if A and B are independent}$$

The sexes of children are independent events as the probability that the next child is a girl is uninfluenced by the sexes of the previous children. An example with dependent events is the probability that a young girl in India is both anaemic and malnourished, since she is much more likely to be anaemic if she is malnourished than if she is not. We explore how **Bayes' rule** can help us understand relations between dependent events in Section 14.4.

Additive rule

We now turn to the **additive rule**, which is used for calculating the probability that at least one of event A or event B occurs. This is equivalent to either (i) A alone occurs, or (ii) B alone occurs, or (iii) both A *and* B occur. For example, consider the probability that the couple will have at least one girl if they have two children. We can see from Table 14.1 that this would happen in three of the four possible outcomes; it would not happen if both children were boys. The probability that the couple would have at least one girl is therefore 3/4. Note that it is *not* simply the *sum* of the probability that the first child is a girl *plus* the probability that the second child is a girl. Both these probabilities are 1/2 and would sum to 1 rather than the correct 3/4. This is because the possibility that both children are girls is included in each of the individual probabilities and has therefore been double-counted.

The **additive rule** for the calculation of the probability of occurrence of at least one of two events A and B is therefore:

$$\text{Prob}\,(\text{A } \textit{or} \text{ B } \textit{or} \text{ both}) = \text{prob}\,(\text{A}) + \text{prob}\,(\text{B}) - \text{prob}\,(\text{both})$$

In Example 14.1

$$\begin{aligned}\text{Prob}\,(\text{at least one girl}) &= \text{prob}\,(\text{1st child is girl}) + \text{prob}\,(\text{2nd child is girl})\\ &\quad - \text{prob}\,(\text{both are girls})\\ &= 1/2 + 1/2 - 1/4 = 3/4\end{aligned}$$

From our example, it is also clear that an *alternative* formulation is:

$$\text{Prob}\,(\text{A } \textit{or} \text{ B } \textit{or} \text{ both}) = 1 - \text{prob}\,(\text{A doesn't occur } \textit{and} \text{ B doesn't occur})$$

since

$$\begin{aligned}&\text{Prob}\,(\text{at least one girl}) = 1 - \text{prob}\,(\text{1st is not a girl and 2nd is not a girl})\\ &\text{or equivalently, } 1 - \text{prob}\,(\text{both children are boys}) = 1 - 1/4 = 3/4\end{aligned}$$

14.4 BAYES' RULE

We will now introduce Bayes' rule, which is the basis of the **Bayesian** approach to statistics, introduced in Section 14.2 and described in Chapter 33. We saw above that the general rule for the probability of *both* of two events is

$$\text{Prob}\,(\text{A } \textit{and} \text{ B}) = \text{prob}\,(\text{A}) \times \text{prob}\,(\text{B given A})$$

where we have written the **conditional probability** prob (B given that A has occurred) more concisely as prob (B given A). We now show how this leads to **Bayes' rule** for relating conditional probabilities. Switching A and B in the above formula gives:

$$\text{Prob}\,(\text{B } \textit{and} \text{ A}) = \text{prob}\,(\text{B}) \times \text{prob}\,(\text{A given B})$$

Since the left hand sides of these two equations are exactly the same, that is the probability that both A and B occur, the right hand sides of the two equations must be equal:

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

Rearranging this by dividing both sides of this equation by prob (A) gives **Bayes' rule** for relating conditional probabilities:

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

This allows us to derive the probability of B given that A has happened from the probability of A given that B has happened. The importance of this will become clear in Chapter 33 on the Bayesian approach to statistics. Here, we will just illustrate the calculation with an example.

Example 14.2

Suppose that we know that 10% of young girls in India are malnourished, and that 5% are anaemic, and that we are interested in the relationship between the two. Suppose that we also know that 50% of anaemic girls are also malnourished. This means that the two conditions are not independent, since if they were then only 10% (not 50%) of anaemic girls would also be malnourished, the same proportion as the population as a whole. However, we don't know the relationship the other way round, that is what percentage of malnourished girls are also anaemic. We can use Bayes' rule to deduce this. Writing out the probabilities gives:

$$\begin{aligned}&\text{Probability (malnourished)} = 0.1\\&\text{Probability (anaemic)} = 0.05\\&\text{Probability (malnourished given anaemic)} = 0.5\end{aligned}$$

Using Bayes rule gives:

$$\begin{aligned}&\text{Prob (anaemic given malnourished)}\\&\quad = \frac{\text{prob (anaemic)} \times \text{prob (manourished given anaemic)}}{\text{prob (malnourished)}}\\&\quad = \frac{0.05 \times 0.5}{0.1} = 0.25\end{aligned}$$

We can thus conclude that 25%, or one quarter, of malnourished girls are also anaemic.

14.5 THE INDEPENDENCE ASSUMPTION

Standard statistical methods assume that the outcome for each individual is **independent** of the outcome for other individuals. In other words, it is assumed that the probability that the outcome occurs for a particular individual in the sample is unrelated to whether or not it has occurred for the other individuals. An example where this assumption is violated is when different individuals in the same family (for example siblings) are sampled, because the outcome for an individual is on average more similar to that for their sibling than to the rest of the population. The data are then **clustered**, and special methods that allow for the clustering must be used. These are described in Chapter 31.

14.6 PROBABILITIES AND ODDS

In this section, we introduce the concept of odds and examine how they relate to probability. The **odds** of an event are commonly used in betting circles. For example, a bookmaker may offer odds of 10 to 1 that Arsenal Football Club will be champions of the Premiership this season. This means that the bookmaker considers the probability that Arsenal will *not* be champions is 10 times the probability that they will be. Most people have a better intuitive understanding of probability than odds, the only common use of odds being in gambling (see below). However, as we will see in Chapters 16 to 21, many of the statistical methods for the analysis of binary outcome variables are based on the odds of an event, rather than on its probability.

More formally, the **odds** of event A are defined as the probability that A *does* happen *divided* by the probability that it *does not* happen:

$$\text{Odds}\,(\text{A}) = \frac{\text{prob}\,(\text{A happens})}{\text{prob}\,(\text{A does not happen})} = \frac{\text{prob}\,(\text{A})}{1 - \text{prob}\,(\text{A})}$$

since $1 - \text{prob}\,(\text{A})$ is the probability that A does not happen. By manipulating this equation, it is also possible to express the probability in terms of the odds:

$$\text{Prob}\,(\text{A}) = \frac{\text{Odds}\,(\text{A})}{1 + \text{Odds}\,(\text{A})}$$

Thus it is possible to derive the odds from the probability, and *vice versa*.

When bookmakers offer bets they do so in terms of the odds that the event *will not* happen, since the probability of this is usually greater than that of the event happening. Thus, if the odds on a horse in a race are 4 to 1, this means that the bookmaker considers the probability of the horse losing to be four times greater than the probability of the horse winning. In other words:

$$\text{Odds}\,(\text{horse loses}) = \frac{\text{prob}\,(\text{horse loses})}{\text{prob}\,(\text{horse wins})} = 4$$

Table 14.2 Values of the odds, for different values of the probability.

Probability	Odds
0	0
0.001	0.001001
0.005	0.005025
0.01	0.010101
0.05	0.052632
0.1	0.111111
0.2	0.25
0.5	1
0.9	9
0.95	19
0.99	99
0.995	199
0.999	999
1	∞

Using the equation above, it follows that $\text{prob (horse loses)} = 4/(1+4) = 0.8$, and the probability that it wins is 0.2.

Table 14.2 shows values of the odds corresponding to different values of the probability. It can be seen that the difference between the odds and the probability is small unless the probability is greater than about 0.1. It can also be seen that while probabilities must lie between 0 and 1, odds can take any value between 0 and infinity (∞). This is a major reason why odds are commonly used in the statistical analysis of binary outcomes. Properties of odds are summarized in the box below.

BOX 14.1 PROPERTIES OF THE ODDS

- Both prob (A) and $1 - \text{prob (A)}$ lie between 0 and 1. It follows that the odds lie between 0 (when $\text{prob (A)} = 0$) and ∞ (when $\text{prob (A)} = 1$)
- When the probability is 0.5, the odds are $0.5/(1 - 0.5) = 1$
- The odds are always bigger than the probability (since $1 - \text{prob (A)}$ is less than one)
- **Importantly**: When the probability is small (about 0.1 or less), the odds are very close to the probability. This is because for a small probability $[1 - \text{prob (A)}] \cong 1$ and so $\text{prob (A)}/[1 - \text{prob (A)}] \cong \text{prob (A)}$

CHAPTER 15

Proportions and the binomial distribution

15.1 INTRODUCTION

In this chapter we start by introducing the notation for binary outcome variables that will be used throughout the book. These are outcomes where for each individual in the sample the outcome is one of two alternatives. For example, at the end of the study a subject may have experienced the particular disease (or event) of interest (D), or remained healthy (H). Throughout this part, we will label the two possible outcomes as D (disease) or H (healthy), regardless of the actual categories. Examples of other outcome variables are that a patient dies (D) or survives (H), or that a specimen is positive (D) or negative (H). It is *not* necessary that D refers to an adverse outcome; for example, in a smoking cessation study, our outcome may be that a participant has (D) or has not (H) successfully quit smoking after 6 months.

Of particular interest is the *proportion* (p) of individuals in our sample in category D, that is the number of subjects who experience the event (denoted by d) divided by the total number in the sample (denoted by n). The total who do not experience the event will be denoted throughout by $h = n - d$.

$$p = \frac{d}{n}$$

We use this **sample proportion** to estimate the **probability** or **risk** (see Section 14.2) that an individual in the *population* as a whole will be in category D rather than H.

Example 15.1

Suppose that in a trial of a new vaccine, 23 of 1000 children vaccinated showed signs of adverse reactions (such as fever or signs of irritability) within 24 hours of vaccination. The proportion exhibiting an adverse reaction was therefore:

$$p = 23/1000 = 0.023 \text{ or } 2.3\%$$

We would then advise parents of children about to be vaccinated that the vaccine is associated with an estimated 2.3% risk of adverse reactions. See Section 15.5 for how to calculate a confidence interval for such a proportion.

The (unknown) probability or risk that the outcome D occurs in the population is denoted by π (Greek letter pi; not related here to the mathematical constant 3.14159). Its estimation is, of course, subject to **sampling variation**, in exactly the same way as the estimation of a population mean from a sample mean, described in Section 4.5. In the following sections, we derive the sampling distribution of a proportion, which is known as the binomial distribution, and then show how it can be approximated by the normal distribution to give a confidence interval and z-test for a single proportion. Finally, we define two types of proportion that are of particular importance in medical research; cumulative incidence (risk) and prevalence.

15.2 BINOMIAL DISTRIBUTION: THE SAMPLING DISTRIBUTION OF A PROPORTION

The **sampling distribution of a proportion** is called the **binomial distribution** and can be calculated from the sample size, n, and the *population* proportion, π, as shown in Example 15.2. π is the probability that the outcome for any one individual is D.

Example 15.2

A man and woman each with sickle cell trait (AS; that is, heterozygous for the sickle cell [S] and normal [A] haemoglobin genes) have four children. What is the probability that none, one, two, three, or four of the children have sickle cell disease (SS)?

For each child the probability of being SS is the probability of having inherited the S gene from each parent, which is $0.5 \times 0.5 = 0.25$ by the multiplicative rule of probabilities (see Section 14.3). The probability of not being SS (i.e. of being AS or AA) is therefore 0.75. We shall call being SS category D and not being SS category H, so $\pi = 0.25$.

The probability that none of the children is SS (i.e. $d = 0$) is $0.75 \times 0.75 \times 0.75 \times 0.75 = 0.75^4 = 0.3164$ (0.75^4 means 0.75 multiplied together four times). This is by the multiplicative rule of probabilities.

The probability that exactly one child is SS (i.e. $d = 1$) is the probability that (first child SS; second, third, fourth not SS) or (second child SS; first, third, fourth not SS) or (third child SS; first, second, fourth not SS) or (fourth child SS; first, second, third not SS). Each of these four possibilities has probability 0.25×0.75^3 (multiplicative rule) and since they cannot occur together the probability of one or other of them occurring is $4 \times 0.25 \times 0.75^3 = 0.4219$, by the additive rule of probabilities (see Section 14.3).

Table 15.1 Calculation of the probabilities of the possible numbers of children who have inherited sickle cell (SS) disease, in a family of four children where both parents have the sickle cell trait. (The probability that an individual child inherits sickle cell disease is 0.25.)

No. of children			Probability
With SS (d)	WithoutSS (h)	No. of ways in which combination could occur	Prob (d events) $= \dfrac{n!}{d!(n-d)!}\pi^d(1-\pi)^{n-d}$
0	4	1	$1 \times 1 \times 0.75^4 = 0.3164$
1	3	4	$4 \times 0.25 \times 0.75^3 = 0.4219$
2	2	6	$6 \times 0.25^2 \times 0.75^2 = 0.2109$
3	1	4	$4 \times 025^3 \times 0.75 = 0.0469$
4	0	1	$1 \times 0.25^4 \times 1 = 0.0039$
			Total $= 1.0000$

In similar fashion, one can calculate the probability that exactly two, three, or four children are SS by working out in each case the different possible arrangements within the family and adding together their probabilities. This gives the probabilities shown in Table 15.1. Note that the sum of these probabilities is 1, which it has to be as one of the alternatives must occur.

The probabilities are also illustrated as a probability distribution in Figure 15.1. This is the **binomial probability distribution** for $\pi = 0.25$ and $n = 4$.

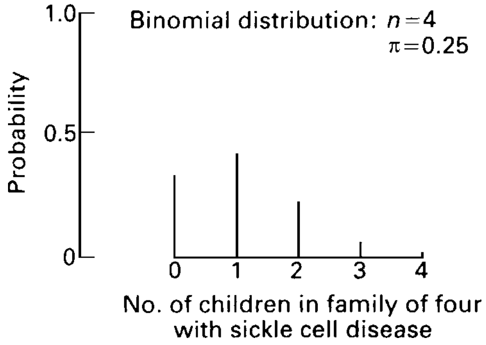

Fig. 15.1 Probability distribution of the number of children in a family of four with sickle cell disease where both parents have the sickle cell trait. The probability that a child inherits sickle cell disease is 0.25.

General formula for binomial probabilities

The general formula for the probability of getting exactly d events in a sample of n individuals when the probability of D for each individual is π is:

$$\text{Prob}\,(d \text{ events}) = \frac{n!}{d!(n-d)!}\pi^d(1-\pi)^{n-d}$$

The first part of the formula represents the number of possible ways in which d events could be observed in a sample of size n, and the second part equals the probability of each of these ways.

- The *exclamation mark* denotes the *factorial* of the number and means all the integers from the number down to 1 multiplied together. (0! is defined to equal 1.)
- π^d means π multiplied together d times or, in mathematical terminology, π to the power d. Any number to the power zero is defined to equal 1.
- Note that when π equals 0.5, $(1-\pi)$ also equals 0.5 and the second part of the formula simplifies to 0.5^n.

The interested reader may like to practise the application of the above formula by checking the calculations presented in Table 15.1. For example, applying the formula in the above example to calculate the probability that exactly two out of the four children are SS gives:

$$\begin{aligned}\text{Prob}\,(2 \text{ SS children}) &= \frac{4!}{2!(4-2)!}0.25^2(1-0.25)^{4-2}\\ &= \frac{4\times 3\times 2\times 1}{2\times 1\times 2\times 1}0.25^2(0.75)^2\\ &= 6\times 0.25^2\times 0.75^2 = 0.2109\end{aligned}$$

The *first part of the formula* may be more easily calculated using the following expression, where $(n-d)!$ has been cancelled into n!

$$\frac{n!}{d!(n-d)!} = \frac{n\times(n-1)\times(n-2)\times\ldots\times(n-d+1)}{d\times(d-1)\times\ldots 3\times 2\times 1}$$

For example, if $n = 18$ and d $= 5$, $(n-d+1) = 18-5+1 = 14$ and the expression equals:

$$\frac{18\times 17\times 16\times 15\times 14}{5\times 4\times 3\times 2\times 1} = \frac{1028160}{120} = 8568$$

Shape of the binomial distribution

Figure 15.2 shows examples of the binomial distribution for various values of π and n. These distributions have been illustrated for d, the number of events in the sample, although they apply equally to p, the proportion of events. For example, when the sample size, n, equals 5, the possible values for d are 0, 1, 2, 3, 4 or 5, and the horizontal axis has been labelled accordingly. The corresponding proportions are 0, 0.2, 0.4, 0.6, 0.8 and 1 respectively. Relabelling the horizontal axis with these values would give the binomial distribution for p. Note that, although p is a fraction, its sampling distribution is discrete and not continuous, since it may take only a limited number of values for any given sample size.

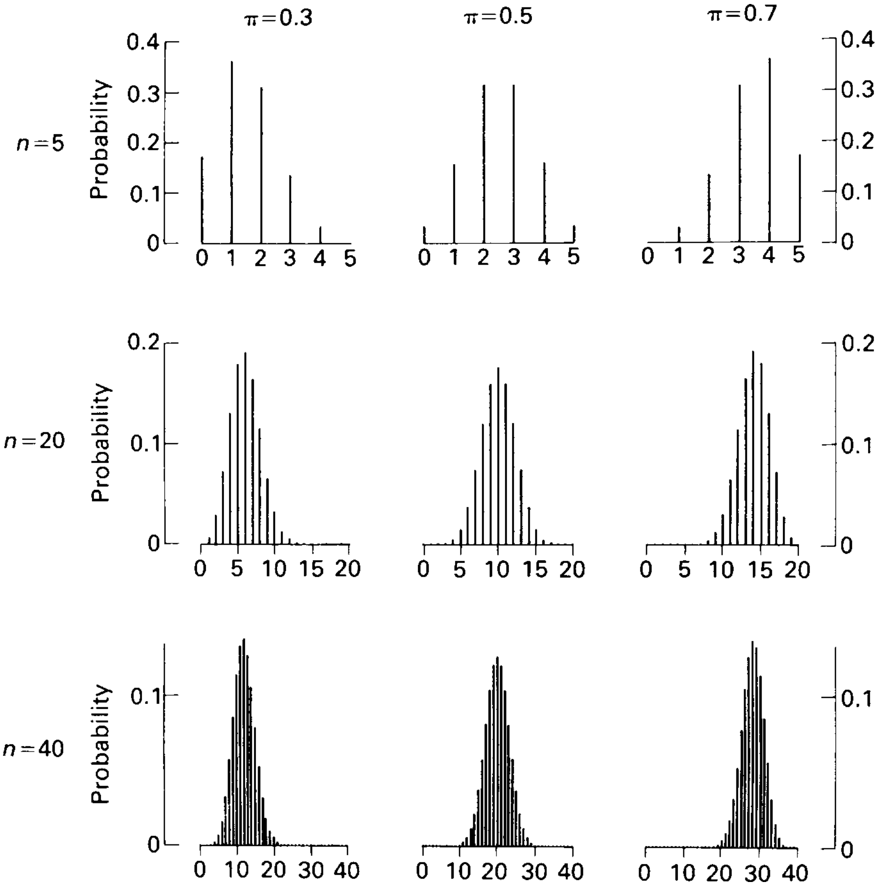

Fig. 15.2 Binomial distribution for various values of π and n. The horizontal scale in each diagram shows values of d.

15.3 STANDARD ERROR OF A PROPORTION

Since the binomial distribution is the **sampling distribution** for the number (or proportion) of D's, its mean equals the population mean and its standard deviation represents the **standard error**, which measures how closely the sample value estimates the population value. The population means and standard errors can be calculated from the binomial probabilities; the results are given in Table 15.2 for the number, proportion and percentage of events. The percentage is, of course, just the proportion multiplied by 100.

Table 15.2 Population mean and standard error for the number, proportion and percentage of D's in a sample.

	Observed value	Population mean	Standard error
Number of events	d	$n\pi$	$\sqrt{[n\pi(1-\pi)]}$
Proportion of events	$p = d/n$	π	$\sqrt{[\pi(1-\pi)/n]}$
Percentage of events	$100p$	100π	$100\sqrt{[\pi(1-\pi)/n]}$

15.4 NORMAL APPROXIMATION TO THE BINOMIAL DISTRIBUTION

As the sample size n increases the binomial distribution becomes very close to a **normal distribution** (see Figure 15.2), and this can be used to calculate confidence intervals and carry out hypothesis tests as described in the following sections. In fact the normal distribution can be used as a reasonable approximation to the binomial distribution if both $n\pi$ and $n - n\pi$ are 10 or more. This approximating normal distribution has the same mean and standard error as the binomial distribution (see Table 15.2).

15.5 CONFIDENCE INTERVAL FOR A SINGLE PROPORTION USING THE NORMAL DISTRIBUTION

The calculation and interpretation of confidence intervals was explained in detail in Chapters 6 and 8. Using the binomial distribution to derive a confidence interval for a proportion is complicated. Methods that do this are known as **exact methods** and are described in more detail by Altman *et al.* (2000), and by Clayton and Hills (1993). The usual approach is to use the approximation to the normal distribution with π estimated by p, the standard error estimated by $\sqrt{[p(1-p)/n]}$ (see Table 15.2), and methods similar to those described in Chapter 6 for means. This is valid providing that both np and $n - np$ are 10 or more, so that the normal approximation to the binomial distribution is sufficiently good. The confidence interval is:

$$\text{CI} = p - (z' \times \text{s.e.}) \text{ to } p + (z' \times \text{s.e.}),$$
$$\text{s.e.} = \sqrt{[p(1-p)/n]}$$

where z' is the appropriate percentage point of the standard normal distribution. For example, for a 95% confidence interval, $z' = 1.96$.

Example 15.3

In September 2001 a survey of smoking habits was conducted in a sample of 1000 teenagers aged 15–16, selected at random from all 15–16 year-olds living in Birmingham, UK. A total of 123 reported that they were current smokers. Thus the proportion of current smokers is:

$$p = 123/1000 = 0.123 = 12.3\%$$

The standard error of p is estimated by $\sqrt{[p(1-p)/n]} = \sqrt{0.123 \times 0.877/1000} = 0.0104$. Thus the 95% confidence interval is:

$$95\%\ \text{CI} = 0.123 - (1.96 \times 0.0104) \text{ to } 0.123 + (1.96 \times 0.0104) = 0.103 \text{ to } 0.143$$

With 95% confidence, in September 2001 the proportion of 15–16 year-olds living in Birmingham who smoked was between 0.103 and 0.143 (or equivalently, between 10.3% and 14.3%).

15.6 *z*-TEST THAT THE POPULATION PROPORTION HAS A PARTICULAR VALUE

The approximating normal distribution (to the binomial sampling distribution) can also be used in a ***z*-test** of the null hypothesis that the population proportion equals a particular value, π. This is valid provided that both $n\pi$ and $n - n\pi$ are greater than or equal to 10. The z-test compares the size of the difference between the sample proportion and the hypothesized value, with the standard error. The formula is:

$$z = \frac{p - \pi}{s.e.(p)} = \frac{p - \pi}{\sqrt{[\pi(1-\pi)/n]}}$$

In exactly the same way as explained in Chapter 8, we then derive a P-value, which measures the strength of the evidence against the null hypothesis that $p = \pi$.

Example 15.3 (continued)

In 1998 the UK Government announced a target of reducing smoking among children from the national average of 13% to 9% or less by the year 2010, with a fall to 11% by the year 2005. Is there evidence that the proportion of 15–16 year-old smokers in Birmingham at the time of our survey in 2001 was below the national average of 13% at the time the target was set?

The null hypothesis is that the population proportion is equal to 0.13 (13%). The sampling distribution for the number of smokers, if the null hypothesis is true, is therefore a binomial distribution with $\pi = 0.13$ and $n = 1000$. The standard error of p under the null hypothesis is:

$$\text{s.e.}(\pi) = \sqrt{[0.13(1 - 0.13)/1000]} = 0.0106. \text{ Therefore } z = \frac{0.123 - 0.13}{0.0106} = -0.658$$

The corresponding P-value is 0.51. There is no evidence that the proportion of teenage smokers in Birmingham in September 2001 was lower than the national 1998 levels.

Continuity correction

When either $n\pi$ or $n - n\pi$ are below 10, but both are 5 or more, the accuracy of hypothesis tests based on the normal approximation can be improved by the introduction of a **continuity correction** (see also Section 17.2). The continuity correction adjusts the numerator of the test statistic so that there is a closer fit between the P-value based on the z-test and the P-value based on an exact calculation using the binomial probabilities. This is illustrated in Figure 15.3 and Table 15.3, which show that incorporating a continuity correction and calculating the area under the normal curve above 8.5 gives a close approximation to the exact binomial probability of observing 9 events or more. In contrast the area of the normal curve above 9

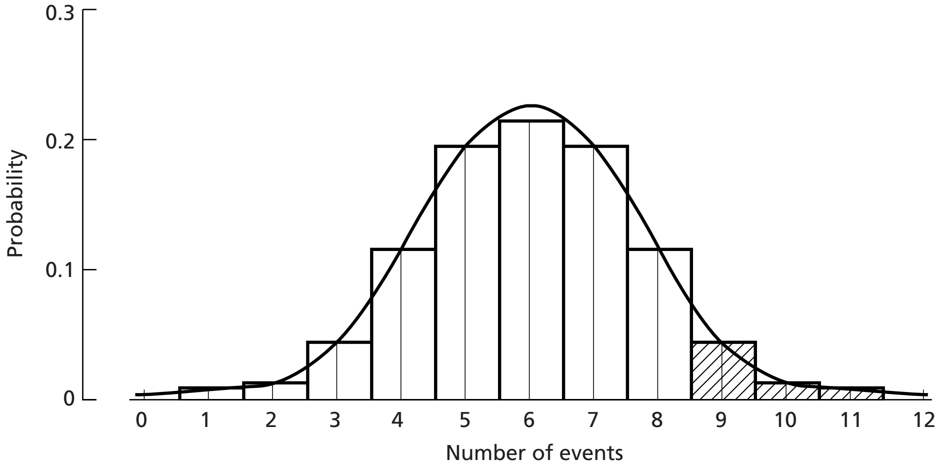

Fig. 15.3 Comparison of the binomial distribution ($n = 12, \pi = 0.5$) with the approximating normal distribution to illustrate the need for a continuity correction for small n. This shows that the area under the normal curve above 8.5 is closer to the shaded exact probabilities than the area above 9.

Table15.3 Comparisons of the different methods of calculating the probability of observing 9 or more events, when $n = 12$ and $\pi = 0.5$.

Probability of observing 9 or more events, when $n = 12$ and $\pi = 0.5$	
Calculated using binomial probabilities:	
9 events	$220 \times 0.5^{12} = 0.0537$
10 events	$66 \times 0.5^{12} = 0.0161$
11 events	$12 \times 0.5^{12} = 0.0029$
12 events	$1 \times 0.5^{12} = 0.0002$
Total of 9+ events	0.0729
Using approximating normal distribution:	
Based on area above 9	0.0418
With continuity correction, based on area above 8.5	0.0749

is not a good approximation. More details are not included here since continuity corrections are not often used in modern medical statistics. This is because they can't be extended to the regression models, described in Chapter 19 and later in the book, which are used to examine the effects of a *numbe*r of exposure variables on a binary outcome.

15.7 INCIDENCE AND PREVALENCE

We now define two particular types of proportion that are of particular relevance in medical research. These are the cumulative incidence (or risk) of a disease event, and the prevalence of a disease.

Cumulative incidence (risk)

The **cumulative incidence** or **risk**, *r*, of a disease event is the probability that the disease event occurs during a specified period of time. It is estimated by the number of new cases of a disease during a specified period of time divided by the number of persons initially disease-free and therefore at risk of contracting the disease.

$$\text{Risk} = \text{cumulative incidence} = \frac{\text{number of new cases of disease in period}}{\text{number initially disease-free}}$$

For example, we might be interested in:

- the risk of death in the five years following diagnosis with prostate cancer;
- the risk of vertical transmission of HIV during pregnancy or childbirth in HIV-infected mothers given antiretroviral therapy during pregnancy.

Risks usually refer to adverse (undesirable) events, though this is not essential.

Example 15.4

Suppose we study 5000 individuals aged 45 to 54, with no existing cardiovascular disease. Ten years later, the same individuals are followed up and we find that 147 have died from or have developed coronary heart disease. Then the **risk** of coronary heart disease is the proportion of individuals who developed the disease: $147/5000 = 0.0294$, or 2.94%.

Prevalence

In contrast, the **prevalence** represents the burden of disease at a particular time, rather than the chance of future disease. It is based on the total number of existing cases among the whole population, and represents the probability that any one individual in the population is currently suffering from the disease.

$$\text{Prevalence} = \frac{\text{number of people with the disease at particular point in time}}{\text{total population}}$$

For example, we might be interested in:

- the prevalence of schistosomiasis among villagers living on the shore of Lake Malawi;
- the prevalence of chronic lower back pain among refuse collectors in Bristol, UK.

Example 15.5

Suppose we study a sample of 2000 individuals aged 15 to 50, registered with a particular general practice. Of these, 138 are being treated for asthma. Then the **prevalence** of diagnosed asthma in the practice population is the proportion of the sample with asthma: $138/2000 = 0.069$, or 6.9%.

Both cumulative incidence and prevalence are usually expressed as a percentage or, when small, as per 1000 population or per 10 000 or 100 000 population. In Chapter 22 we define the **incidence rate**, the measure used in longitudinal studies with variable lengths of follow up.

CHAPTER 16

Comparing two proportions

16.1 INTRODUCTION

In Chapter 15 we saw how the sampling distribution of a proportion can be approximated by the normal distribution to give a confidence interval and *z*-test for a single proportion. In this chapter we deal with the more common situation where we wish to compare the occurrence of a binary outcome variable between *two exposure (or treatment) groups*. We will use the same notation for these two groups as was introduced in Chapter 7 for the comparison of two means. Group 1 denotes individuals *exposed* to a risk factor, and group 0 denotes those *unexposed*. In clinical trials, group 1 denotes the *treatment* group, and group 0 the *control*, or *placebo* group (a **placebo** is a preparation made to be as similar as possible to the treatment in all respects, but with no effective action). For example,

- In a study of the effects of bacterial infection during pregnancy, we may wish to compare the risk of premature delivery for babies born to women infected during the first trimester (the exposed group, 1) with that for babies born to uninfected women (the unexposed group, 0).
- In a trial of a new influenza vaccine, the comparison of interest might be the proportion of participants who succumbed to influenza during the winter season in the vaccine group (the treatment group, 1), compared to the proportion in the placebo group (the control group, 0).

We start by showing how the data can be displayed in a **2 × 2 table**, with individuals in the sample classified according to whether they experienced the disease outcome (or not), and according to whether they were exposed (or not). We then explain three different measures for comparing the outcome between the two groups: the difference in the two proportions, the risk ratio and the odds ratio. We describe how to calculate a confidence interval and carry out a hypothesis test for each of them, and outline their relative advantages and disadvantages.

16.2 THE 2 × 2 TABLE, AND MEASURES OF EXPOSURE EFFECT

In Section 3.4, we described how the relationship between two categorical variables can be examined by cross-tabulating them in a **contingency table**. We noted that a useful convention is for the rows of the table to correspond to the exposure values and the columns to the outcomes. To compare the occurrence of a binary outcome variable between two exposure groups, we therefore display the data in a **2 × 2 table**. Table 16.1 shows the notation that we will use for the number of individuals in each group. As introduced in the last chapter, we use letter d to denote the number of subjects who experience the outcome event, h to denote the number of subjects who do not experience the outcome event, and n for the total number in the sample. In addition, we use the subscripts 1 and 0 to denote the exposed and unexposed groups respectively.

As explained in Section 3.4, it is recommended that the table also shows the proportion (or percentage) in each outcome category, within each of the exposure groups. Thus, if the exposure is the row variable (as here) then row percentages should be presented, while if it is the column variable then column percentages should be presented. Following the notation introduced in Chapter 15, the overall proportion is denoted by $p = d/n$, and the proportions in the exposed and unexposed groups are denoted by $p_1 = d_1/n_1$ and $p_0 = d_0/n_0$, respectively.

Example 16.1

Consider the following results from an influenza vaccine trial carried out during an epidemic. Of 460 adults who took part, 240 received influenza vaccination and 220 placebo vaccination. Overall 100 people contracted influenza, of whom 20 were in the vaccine group and 80 in the placebo group. We start by displaying the results of the trial in a 2 × 2 table (Table 16.2). In Table 16.2 the exposure is vaccination (the row variable) and the outcome is whether the subject contracts influenza (the column variable). We therefore also include *row* percentages in the

Table 16.1 Notation to denote the number of individuals in each group for the 2 × 2 table comparing a binary outcome variable between two exposure groups.

	Outcome		
Exposure	Experienced event: D (Disease)	Did not experience event: H (Healthy)	Total
Group 1 (exposed)	d_1	h_1	n_1
Group 0 (unexposed)	d_0	h_0	n_0
Total	d	h	n

Table 16.2 2 × 2 table showing results from an influenza vaccine trial.

	Influenza		
	Yes	No	Total
Vaccine	20 (8.3%)	220 (91.7%)	240
Placebo	80 (36.4%)	140 (63.6%)	220
Total	100 (21.7%)	360 (78.3%)	460

table. Overall, 21.7% of subjects contracted influenza. We can see that the percentage contracting influenza was much lower in the vaccine group (8.3%), than in the placebo group (36.4%). We can use these data to answer the following related questions.

1 How effective was the vaccine in preventing influenza in our trial? The size of this effect can be measured in three different ways:
   (a) The **difference** between the **risks** of contracting influenza in the vaccine group compared to the placebo group.
   (b) The **ratio** of the **risks** of contracting influenza in the vaccine group compared to the placebo group. This is also known as the **relative risk**.
   (c) The **ratio** of the **odds** of contracting (to not contracting) influenza in the vaccine group, compared to the placebo group.

2 What does the effect of the vaccine in our trial tell us about the size of its effect in preventing influenza more generally in the population? This is addressed by calculating a **confidence interval** for the size of the effect.

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

The use of confidence intervals and *P*-values to interpret the results of statistical analyses is discussed in detail in Chapter 8, and readers may wish to refer to that chapter at this point.

The three different measures for comparing a binary outcome between two exposure (or treatment) groups are summarized in Table 16.3, together with the results for the influenza vaccine trial. All three measures indicate a benefit of the vaccine. The risk difference is −0.281, meaning that the *absolute* risk of contracting influenza was 0.281 *lower* in the vaccine group compared to the placebo group. The risk ratio equals 0.228, meaning that the risk of contracting influenza in the vaccine group was only 22.8% of the risk in the placebo group. Equivalently, we could say the vaccine prevented 77.2% (100 – 22.8%) of influenza cases. This is called the **vaccine efficacy**; it is discussed in more detail in Chapter 37. The odds

Table 16.3 Three different measures for comparing a binary outcome between two exposure (or treatment) groups, together with the results for the vaccine trial data in Table 16.2.

Measure of comparison	Formula	Result for influenza vaccine trial
Risk difference	$p_1 - p_0$	$0.083 - 0.364 = -0.281$
Risk ratio (relative risk)	p_1/p_0	$0.083/0.364 = 0.228$
Odds ratio	$\dfrac{d_1/h_1}{d_0/h_0} = \dfrac{d_1 \times h_0}{d_0 \times h_1}$	$\dfrac{20/220}{80/140} = \dfrac{20 \times 140}{80 \times 220} = 0.159$

ratio in the trial was 0.292 meaning that the *odds* of contracting influenza in the vaccine group were 29.2% of the odds in the placebo group.

The following sections describe how to calculate confidence intervals and carry out hypothesis tests for each of these three measures. They also discuss their relative advantages and disadvantages. When to use which measure is also discussed in Chapter 37 ('Measures of association and impact').

16.3 RISK DIFFERENCES

We will start with the first of the three measures of effect, the difference between the two proportions. From now on we will refer to this as a **risk difference**, though the methods apply to any type of proportion. We will see how to derive a confidence interval for the difference, and carry out a test of the null hypothesis that there is no difference between the proportions in the population from which the sample was drawn. As in the case of a single proportion we will use methods based on the normal approximation to the sampling distribution of the two proportions. These will be illustrated in the context of the influenza vaccine trial described in Example 16.1 above.

Sampling distribution of the difference between two proportions

Before we can construct a confidence interval for the difference between two proportions, or carry out the related hypothesis test, we need to know the sampling distribution of the difference. The difference, $p_1 - p_0$, between the proportions in the exposed and unexposed groups in our sample provides an estimate of the underlying difference, $\pi_1 - \pi_0$, between the exposed and unexposed groups in the population. It is of course subject to *sampling variation*, so that a different sample from the same population would give a different value of $p_1 - p_0$. Note that:

1 The normal distribution is a reasonable approximation to the sampling distribution of the difference $p_1 - p_0$, provided $n_1 p_1$, $n_1 - n_1 p_1$, $n_0 p_0$ and $n_0 - n_0 p_0$ are each greater than 10, and will improve as these numbers get larger.

2 The mean of this sampling distribution is simply the difference between the two population means, $\pi_1 - \pi_0$.

3 The standard error of $p_1 - p_0$ is based on a combination of the standard errors of the individual proportions:

$$\text{s.e.}(p_1 - p_0) = \sqrt{[p_1(1 - p_1)/n_1 + p_0(1 - p_0)/n_0]} = \sqrt{[\text{s.e.}(p_1)^2 + \text{s.e.}(p_0)^2]}$$

The confidence interval for the difference between two proportions is given by:

$$\text{CI} = (p_1 - p_0) - z' \times \text{s.e.}(p_1 - p_0) \text{ to } (p_1 - p_0) + z' \times \text{s.e.}(p_1 - p_0)$$

where z' is the appropriate percentage point of the normal distribution.

Example 16.1 (continued)

The difference in proportions between the vaccine and placebo groups is 0.083−0.364 = −0.281. Its standard error is:

$$\text{s.e.}(p_1 - p_0) = \sqrt{[0.083(1 - 0.083)/240 + 0.364(1 - 0.364)/220]} = 0.037$$

and so the approximate 95% confidence interval for this reduction is:

$$\begin{aligned} 95\%\ \text{CI} &= -0.281 - (1.96 \times 0.037) \text{ to } -0.281 + (1.96 \times 0.037) \\ &= -0.353 \text{ to } -0.208 \end{aligned}$$

That is, we are 95% confident that in the population the vaccine would reduce the risk of contracting influenza by between 0.208 and 0.353.

Test that the difference between two proportions is zero

The normal test to compare two sample proportions is based on:

$$z = \frac{p_1 - p_0}{\text{s.e.}(p_1 - p_0)}$$

The standard error used in the test is different to that used in the confidence interval because it is calculated assuming that the null hypothesis is true (i.e. that $\pi_1 = \pi_0 = \pi$). Under the null hypothesis that the population proportions are equal:

$$\text{s.e.}(p_1 - p_0) = \sqrt{[\pi(1 - \pi)(1/n_1 + 1/n_0)]}$$

π is estimated by the overall proportion in both samples, that is by:

$$p = \frac{d_0 + d_1}{n_0 + n_1} = \frac{d}{n}$$

The formula for the z-test is therefore:

$$z = \frac{p_1 - p_0}{\sqrt{[p(1-p)(1/n_1 + 1/n_0)]}}$$

This test is a valid approximation provided that either $n_1 + n_0$ is greater than 40 or $n_1 p,\ n_1 - n_1 p,\ n_2 p$ and $n_2 - n_2 p$ are all 10 or more. If this condition is not satisfied, but $n_1 p,\ n_1 - n_1 p,\ n_2 p$ and $n_2 - n_2 p$ are all 5 or more, then a modified version of the z-test incorporating a **continuity correction**, or the equivalent chi-squared test with a continuity correction, can be used (see Section 17.2). If none of these conditions are satisfied, the exact test described in Section 17.3 should be used.

Example 16.1 (continued)

The overall proportion that contracted influenza was 0.217 or 21.7%. Therefore:

$$z = \frac{(0.083 - 0.364)}{\sqrt{[0.217(1 - 0.217)(1/240 + 1/220)]}} = \frac{-0.281}{0.0385} = -7.299$$

The corresponding P-value is < 0.0001. Thus there is strong evidence that there was a reduction in the risk of contracting influenza following vaccination with the influenza vaccine.

16.4 RISK RATIOS

We now turn to the second measure of effect introduced in Section 16.2, the ratio of the two proportions. We will refer to this as the **risk ratio**, although the methods apply to ratios of any proportions, and not just those that estimate risks. The risk ratio is often abbreviated to **RR**, and is also known as the **relative risk**.

$$\text{RR} = \frac{p_1}{p_0} = \frac{d_1/n_1}{d_0/n_0}$$

Example 16.2

Table 16.4 shows hypothetical data from a study to investigate the association between smoking and lung cancer. 30 000 smokers and 60 000 non-smokers were followed for a year, during which time 39 of the smokers and 6 of the non-smokers developed lung cancer, giving risks of 0.13% and 0.01% respectively. Thus the risk of lung cancer was considerably higher among smokers than non-smokers.

Table 16.4 Hypothetical data from a cohort study to investigate the association between smoking and lung cancer. The calculations of risk ratio (RR) and risk difference are illustrated.

	Lung cancer	No lung cancer	Total	Risk
Smokers (exposed)	39	29 961	30 000	$p_1 - 39/30\,000 = 0.0013$ (0.13%)
Non-smokers (unexposed)	6	59 994	60 000	$p_0 - 6/60\,000 = 0.0001$ (0.01%)
Total	45	89 955	90 000	
				Risk difference = 0.13% − 0.01% = 0.12%
				Risk ratio = 0.0013/0.0001 = 13

The risk ratio is:

$$\text{RR} = \frac{p_1}{p_0} = \frac{0.0013}{0.0001} = 13$$

Interpreting the risk ratio

In an *epidemiological study*, comparing an exposed group with an unexposed, the risk ratio is a good indicator of the strength of the association between the exposure and the disease outcome. It equals:

$$\text{Risk ratio (RR)} = \frac{\text{risk in exposed group}}{\text{risk in unexposed group}}$$

In a *clinical trial* to assess the impact of a new treatment, procedure or preventive intervention on disease outcome or occurrence, the risk ratio equals:

$$\text{Risk ratio (RR)} = \frac{\text{risk in treatment group}}{\text{risk in control group}}$$

A risk ratio of 1 occurs when the risks are the same in the two groups and is equivalent to no association between the risk factor and the disease. A risk ratio greater than 1 occurs when the risk of the outcome is higher among those exposed to the factor (or treatment) than among the non-exposed, as in Example 16.2 above, with exposed referring to smoking. A risk ratio less than 1 occurs when the risk is lower among those exposed, suggesting that the factor (or treatment) may be protective. An example is the reduced risk of infant death observed among infants that are breast-fed compared to those that are not. The further the risk ratio is from 1, the stronger the association between exposure (or treatment) and outcome. Note that a risk ratio is always a positive number.

Relationship between risk ratios and risk differences

The risk ratio is more commonly used to measure of the strength of an association than is the difference in risks. This is because the amount by which an exposure (risk factor) multiplies the risk of an event is interpretable regardless of the size of the risk. For example, suppose we followed the population in Example 16.2 above for two years instead of one, and therefore observed exactly double the number of events in each group (here we are ignoring the small number of individuals lost to follow-up because they died in the first year). The risks are now 0.26% in smokers and 0.02% in non-smokers. The risk ratio is 0.26/0.02 = 13; exactly as before. However, the risk difference is now 0.26 − 0.02% = 0.24%, double that observed when there was only one year's follow-up. The use and interpretation of ratio and difference measures of the size of exposure effects is discussed in Chapter 37.

16.5 RISK RATIOS: CONFIDENCE INTERVALS AND HYPOTHESIS TESTS

Standard error and confidence interval for ratio measures

Until now, we have followed exactly the same procedure whenever we wish to calculate a confidence interval. We derive the standard error (s.e.) of the quantity, q, in which we are interested, and determine the multiplier z_α corresponding to the appropriate percentage point of the sampling distribution:

$$\text{CI} = q - z_\alpha \times \text{s.e. to } q + z_\alpha \times \text{s.e}$$

When the sampling distribution is normal, z_α is 1.96 for a 95% confidence interval and:

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

For *ratio measures* such as risk ratios, this can lead to problems when the standard error is large and q is close to zero, because the lower limit of the confidence interval may come out negative despite the fact that the risk ratio is always positive. To overcome this problem, we adopt the following procedure:

1 Calculate the *logarithm* of the risk ratio, and its standard error. The formula for this standard error is derived using the **delta method** (see Box 16.1), and is:

$$\text{s.e.}(\log \text{RR}) = \sqrt{[1/d_1 - 1/n_1 + 1/d_0 - 1/n_0]}$$

Note that s.e.(log RR) should be interpreted as 'standard error of the log RR', and that throughout this book, all logs are to the base e (natural logarithms) unless explicitly denoted by $\log_{10}$ as being logs to the base 10. See Section 13.1 for an explanation of logarithms and the exponential function.

2 Derive a confidence interval for the *log risk ratio* in the usual way:

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

3 *Antilog the confidence limits* obtained, to convert this into a confidence interval for the risk ratio.

$$95\%\ \text{CI (RR)} = \exp[\log \text{RR} - 1.96 \times \text{s.e.}(\log \text{RR})] \text{ to } \exp[\log \text{RR} + 1.96 \times \text{s.e.}(\log \text{RR})]$$

4 Use the rules of logarithms and antilogs to make this simpler. The rules are:

Rules of logarithms:

$$\log(\text{a}) + \log(\text{b}) = \log(\text{a} \times \text{b})$$
$$\log(\text{a}) - \log(\text{b}) = \log(\text{a}/\text{b})$$

Rules of antilogs:

exp(a) means e^{a}; it is the antilog (exponential) function

$$\exp[\log(\text{a})] = \text{a}$$
$$\exp(\text{a} + \text{b}) = \exp(\text{a}) \times \exp(\text{b})$$
$$\exp(\text{a} - \text{b}) = \exp(\text{a})/\exp(\text{b})$$

Following these rules, and noting that $\exp(\log \text{RR}) = \text{RR}$, gives:

$$95\%\ \text{CI (RR)} = \text{RR}/\exp[1.96 \times \text{s.e.}(\log \text{RR})] \text{ to } \text{RR} \times \exp[1.96 \times \text{s.e.}(\log \text{RR})]$$

The quantity exp[1.96 × s.e. (log RR)] is known as an **error factor (EF)**; it is always greater than 1, because exp(x) is greater than 1 if x is greater than zero. The 95% confidence interval can therefore be written more simply as:

$$95\%\ \text{CI (RR)} = \text{RR}/\text{EF} \text{ to } \text{RR} \times \text{EF}$$

Putting all of this together, the formula for the **95% confidence interval for the risk ratio** is:

$$95\%\ \text{CI (RR)} = \text{RR}/\text{EF} \text{ to } \text{RR} \times \text{EF},$$
$$\text{where EF} = \exp[1.96 \times s.e.(\log RR)]$$
$$\text{and s.e.}(\log \text{RR}) = \sqrt{}[1/d_1 - 1/n_1 + 1/d_0 - 1/n_0]$$

BOX 16.1 DERIVATION OF THE FORMULA FOR THE STANDARD ERROR OF THE LOG(RISK RATIO)

This box is intended for those who wish to understand the mathematics behind the approximate formula for the *standard error of the log (risk ratio)* used in step 1 of the procedure described in Section 16.5, for calculating a confidence interval for the risk ratio.

The formula was derived using the **delta method**. This is a technique for calculating the standard error of a *transformed* variable from the mean and standard error of the original *untransformed* variable. In this Box, we briefly outline how this method is used to give (a) an approximate formula for the standard error of a *log transformed variable*, and in particular (b) the formula for the standard error of a *log transformed proportion*. We then show how this result can be used to derive (c) an approximate *formula for the standard error of the log(risk ratio)*.

(a) Deriving the formula for the standard error of a log transformed variable:

The delta method uses a mathematical technique known as a Taylor series expansion to show that:

$$\log(\mathrm{X}) \simeq \log(\mu) + (\mathrm{X} - \mu)(\log'(\mu))$$

where $\log'(\mu)$ denotes the first derivative of $\log(\mu)$, the slope of the graph of $\log(\mu)$ against μ. This approximation works provided that the variance of variable X is small compared to its mean.

As noted in Section 4.3, adding or subtracting a constant to a variable leaves its standard deviation (and variance) unaffected, and multiplying by a constant has the effect of multiplying the standard deviation by that constant (or equivalently multiplying the variance by the square of the constant). By applying these in the formula above, and further noting that $\log'(\mu) = 1/\mu$, we can deduce that

$$\text{s.e.}(\log(\mathrm{X})) \simeq \text{s.e.}(\mathrm{X}) \times \log'(\mu) = \text{s.e.}(\mathrm{X})/\mu$$

(b) Formula for the standard error of the log(proportion):

Recall from Section 15.3 that the mean of the sampling distribution for a proportion is estimated by $p = d/n$ and the standard error by $\sqrt{[p(1-p)/n]}$. Therefore:

$$\text{s.e.}(\log p) \simeq \frac{\sqrt{[p(1-p)/n]}}{d/n} = \sqrt{[1/d - 1/n]}$$

(c) Formula for the standard error of the log(risk ratio):

$$\text{Risk ratio (RR)} = \frac{p_1}{p_0}$$

Using the rules of logarithms given above the log risk ratio is given by:

$$\log \mathrm{RR} = \log(p_1) - \log(p_0)$$

Since the standard error of the difference between two variables is the square root of the sum of their variances (see Section 7.2), it follows that the standard error of log RR is given by:

$$\text{s.e.}(\log \mathrm{RR}) = \sqrt{[\mathrm{var}(\log(p_1) + \mathrm{var}(\log(p_0)]} = \sqrt{[1/d_1 - 1/n_1 + 1/d_0 - 1/n_0]}$$

Example 16.2 (continued)

Consider the data presented in Table 16.4, showing a risk ratio of 13 for the association between smoking and risk of lung cancer. The standard error of the log RR is given by:

$$\text{s.e.}(\log \text{RR}) = \sqrt{}[(1/39 - 1/30000 + 1/6 - 1/60000)] = 0.438$$

The error factor is given by:

$$\text{EF} = \exp(1.96 \times 0.438) = 2.362$$

The 95% confidence interval for the risk ratio is therefore:

$$95\%\ \text{CI} = (13/2.362 \text{ to } 13 \times 2.362) = 5.5 \text{ to } 30.7$$

Test of the null hypothesis

If the null hypothesis of no difference between the risks in the two groups is true, then the RR = 1 and hence log RR = 0. We use the log RR and its standard error to derive a z statistic and test the null hypothesis in the usual way:

$$z = \frac{\log RR}{\text{s.e.}(\log RR)}$$

Example 16.2 (continued)

In the smoking and lung cancer example,

$$z = 2.565/0.438 = 5.85$$

This corresponds to a P-value of < 0.0001. There is therefore strong evidence against the null hypothesis that the RR = 1.

Further analyses of risk ratios

The risk ratio is a measure that is easy to interpret, and the analyses based on risk ratios described in this chapter are straightforward. Perhaps surprisingly, however, more complicated analyses of associations between exposures and binary outcomes are rarely based on risk ratios. It is much more common for these to be based on *odds ratios*, as discussed in the next section, and used throughout

Chapters 17 to 21. In Section 20.4, we briefly describe how to conduct regression analyses based on risk ratios, rather than odds ratios, and why this is not usually the preferred method.

16.6 ODDS RATIOS

We now turn to the third and final measure of effect introduced in Section 16.2, the *ratio* of the *odds* of the outcome event in the exposed group compared to the odds in the unexposed group (or in the case of a clinical trial, in the treatment group compared to the control group). Recall from Section 14.6 that the odds of an outcome event D are defined as:

$$\text{Odds} = \frac{\text{prob(D happens)}}{\text{prob(D does not happen)}} = \frac{\text{prob(D)}}{1 - \text{prob(D)}}$$

The odds are *estimated* by:

$$\text{Odds} = \frac{p}{1-p} = \frac{d/n}{(1-d/n)} = \frac{d/n}{h/n} = \frac{d}{h}$$

i.e. by the number of individuals who experience the event divided by the number who do not experience the event. The **odds ratio** (often abbreviated to **OR**) is estimated by:

$$\text{OR} = \frac{\text{odds in exposed group}}{\text{odds in unexposed group}} = \frac{d_1/h_1}{d_0/h_0} = \frac{d_1 \times h_0}{d_0 \times h_1}$$

It is also known as the **cross-product ratio** of the 2 × 2 table.

Example 16.3

Example 15.5 introduced a survey of 2000 patients aged 15 to 50 registered with a particular general practice, which showed that 138 (6.9%) were being treated for asthma. Table 16.5 shows the number diagnosed with asthma according to their gender. Both the prevalence (proportion with asthma) and odds of asthma in women and men are shown, as are their ratios.

The odds ratio of 1.238 indicates that asthma is more common among women than men. In this example the odds ratio is close to the ratio of the prevalences; this is because the prevalence of asthma is low (6% to 8%). Properties of odds ratios are summarized in Box 16.2.

Table 16.5 Hypothetical data from a survey to examine the prevalence of asthma among patients at a particular general practice.

	Asthma	No asthma	Total	Prevalence	Odds
Women	81	995	1076	0.0753	0.0814
Men	57	867	924	0.0617	0.0657
Total	138	1862	2000	$\text{RR} = \dfrac{0.0753}{0.0617} = 1.220$	$\text{OR} = \dfrac{0.0814}{0.0657} = 1.238$

BOX 16.2 PROPERTIES OF ODDS RATIOS

The minimum possible value is zero, and the maximum possible value is infinity.

- An odds ratio of 1 occurs when the odds, and hence the proportions, are the same in the two groups and is equivalent to no association between the exposure and the disease.
- The odds ratio is always further away from 1 than the corresponding risk (or prevalence) ratio. Thus:

$$\text{if RR} > 1 \text{ then OR} > \text{RR}$$

$$\text{if RR} < 1 \text{ then OR} < \text{RR}$$

- For a rare outcome (one in which the probability of the event not occurring is close to 1) the odds ratio is approximately equal to the risk ratio (since the odds are approximately equal to the risk, see Section 14.6).
- The odds ratio for the occurrence of disease is the reciprocal of the odds ratio for non-occurrence.
- The odds ratio for exposure, that is the odds of disease in the exposed compared to the odds in the unexposed group, *equals* the odds ratio for disease, that is the odds of exposure in the disease compared to the odds in the healthy group. (This equivalence is fundamental for the analysis of case- control studies.)

Comparison of odds ratios and risk ratios

As mentioned in Section 16.2, both the risk difference and the risk ratio have immediate intuitive interpretations. It is relatively easy to explain that, for example, moderate smokers have twice the risk of cardiovascular disease than non-smokers (RR = 2). In contrast, interpretation of odds ratios often causes problems; except for gamblers, who tend to be extremely familiar with the meaning of odds (see Chapter 14).

Table 16.6 Values of the risk ratio when the odds ratio = 2, and the odds ratio when the risk ratio = 2, given different values of the risk in the unexposed group.

Odds ratio = 2		Risk ratio = 2	
Risk in the unexposed group	Corresponding risk ratio	Risk in the unexposed group	Corresponding odds ratio
0.001	1.998	0.001	2.002
0.005	1.99	0.005	2.010
0.01	1.980	0.01	2.020
0.05	1.905	0.05	2.111
0.1	1.818	0.1	2.25
0.5	1.333	0.3	3.5
0.9	1.053	0.4	6.0
0.95	1.026	0.45	11.0
0.99	1.005	0.5*	∞

*When π_0 is greater than 0.5, the risk ratio must be less than 2, since $\pi_1 = \text{RR} \times \pi_0$, and probabilities cannot exceed 1.

A common mistake in the literature is to interpret an odds ratio as if it were a risk ratio. For rare outcomes, this is not a problem since the two are numerically equal (see Box 16.2 and Table 16.6). However, for common outcomes, this is not the case; the interpretation of odds ratios diverges from that for risk ratios. Table 16.6 shows values of the risk ratio for an odds ratio of 2, and conversely the values of the odds ratio for a risk ratio of 2, for different values of the risk in the unexposed group. For example, it shows that if the risk in the exposed group is 0.5, then an odds ratio of 2 is equivalent to a risk ratio of 1.33. When the outcome is common, therefore, an odds ratio of (for example) 2 or 5 *must not* be interpreted as meaning that the risk is multiplied by 2 or 5.

As the risk in the unexposed group becomes larger, the maximum possible value of the risk ratio becomes constrained, because the maximum possible value for a risk is 1. For example, if the risk in the unexposed group is 0.33, the maximum possible value of the RR is 3. Because there is no upper limit for the odds, the OR is not constrained in this manner. Note that as the risk in the unexposed group increases the odds ratio becomes much larger than the risk ratio and, as explained above, should no longer be interpreted as the amount by which the risk factor multiplies the risk of the disease outcome.

The constraint on the value of the risk ratio can cause problems for statistical analyses using risk ratios when the outcome is not rare, because it can mean that the risk ratio differs between population strata. For example, in a low-risk stratum the risk of disease might be 0.2 (20%) in the unexposed group and 0.5 (50%) in the exposed group. The risk ratio in that stratum is therefore $0.5/0.2 = 2.5$. If the risk of disease in a high-risk stratum is 0.5 then the risk ratio can be at most 2 in that stratum, since the maximum possible risk of disease is 1, and $1/0.5 = 2$.

A further difficulty with risk ratios is that the interpretation of results may depend on whether the occurrence of an event, or its non-occurrence, is considered as the outcome. For odds ratios this presents no problems, since:

$$\text{OR(disease)} = 1/\text{OR(healthy)}$$

However no such relationship exists for risk ratios. For instance, consider the low-risk stratum in which the risk ratio is $0.5/0.2 = 2.5$. If the non-occurrence of disease (healthy) is considered as the outcome, then the risk ratio is $(1-0.5)/(1-0.2) = 0.5/0.8 = 0.625$. This is *not* the same as $1/2.5 = 0.4$.

Example 16.4

Consider a study in which we monitor the risk of severe nausea during chemotherapy for breast cancer. A new drug is compared with standard treatment. The hypothetical results are shown in Table 16.7.

The risk of severe nausea is 88% in the group treated with the new drug and 71% in the group given standard treatment, so the risk ratio is $0.88/0.71 = 1.239$, an apparently moderate increase in the prevalence of nausea. In contrast the odds ratio is 2.995, a much more dramatic increase. Note, however, that the risk ratio is constrained: it cannot be greater than $1/0.71 = 1.408$.

Suppose now that we consider our outcome to be *absence* of nausea. The risk ratio is $0.12/0.29 = 0.414$: the proportion of patients without severe nausea has more than halved. The odds ratio is 0.334: exactly the inverse of the odds ratio for nausea ($1/2.995 = 0.334$).

Table 16.7 Risk of severe nausea following chemotherapy for breast cancer.

	Number with severe nausea	Number without severe nausea	Total
New drug	88 (88%)	12	100
Standard treatment	71 (71%)	29	100

Rationale for the use of odds ratios

In the recent medical literature, *the statistical analysis of binary outcomes* is almost always based on *odds ratios*, even though they are less easy to interpret than risk ratios (or risk differences). This is for the following three reasons:

1 When the **outcome is rare**, *the odds ratio is the same as the risk ratio*. This is because the *odds* of occurrence of a rare outcome are numerically equivalent to its *risk*. Analyses based on odds ratios therefore give the same results as analyses based on risk ratios.

2 When the **outcome is common**, risk ratios are *constrained* but odds ratios are not. Analyses based on risk ratios, particularly those examining the effects of more than one exposure variable, can cause computational problems and are difficult to interpret. In contrast, these problems do not occur in analyses based on odds ratios.

3 For odds ratios, the conclusions are identical whether we consider our outcome as the occurrence of an event, or the absence of the event.

Taken together, these mean that analyses of binary outcomes controlling for possible confounding (see Chapter 18), or which use regression modelling (see Chapters 19 to 21), usually report exposure effects as odds ratios, regardless of whether the outcome is rare or common.

In addition, odds ratios are the measure of choice in **case–control studies**. In fact, it is in this context that they were first developed and used. In case–control studies we recruit a group of people with the disease of interest (cases) and a random sample of people without the disease (the controls). The distribution of one or more exposures in the cases is then compared with the distribution in the controls. Because the controls usually represent an unknown fraction of the whole population, it is not possible to estimate the risk of disease in a case–control study, and so risk differences and risk ratios cannot be derived. The odds ratio can be used to compare cases and controls because the ratio of the *odds of exposure* (d_1/d_0) among the *diseased* group compared to the odds of exposure among the *healthy* group (h_1/h_0), is equivalent to the ratio of the odds of disease in exposed compared to unexposed:

$$\text{OR} = \frac{d_1/h_1}{d_0/h_0} = \frac{d_1 \times h_0}{d_0 \times h_1} = \frac{d_1/d_0}{h_1/h_0}$$

16.7 ODDS RATIOS: CONFIDENCE INTERVALS AND HYPOTHESIS TESTS

Confidence interval for the odds and the odds ratio

We saw in Section 16.5 how a confidence interval for the risk ratio is derived by calculating a confidence interval for the log risk ratio and then converting this to a confidence interval for the risk ratio. Confidence intervals for the odds, and the odds ratio, are calculated in exactly the same way. The results are shown in Table 16.8. Note that s.e.(log OR) should be interpreted as 's.e. of the log OR'. The formula for s.e.(log OR) is also known as **Woolf's formula**.

Table 16.8 Formulae for calculation of 95% confidence intervals for the odds and the odds ratio.

Odds	Odds ratio (OR)
95% CI = odds/EF to odds × EF, where EF = $\exp[1.96 \times \text{s.e.}(\log \text{odds})]$ and $\text{s.e.}(\log\ \text{odds}) = \sqrt{[1/d + 1/h]}$	95% CI = OR/EF to OR × EF, where EF = $\exp[1.96 \times \text{s.e.}(\log \text{OR})]$ and $\text{s.e.}(\log \text{OR}) = \sqrt{[1/d_1 + 1/h_1 + 1/d_0 + 1/h_0]}$

Example 16.3 (continued)

Consider the data from the asthma survey presented in Table 16.5. The standard error of the log OR is given by:

$$\text{s.e.}(\log \text{OR}) = \sqrt{[1/57 + 1/867 + 1/81 + 1/995]} = 0.179$$

The error factor is given by:

$$\text{EF} = \exp(1.96 \times 0.179) = 1.420$$

The 95% confidence interval for the odds ratio is therefore:

$$95\%\ \text{CI} = 1.238/1.420 \text{ to } 1.238 \times 1.420 = 0.872 \text{ to } 1.759$$

With 95% confidence, the odds ratio in the population lies between 0.872 and 1.759.

Test of the null hypothesis

We use the log OR and its standard error to derive a z statistic and test the null hypothesis in the usual way:

$$z = \frac{\log\ \text{OR}}{\text{s.e.}(\log \text{OR})}$$

The results are identical to those produced by simple logistic regression models (see Chapter 19).

Example 16.3 (continued)

The z statistic is given by $z = 0.214/0.179 = 1.194$. This corresponds to a P-value of 0.232. There is no clear evidence against the null hypothesis that the OR = 1, i.e. that the prevalence of asthma is the same in men and women.

CHAPTER 17

Chi-squared tests for 2 × 2 and larger contingency tables

17.1 INTRODUCTION

We saw in the last chapter that when both exposure and outcome variables have only two possible values (binary variables) the data can be displayed in a **2×2 table**. As described in Section 3.4, **contingency tables** can also be used to display the association between two categorical variables, one or both of which has more than two possible values. The categories for one variable define the rows, and the categories for the other variable define the columns. Individuals are assigned to the appropriate cell of the contingency table according to their values for the two variables. A contingency table is also used for discrete numerical variables, or for continuous numerical variables whose values have been grouped. These larger tables are generally called ***r* × *c* tables**, where *r* denotes the number of rows in the table and *c* the number of columns. If the variables displayed are an *exposure* and an *outcome*, then it is usual to arrange the table with exposure as the row variable and outcome as the column variable, and to display percentages corresponding to the *exposure* variable.

In this chapter, we describe how to use a **chi-squared (χ^2) test** to examine whether there is an association between the row variable and the column variable or, in other words, whether the distribution of individuals among the categories of one variable is independent of their distribution among the categories of the other. We explain this for 2 × 2 tables, and for larger *r* × *c* tables. When the table has only two rows and two columns the χ^2 test is equivalent to the *z*-test for the difference between two proportions. We also describe the **exact test** for a 2 × 2 table when the sample size is too small for the *z*-test or the χ^2 test to be valid. Finally, we describe the use of a **χ^2 test for trend**, for the special case where we have a binary outcome variable and several exposure categories, which have a natural order.

17.2 CHI-SQUARED TEST FOR A 2×2 TABLE

Example 17.1

Table 17.1 shows the data from the influenza vaccination trial described in the last chapter (see Example 16.1). Since the exposure is vaccination (the row variable), the table includes row percentages. We now wish to assess the strength of the evidence that vaccination affected the probability of contracting influenza.

Table 17.1 2 × 2 table showing results from an influenza vaccine trial.

(a) Observed numbers.

	Influenza		
	Yes	No	Total
Vaccine	20 (8.3%)	220 (91.7%)	240
Placebo	80 (36.4%)	140 (63.6%)	220
Total	100 (21.7%)	360 (78.3%)	460

(b) Expected numbers.

	Influenza		
	Yes	No	Total
Vaccine	52.2	187.8	240
Placebo	47.8	172.2	220
Total	100	360	460

The **chi-squared test** compares the observed numbers in each of the four categories in the contingency table with the numbers to be expected if there were no difference in efficacy between the vaccine and placebo. Overall 100/460 people contracted influenza and, if the vaccine and the placebo were equally effective, one would expect this same proportion in each of the two groups; that is $100/460 \times 240 = 52.2$ in the vaccine group and $100/460 \times 220 = 47.8$ in the placebo group would have contracted influenza. Similarly $360/460 \times 240 = 187.8$ and $360/460 \times 220 = 172.2$ would have escaped influenza. These expected numbers are shown in Table 17.1(b). They add up to the same row and column totals as the observed numbers. The chi-squared value is obtained by calculating

$$(\text{observed} - \text{expected})^2/\text{expected}$$

for each of the four cells in the contingency table and then summing them.

$$\chi^2 = \Sigma \frac{(O - E)^2}{E}, \text{d.f.} = 1 \text{ for a } 2 \times 2 \text{ table}$$

This is exactly the same formula as was given for the chi-squared goodness of fit test, which was described in Chapter 12. The greater the differences between the observed and expected numbers, the larger the value of χ^2. The percentage points of the chi-squared distribution are given in Table A5 in the Appendix. The values depend on the degrees of freedom, which equal 1 for a 2×2 table (the number of rows minus 1 multiplied by the number of columns minus 1). In this example:

$$\begin{aligned}\chi^2 &= \frac{(20-52.2)^2}{52.2}+\frac{(80-47.8)^2}{47.8}+\frac{(220-187.8)^2}{187.8}+\frac{(140-172.2)^2}{172.2}\\ &= 19.86+21.69+5.52+6.02=53.09\end{aligned}$$

53.09 is greater than 10.83, the 0.1% point for the chi-squared distribution with 1 degree of freedom so that the P-value for the test is < 0.001. This means that the probability is less than 0.001, or 0.1%, that such a large observed difference in the percentages contracting influenza could have arisen by chance, if there was no real difference between the vaccine and the placebo. Thus there is strong evidence against the null hypothesis of no effect of the vaccine on the probability of contracting influenza. It is therefore concluded that the vaccine is effective.

Quick formula

Using our standard notation for a 2×2 table (see Table 16.1), a quicker formula for calculating chi-squared on a 2×2 table is:

$$\chi^2 = \frac{n(d_1h_0 - d_0h_1)^2}{dhn_1n_0}, \text{d.f.} = 1$$

In the example,

$$\chi^2 = \frac{460 \times (20 \times 140 - 80 \times 220)^2}{100 \times 360 \times 240 \times 220} = 53.01$$

which, apart from rounding error, is the same as the value of 53.09 obtained above.

Relation with normal test for the difference between two proportions

The square of the z statistic (normal test) for the difference between two proportions and the chi-squared statistic for a 2×2 contingency table are in fact mathematically equivalent ($\chi^2 = z^2$), and the P-values from the two tests are identical. In Example 16.1 (Section 16.3) the z-test gave a value of -7.281 for the influenza vaccine data; $z^2 = (-7.281)^2 = 53.01$ which, apart from rounding error, is the same as the χ^2 value of 53.09 calculated above.

We will show below that the chi-squared test can be extended to larger contingency tables. Note that the percentage points given in Table A5 for a chi-squared distribution with 1 degree of freedom correspond to the two-sided percentage points presented in Table A2 for the standard normal distribution (see Appendix). (The concepts of one- and two-sided tests do not extend to chi-squared tests with larger degrees of freedom as these contain multiple comparisons.)

Continuity correction

The chi-squared test for a 2 × 2 table can be improved by using a continuity correction, often called **Yates' continuity correction**. The formula becomes:

$$\chi^2 = \Sigma \frac{(|O - E| - 0.5)^2}{E}, \text{d.f.} = 1$$

resulting in a smaller value for χ^2. $|O - E|$ means the absolute value of $O - E$ or, in other words, the value of $O - E$ ignoring its sign.

In the example the value for χ^2 becomes:

$$\begin{aligned}\chi^2 &= \frac{(32.2 - 0.5)^2}{52.2} + \frac{(32.2 - 0.5)^2}{47.8} + \frac{(32.2 - 0.5)^2}{187.8} + \frac{(32.2 - 0.5)^2}{172.2} \\ &= 19.25 + 21.02 + 5.35 + 5.84 = 51.46, P < 0.001\end{aligned}$$

compared to the uncorrected value of 53.09.

The rationale of the continuity correction is explained in Figure 15.3, where the normal and binomial distributions are superimposed. It makes little difference unless the total sample size is less than 40, or the expected numbers are small. However there is no analogue of the continuity correction for the Mantel–Haenszel and regression analyses described later in this part of the book. When the expected numbers are *very* small, then the exact test described in Section 17.3 should be used; see discussion on *validity* below.

Validity

When the expected numbers are very small the chi-squared test (and the equivalent *z*-test) is not a good enough approximation and the alternative exact test for a 2 × 2 table should be used (see Section 17.3). Cochran (1954) recommended the use of the **exact test** when:

1 the overall total of the table is less than 20, *or*

2 the overall total is between 20 and 40 and the smallest of the four *expected* numbers is less than 5.

Thus the chi-squared test is *valid* when the overall total is more than 40, regardless of the expected values, and when the overall total is between 20 and 40 provided all the expected values are at least 5.

17.3 EXACT TEST FOR 2×2 TABLES

The exact test to compare two proportions is needed when the numbers in the 2×2 table are very small; see the discussions concerning the validity of the z-test to compare two proportions (Section 16.3) and of the chi-squared test for a 2×2 table (Section 17.2 above). It is most easily described in the context of a particular example.

Example 17.2

Table 17.2 shows the results from a study to compare two treatment regimes for controlling bleeding in haemophiliacs undergoing surgery. Only one (8%) of the 13 haemophiliacs given treatment regime A suffered bleeding complications, compared to three (25%) of the 12 given regime B. These numbers are too small for the chi-squared test to be valid; the overall total, 25, is less than 40, and the smallest expected value, 1.9 (complications with regime B), is less than 5. The exact test is therefore indicated.

Table 17.2 Comparison of two treatment regimes for controlling bleeding in haemophiliacs undergoing surgery.

	Bleeding complications		
Treatment regime	Yes	No	Total
A (group 1)	1 (d_1)	12 (h_1)	13 (n_1)
B (group 0)	3 (d_0)	9 (h_0)	12 (n_0)
Total	4 (d)	21 (h)	25 (n)

The exact test is based on calculating the exact probabilities of the observed table and of more 'extreme' tables with the same row and column totals, using the following formula:

$$\text{Exact probability of } 2 \times 2 \text{ table} = \frac{d!h!n_1!n_0!}{n!d_1!d_0!h_1!h_0!}$$

where the notation is the same as that defined in Table 16.1. The exclamation mark denotes the *factorial* of the number and means all the integers from the number down to 1 multiplied together. (0! is defined to equal 1.) Many calculators have a key for factorial, although this expression may be easily computed by cancelling factors in the top and bottom. The exact probability of Table 17.2 is therefore:

$$\frac{4!21!13!12!}{25!1!3!12!9!} = \frac{4 \times 13 \times 12 \times 11 \times 10}{25 \times 24 \times 23 \times 22} = 0.2261$$

(21! being cancelled into 25!, for example, leaving 25 × 24 × 23 × 22).

In order to test the null hypothesis that there is no difference between the treatment regimes, we need to calculate not only the probability of the observed table but also the probability that a more extreme table could occur by chance. Altogether there are five possible tables that have the same row and column totals as the data. These are shown in Table 17.3 together with their probabilities, which total 1. The observed case is Table 17.3(b) with a probability of 0.2261.

Table 17.3 All possible tables with the same row and column totals as Table 17.2, together with their probabilities.

(a)			Total
	0	13	13
	4	8	12
Total	4	21	25

$P = 0.0391$

(b)			Total
	1	12	13
	3	9	12
Total	4	21	25

$P = 0.2261$

(c)			Total
	2	11	13
	2	10	12
Total	4	21	25

$P = 0.4070$

(d)			Total
	3	10	13
	1	11	12
Total	4	21	25

$P = 0.2713$

(e)			Total
	4	9	13
	0	12	12
Total	4	21	25

$P = 0.0565$

There are two approaches to calculating the *P*-value. In the first approach, more extreme is defined as less probable; more extreme tables are therefore 17.3(a) and 17.3(e) with probabilities 0.0391 and 0.0565 respectively. The total probability needed for the *P*-value is therefore $0.2261 + 0.0391 + 0.0565 = 0.3217$, and so there is clearly no evidence against the null hypothesis of no difference between the regimes.

P-value (approach I) = probability of observed table + probability of less probable tables
P-value (approach II) = 2 × (probability of observed table + probability of more extreme tables in the same direction)

The alternative approach is to restrict the calculation to extreme tables showing differences in the same direction as that observed, and then to double the resulting probability in order to cover differences in the other direction. In this example, the *P*-value thus obtained would be twice the sum of the probabilities of Tables 17.3(a) and 17.3(b), namely $2 \times (0.0391 + 0.2261) = 0.5304$. Neither method is clearly superior to the other, but the second method is simpler to carry out. Although the two approaches give different results, the choice is unlikely, in practice, to affect the assessment of whether the observed difference is due to chance or to a real effect.

17.4 LARGER CONTINGENCY TABLES

So far, we have dealt with 2×2 tables, which are used to display data classified according to the values of two binary variables. The chi-squared test can also be applied to larger tables, generally called ***r* × *c* tables**, where *r* denotes the number of rows in the table and *c* the number of columns.

$$\chi^2 = \Sigma \frac{(O - E)^2}{E}, \text{d.f.} = (r - 1) \times (c - 1)$$

There is no continuity correction or exact test for contingency tables larger than 2×2. Cochran (1954) recommends that the approximation of the chi-squared test is valid provided that less than 20% of the expected numbers are under 5 and none is less than 1. This restriction can sometimes be overcome by combining rows (or columns) with low expected numbers, providing that these combinations make biological sense.

There is no quick formula for a general $r \times c$ table. The expected numbers must be computed for each cell. The reasoning employed is the same as that described above for the 2×2 table. The general rule for calculating an expected number is:

$$E = \frac{\text{column total} \times \text{row total}}{\text{overall total}}$$

It is worth pointing out that the chi-squared test is only valid if applied to the actual numbers in the various categories. It must never be applied to tables showing just proportions or percentages.

Example 17.3

Table 17.4(a) shows the results from a survey to compare the principal water sources in three villages in West Africa. These data were also presented when we introduced cross-tabulations in Chapter 3. The numbers of households using a river, a pond, or a spring are given. We will treat the water source as outcome and village as exposure, so column percentages are displayed. For example, in village A, 40.0% of households use mainly a river, 36.0% a pond and 24.0% a spring. Overall, 70 of the 150 households use a river. If there were no difference between villages one would expect this same proportion of river usage in each village. Thus the expected numbers of households using a river in villages A, B and C, respectively, are:

$$\frac{70}{150} \times 50 = 23.3, \quad \frac{70}{150} \times 60 = 28.0 \quad \text{and} \quad \frac{70}{150} \times 40 = 18.7$$

The expected numbers can also be found by applying the general rule. For example, the expected number of households in village B using a river is:

$$\frac{\text{row total (B)} \times \text{column total (river)}}{\text{overall total}} = \frac{60 \times 70}{150} = 28.0$$

The expected numbers for the whole table are given in Table 17.4(b).

Table 17.4 Comparison of principal sources of water used by households in three villages in West Africa.

(a) Observed numbers.

	Water source			
Village	River	Pond	Spring	Total
A	20 (40.0%)	18 (36.0%)	12 (24.0%)	50 (100.0%)
B	32 (53.3%)	20 (33.3%)	8 (13.3%)	60 (100.0%)
C	18 (45.0%)	12 (30.0%)	10 (25.0%)	40 (100.0%)
Total	70 (46.7%)	50 (33.3%)	30 (20.0%)	150 (100.0%)

(b) Expected numbers.

	Water source			
Village	River	Pond	Spring	Total
A	23.3	16.7	10.0	50
B	28.0	20.0	12.0	60
C	18.7	13.3	8.0	40
Total	70	50	30	150

$$\begin{aligned}
\chi^2 &= \Sigma \frac{(O-E)^2}{E} \\
&= (20-23.3)^2/23.3 + (18-16.7)^2/16.7 + (12-10.0)^2/10.0 + \\
&\quad (32-28.0)^2/\ 28.0 + (18-18.7)^2/18.7 + (20-20.0)^2/20.0 + \\
&\quad (8-12.0)^2/12.0 + (12-13.3)^2/13.3 + (10-8.0)^2/8.0 \\
&= 3.53 \\
\text{d.f.} &= (r-1)\times(c-1) = 2\times 2 = 4
\end{aligned}$$

The corresponding P-value (derived using a computer) is 0.47, so we can conclude that there is no evidence of a difference between the villages in the proportion of households using different water sources. Alternatively, we can see from the fourth row of Table A5 (see Appendix) that since 3.53 lies between 3.36 and 5.39, the P-value lies between 0.25 and 0.5.

17.5 ORDERED EXPOSURES: χ^2 TEST FOR TREND

We now consider the special case where we have a binary outcome variable and several exposure categories, which have a natural order. The standard chi-squared test for such data is a general test to assess whether there are differences among the proportions in the different exposure groups. The **χ^2 test for trend**, described now, is a more sensitive test that assesses whether there is an increasing (or decreasing) trend in the proportions over the exposure categories.

Example 17.4

Table 17.5 shows data from a study that examined the association between obesity and age at menarche in women. The outcome was whether the woman was aged < 12 years at menarche (event D) or aged $> 12+$ years (event H). The exposure, obesity, is represented by triceps skinfold, categorised into three groups. Although it is conventional that the exposure variable is the row variable, this is not an absolute rule. For convenience, we have not followed this convention, and have

Table 17.5. Relationship between triceps skinfold and early menarche. Data from a study on obesity in women (Beckles *et al*. (1985) *International Journal of Obesity* **9**: 127–35).

	Triceps skinfold group			
Age at menarche	Small	Intermediate	Large	Total
< 12 years (D)	15 (8.8%)	29 (12.8%)	36 (19.4%)	80
12+ years (H)	156 (91.2%)	197 (87.2%)	150 (80.6%)	503
Total	171 (100%)	226 (100%)	186 (100%)	583
Exposure group score (*x*)	0	1	2	
Odds of early menarche	0.10 (0.06 to 0.16)	0.15 (0.10 to 0.22)	0.24 (0.17 to 0.35)	
Log odds	−2.34 (−2.87 to −1.81)	−1.92 (−2.31 to −1.53)	−1.43 (−1.79 to −1.06)	

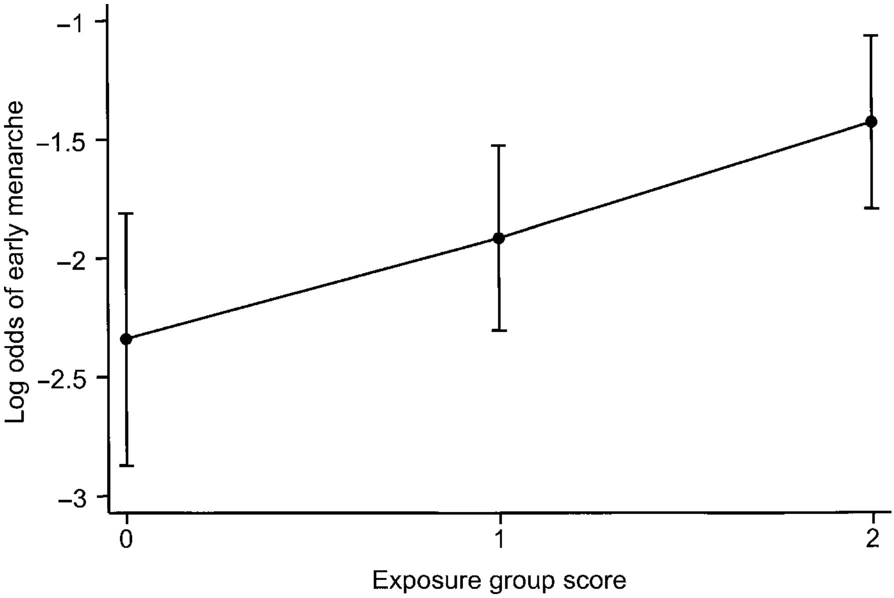

Fig. 17.1 Log odds of early menarche according to skinfold thickness group.

presented the exposure in the columns and the outcome in the rows. It can be seen that the proportion of women who had experienced early menarche increased with triceps skinfold size. This can be examined using the **χ^2 test for trend**.

The first step is to assign scores to the exposure groups. The usual choice is simply to number the columns 0, 1, 2, etc., as shown here (or equivalently 1, 2, 3, etc.). This is equivalent to assuming that the *log odds* goes up (or down) by equal amounts between the exposure groups, or in other words that there is a linear relationship between the two. The odds and log odds of early menarche are shown below the exposure scores, and the log odds with 95% confidence intervals are plotted in Figure 17.1. It is clear that the assumption of a linear increase in log odds, with exposure group is reasonable. The difference in log odds is $(-1.92 - -2.34) = 0.42$ between groups 1 and 0, and $(-1.43 - -1.92) = 0.49$ between groups 2 and 1.

Another possibility would have been to use the means or medians of the triceps skinfold measurements in each group. The assumption here would be a linear relationship between log odds and triceps skinfold measurement. The two approaches will give similar results if the differences between the means (or medians) are similar between the triceps skinfold groups.

The next step is to calculate three quantities for *each exposure group* in the table and to sum the results of each. These are:

1 dx, the product of the *observed* number, d, with outcome D, and the exposure group score, x;

2 nx, the product of the total, n, in the exposure group and its score, x; and

3 nx^2, the product of the total, n, in the exposure group and the square of its score, x^2.

Using N to denote the overall total and O the total observed number of events (the total of the top row), we then calculate:

$$U = \Sigma(dx) - \frac{O}{N}\Sigma(nx) \quad \text{and} \quad V = \frac{O(N-O)}{N^2(N-1)}[N\Sigma(nx^2) - (\Sigma nx)^2]$$

The increase in log odds ratio per group is estimated by U/V, with standard error $\sqrt{(1/V)}$. The formula for the chi-squared statistic is:

$$\chi^2_{\text{trend}} = \frac{U^2}{V}, \text{d.f.} = 1$$

This tests the null hypothesis that the linear increase in log odds per exposure group is zero.

There are various different forms for this test, most of which are algebraically equivalent. The only difference is that in some forms $(N - 1)$ is replaced by N in the calculation of V. This difference is unimportant.

Example 17.4 (continued)

The calculations for the data presented in Table 17.5 are as follows:

$$\begin{aligned}
\Sigma(dx) &= 15\times 0 + 29\times 1 + 36\times 2 = 101\\
\Sigma(nx) &= 171\times 0 + 226\times 1 + 186\times 2 = 598\\
\Sigma(nx^2) &= 171\times 0 + 226\times 1 + 186\times 4 = 970\\
O &= 80,\ N = 583,\ N - O = 503\\
U &= 101 - \left(\frac{80}{583}\times 598\right) = 18.9417\\
V &= \left(\frac{80\times 503}{583^2\times 582}\right)\times(583\times 970 - 598^2) = 42.2927
\end{aligned}$$

The increase in log odds ratio per group is $U/V = 0.445$: approximately an average of the differences between groups 1 and 0, and 2 and 1 (see above). Its standard error is $\sqrt{(1/V)} = 0.154$ and the 95% CI (derived in the usual way) is 0.146 to 0.749. This converts to an *odds ratio per exposure group* of 1.565 (95% CI 1.158 to 2.115). The chi-squared statistic is:

$$\chi^2_{\text{trend}} = \frac{(18.9417)^2}{42.2927} = 8.483, \quad \text{d.f.} = 1, \quad P = 0.0036.$$

There is therefore strong evidence that the odds of early menarche increased with increasing triceps skinfold.

This is a simple example of a **dose–response model** for the association between an exposure and a binary outcome. We show in Chapter 19 that a logistic regression model for this association gives very similar results. Note that the difference between the standard χ^2 value and the trend test χ^2 value provides a chi-squared value with $(c - 2)$ degrees of freedom to test for **departures from linear trend**, where c is the number of exposure groups. Such tests are described in more detail, in the context of regression modelling, in Section 29.6.

CHAPTER 18

Controlling for confounding: stratification

18.1 INTRODUCTION

Previous chapters in this part of the book have presented methods to examine the association between a binary outcome and two or more exposure (or treatment) groups. We have used confidence intervals and P-values to assess the likely size of the association, and the evidence that it represents a real difference in disease risk between the exposure groups. However, before attributing any difference in outcome between the exposure groups to the exposure itself, it is important to examine whether the exposure–outcome association has been affected by other factors that differ between the exposure groups and which also affect the outcome. Such factors are said to *confound* the association of interest. Failure to control for them can lead to **confounding bias**. This fundamental problem is illustrated by an example in the next section.

In this chapter, we describe the Mantel–Haenszel method that uses stratification to *control for confounding* when both the exposure and outcome are *binary* variables. In Chapter 11, on multiple regression for the analysis of numerical outcomes, we briefly described how regression models can be used to control for confounding. We will explain this in much more detail in Chapter 20 in the context of *logistic regression* for the analysis of binary outcomes.

18.2 CONFOUNDING

Example 18.1

Table 18.1 shows hypothetical results from a survey carried out to compare the prevalence of antibodies to leptospirosis in rural and urban areas of the West Indies, with rural residence as the exposure of interest.

Table 18.1 Results of a survey of the prevalence of leptospirosis in rural and urban areas of the West Indies.

	Leptospirosis antibodies			
Type of area	Yes	No	Total	Odds
Rural	60 (30%)	140 (70%)	200	0.429
Urban	60 (30%)	140 (70%)	200	0.429
Total	120	280	400	

Since the numbers of individuals with and without antibodies are identical in urban and rural areas, the odds ratio is exactly 1 and we would conclude that there is no association between leptospirosis antibodies and urban/rural residence. However, Table 18.2 shows that when the same sample is subdivided according to gender, the risk of having antibodies is higher in rural areas *for both males and females*. The disappearance of this effect when the genders are combined is caused by a combination of two factors:

1. Females in both areas are much less likely than males to have antibodies.
2. The samples from the rural and urban areas have different gender compositions. The proportion of males is 100/200 (50%) in the urban sample but only 50/200 (25%) in the rural sample.

Table 18.2 Association between antibodies to leptospirosis (the outcome variable) and rural/urban residence (the exposure variable), separately in males and females.

(a) Males.

	Antibodies			
Type of area	Yes	No	Total	Odds
Rural	36 (72%)	14 (28%)	50	2.57
Urban	50 (50%)	50 (50%)	100	1.00
Total	86	64	150	

OR $= 2.57/1 = 2.57$ (95% CI $= 1.21$ to 5.45), $P = 0.011$

(b) Females.

	Antibodies			
Type of area	Yes	No	Total	Odds
Rural	24 (16%)	126 (84%)	150	0.19
Urban	10 (10%)	90 (90%)	100	0.11
Total	34	216	250	

OR $= 0.19/0.11 = 1.71$ (95% CI $= 0.778$ to 3.78), $P = 0.176$

Gender is said to be a **confounding** variable because it is related both to the outcome variable (presence of antibodies) and to the exposure groups being compared (rural and urban). Ignoring gender in the analysis leads to a **bias** in the results. Analysing males and females separately provides evidence of a difference between the rural and urban areas for males but not for females (Table 18.2). However, we would like to be able to combine the information in the two tables to estimate the association between leptospirosis antibodies and urban/rural residence, *having allowed for* the association of each of these with gender. We describe how to do this in the next section.

In general confounding occurs when a confounding variable, C, is associated with the exposure, E, and also influences the disease outcome, D. This is illustrated in Figure 18.1. We are interested in the E–D association, but the E–C and C–D associations may bias our estimate of the E–D association unless we take them into account in our analysis.

In our example, failure to allow for gender masked an association with urban/rural residence. In other situations similar effects could suggest a difference or association where none exists, or could even suggest a difference the opposite way around to one that does exist. For example, in the assessment of whether persons suffering from schistosomiasis have a higher mortality rate than uninfected persons, it would be important to take age into account since both the risk of dying and the risk of having schistosomiasis increase with age. If age were not allowed for, schistosomiasis would appear to be associated with increased mortality, even if it were not, as those with schistosomiasis would be on average older and therefore more likely to die than younger uninfected persons.

Note that *a variable that is part of the causal chain leading from E to D is not a confounder*. That is, if E affects C, which in turn affects D, then we should not adjust for the effect of C in our analysis of the E–D association (unless we wish to estimate the effect of E on D which is not caused by the E–C association). For example, even though smoking during pregnancy is related both to socio-economic status and the risk of having a low birth-weight baby, it would be incorrect to control for it when examining socio-economic differences in the risk of low birth

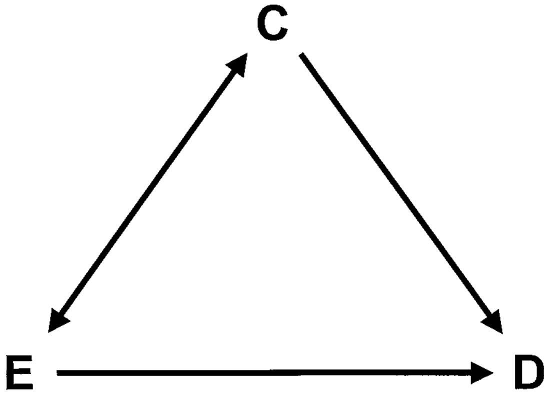

Fig. 18.1 Situation in which C may confound the affect of the E–D association.

weight, since it is on the causal path. Controlling for it in the analysis would lead to an underestimate of any socio-economic differences in risk. These issues are discussed in more detail in Section 38.5.

Note that in clinical trials (and other experimental studies), **randomization** is used to allocate individuals to the different treatment groups (see Chapter 34). Provided that such trials are large enough to ensure that chance differences between the groups are small, the problem of confounding is thus avoided, because the treatment and control groups will be similar in all respects other than those under trial.

18.3 STRATIFICATION TO CONTROL FOR CONFOUNDING

One way to solve the problem of confounding in the analysis is to restrict comparisons to individuals who have the same value of the confounding variable C. Among such individuals associations with C cannot bias the E–D association, because there is no variation in C. Thus in Example 18.1 above, the association between leptospirosis antibodies and urban/rural residence was examined separately for males and females. The subsets defined by the levels of C are called **strata**, and so this process is known as **stratification**. It leads to separate estimates of the odds ratio for the E–D association in each stratum. There is no reason why C should be a binary variable: for example we might allow for the confounding effects of age by splitting a sample of adults aged 15 to 50 years into seven five-year age groups.

Unless it appears that the association between the exposure and outcome varies markedly between the strata (see Section 18.5), we will usually wish to combine the evidence from the separate strata and summarize the association, **controlling** for the confounding effect of C. The simplest approach would be to calculate an average of the estimates of the odds ratios of the E–D association from the different strata. However, we know that, in general, strata in which there are more individuals will tend to have a more precise estimate of the association (i.e. one with a smaller standard error) than strata in which there are fewer individuals. We therefore calculate a **weighted average**, in which greater weight is given to the strata with more data.

$$\text{Weighted average OR} = \frac{\Sigma(w_i \times \text{OR}_i)}{\Sigma w_i}$$

where OR_i is the odds ratio in stratum i, and w_i is the weight it is given in the calculation of the weighted average odds ratio. This is also known as the **summary odds ratio**. Note that in a weighted average, the weights (w_i) are always positive numbers. The larger the value of w_i, the more OR_i influences the weighted average OR. Also note that if all the weights were equal to 1, then the weighted average OR would be equal to the mean OR.

The most widely used weighting scheme is that proposed by Mantel and Haenszel, as described in the next section.

18.4 MANTEL–HAENSZEL METHOD FOR 2 × 2 TABLES

Mantel–Haenszel methods can be used to combine the evidence from the separate strata, and summarize the association, **controlling** for the confounding effect of C. We will describe their use when *both* the outcome and exposure are *binary* variables. In this case, the stratified data will consist of c separate 2 × 2 tables, where c is the number of different values the confounding variable can take. Table 18.3 shows the notation we will use for the 2 × 2 table in stratum i. It is exactly the same as that in Table 16.1 for a single 2 × 2 table, but with the subscript i added, to refer to the stratum i. The estimate of the odds ratio for stratum i is:

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

In Table 18.2, gender is the confounding variable; $c = 2$, and we have two tables of the association between rural/urban residence and presence of leptospirosis antibodies, one for males and one for females.

Table 18.3 Notation for the 2 × 2 table in stratum i.

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

Mantel–Haenszel estimate of the odds ratio controlled for confounding

The Mantel–Haenszel estimate of the summary odds ratio, which we shall denote as OR_{MH}, is a weighted average of the odds ratios from the separate strata, with weights:

$$w_i = \frac{d_{0i} \times h_{1i}}{n_i}$$

Since the numerator of the weight is the same as the denominator of the odds ratio (OR_i) in stratum i, $w_i \times \mathrm{OR}_i = (d_{1i} \times h_{0i})/n_i$. Using these weights therefore leads to the following formula for the **Mantel–Haenszel estimate** of the odds ratio:

$$\mathrm{OR}_{MH} = \frac{\Sigma(w_i \times \mathrm{OR}_i)}{\Sigma w_i} = \frac{\Sigma \dfrac{d_{1i} \times h_{0i}}{n_i}}{\Sigma \dfrac{d_{0i} \times h_{1i}}{n_i}}$$

Following the notation of Clayton and Hills (1993), this can alternatively be written as:

$$\mathrm{OR}_{MH} = Q/R, \text{where}$$
$$Q = \Sigma \frac{d_{1i} \times h_{0i}}{n_i} \text{ and } R = \Sigma \frac{d_{0i} \times h_{1i}}{n_i}$$

Example 18.1 (continued)

Table 18.4 shows the results of the calculations required to derive the Mantel–Haenszel odds ratio combining the data presented separately for males and females in Table 18.2 on the association between antibodies to leptospirosis (the outcome variable) and rural/urban residence (the exposure variable). This Mantel–Haenszel estimate of the odds ratio controlling for gender equals:

$$\mathrm{OR}_{MH} = \frac{Q}{R} = \frac{20.64}{9.71} = 2.13$$

After controlling for the confounding effect of gender, the odds of leptospirosis antibodies are more than doubled in rural compared to urban areas. The summary OR (2.13) is, as expected, in between the odds ratios from the two strata, but is marginally closer to the OR for females (1.71) than it is to the OR for males (2.57). This is because the weight allocated to the estimate for females (5.04) is a little higher than that for males (4.67).

Table 18.4 Calculations required for deriving the Mantel–Haenszel OR, with associated confidence interval and *P*-value.

Stratum i	OR_i	$w_i = \dfrac{d_{0i} \times h_{1i}}{n_i}$	$w_i\mathrm{OR}_i = \dfrac{d_{1i} \times h_{0i}}{n_i}$	V_i	d_{1i}	E_{1i}
Males (i = 1)	2.57	$\dfrac{50 \times 14}{150} = 4.67$	12.00	8.21	36	28.67
Females (i = 2)	1.71	$\dfrac{10 \times 126}{250} = 5.04$	8.64	7.08	24	20.40
Total		$R = 9.71$	$Q = 20.64$	$V = 15.29$	O=60	$E = 49.07$

Standard error and confidence interval of the Mantel–Haenszel OR

The 95% confidence interval for OR_{MH} is derived using the standard error of $\log \text{OR}_{MH}$, denoted by s.e._{MH}, in exactly the same way as that for a single odds ratio (see Section 16.7):

$$95\%\ \text{CI} = \text{OR}_{MH}/\text{EF to } \text{OR}_{MH} \times \text{EF},$$
$$\text{where the error factor EF} = \exp(1.96 \times \text{s.e.}_{MH})$$

The simplest formula for the **standard error of log OR_{MH}** (Clayton and Hills 1993) is:

$$\text{s.e.}_{MH} = \surd[V/(Q \times R)],$$
$$Q = \Sigma \frac{d_{1i} \times h_{0i}}{n_i},\ \ R = \Sigma \frac{d_{0i} \times h_{1i}}{n_i},\ \ V = \Sigma V_i = \Sigma \frac{d_i \times h_i \times n_{0i} \times n_{1i}}{n_i^2 \times (n_i - 1)}$$

V is the sum across the strata of the variances V_i for the number of exposed individuals experiencing the outcome event, i.e. the variances of the d_{1i}'s. Note that the formula for the variance V_i of d_{1i} for stratum i is based solely on the marginal totals of the table. It therefore gives the same value for each of the four cells in the table, implying they have equal variances. This is the case because once we know one cell value, we can deduce the others from the appropriate marginal totals.

Example 18.1 (continued)

Using the results of the calculations for Q, R and V shown in Table 18.4, we find that:

$$\text{s.e.}_{MH} = \surd[V/(Q \times R)] = \surd[15.287/(20.640 \times 9.71)] = 0.276$$

so that $\text{EF} = \exp(1.96 \times 0.276) = 1.72$, $\text{OR}_{MH}/\text{EF} = 2.13/1.72 = 1.24$ and $\text{OR}_{\text{MH}} \times \text{EF} = 2.13 \times 1.72 = 3.65$. The 95% CI is therefore:

$$95\%\ \text{CI for } \text{OR}_{MH} = 1.24 \text{ to } 3.65$$

With 95% confidence, the odds of leptospirosis is between 1.24 and 3.65 times higher in rural than urban areas, having controlled for the confounding effect of gender.

Mantel–Haenszel χ^2 test

Finally, we test the null hypothesis that $\text{OR}_{MH} = 1$ by calculating the **Mantel–Haenszel χ^2 test statistic**:

$$\chi^2_{MH} = \frac{(\Sigma d_{1i} - \Sigma E_{1i})^2}{\Sigma V_i} = \frac{(O-E)^2}{V} = \frac{U^2}{V}\text{; d.f.} = 1$$

This is based on a comparison in each stratum of the number of exposed individuals *observed* to have experienced the event (d_{1i}), with the *expected* number in this category (E_{1i}) if there were no difference in the risks between exposed and unexposed. The expected numbers are calculated in exactly the same way as that described for the standard χ^2 test in Chapter 17:

$$E_{1i} = \frac{d_i \times n_{1i}}{n_i}$$

The formula has been simplified by writing O for the sum of the observed numbers, E for the sum of the expected numbers and U for the difference between them:

$$O = \Sigma d_{1i},\ E = \Sigma E_{1i} \text{ and } U = O - E$$

Note that χ^2_{MH} *has just 1 degree of freedom irrespective of how many strata are summarized.*

Example 18.1 (continued)

The calculations for the data presented in Table 18.2 are laid out in Table 18.4. A total $O = 60$ persons in rural areas had antibodies to leptospirosis compared with an expected total of $E = 49.07$, based on assuming no difference in prevalence between rural and urban areas. Thus the Mantel–Haenszel χ^2 statistic is:

$$\chi^2_{MH} = \frac{U^2}{V} = \frac{(60-49.07)^2}{15.29} = 7.82,\ \text{d.f.} = 1,\ P = 0.0052$$

After controlling for gender, there is good evidence of an increase in the prevalence of antibodies to leptospirosis among those living in rural compared to urban areas.

It may seem strange that this test appears to be based entirely on the observed and expected values of d_{1i} and not also on the other cells in the tables. This is not really the case, however, since once the value of d_{1i} is known the values of h_{1i}, d_{0i} and h_{0i} can be calculated from the totals of the table. If the Mantel–Haenszel test is applied to a single 2 × 2 table, the χ^2 value obtained is close to, but not exactly equal to, the standard χ^2 value. It is slightly smaller, equalling $(n-1)/n$ times the standard value. This difference is negligible for values of n of 20 or more, as required for the validity of the chi-squared test.

Validity of Mantel–Haenszel methods

The Mantel–Haenszel estimate of the odds ratio is valid even for small sample sizes. However, the formula that we have given for the standard error of log OR_{MH} will be inaccurate if the overall sample size is small. A more accurate estimate, which is more complicated to calculate, was given by Robins *et al.* (1986).

The **validity** of the Mantel–Haenszel χ^2 test can be assessed by the following 'rule of 5'. Two additional values are calculated for each table and summed over the strata. These are:

1 $\min(d_i,\ n_{1i})$, that is the smaller of d_i and n_{1i}, and

2 $\max(0,\ n_{1i}-h_i)$, which equals 0 if n_{1i} is smaller than or equal to h_i, and $(n_{1i}-h_i)$, if n_{1i} is larger than h_i.

Both sums must differ from the total of the expected values, E, by at least 5 for the test to be valid. The details of these calculations for the leptospirosis data are shown in Table 18.5. The two sums, 84 and 0, both differ from 70.933 by 5 or more, validating the use of the Mantel–Haenszel χ^2 test.

Table 18.5 Rule of 5, to check validity.

Stratum i	$\text{Min}(d_i,\ n_{1i})$,	$\text{Max}(0,\ n_{1i}-h_i)$	E_i
Males ($i=1$)	Min(86, 50) = 50	Max(0, −14) = 0	57.333
Females ($i=2$)	Min(34, 150) = 34	Max(0, −116) = 0	13.600
Total	84	0	70.933

18.5 EFFECT MODIFICATION

When we use Mantel–Haenszel methods to control for confounding we are making an important assumption; namely that the Exposure–Disease (E–D) association is really the same in each of the strata defined by the levels of the confounding variable, C. If this is not true, then it makes little sense to combine the odds ratios (the estimates of the effect of E on D) from the different strata. If the effect of E on D varies according to the level of C then we say that C modifies the effect of E on D: in other words there is **effect modification**. A number of different terms are used to describe effect modification:

- **Effect modification**: C modifies the effect of E on D.
- **Interaction**: there is interaction between the effects of E and C (on D).
- **Heterogeneity between strata**: the estimates of the E–D association differ between the strata.

Similarly, you may see tests for effect modification described as either **tests for interaction** or **tests of homogeneity** across strata.

Testing for effect modification

The use of regression models to examine effect modification (or equivalently interaction) is discussed in Section 29.5. This is the most flexible approach. When we are using Mantel–Haenszel methods to control for confounding, an alternative is to use a χ^2 test for effect modification. This is equivalently, and more commonly, called a **χ^2 test of heterogeneity**. Under the null hypothesis of no effect modification, all the individual stratum odds ratios would equal the overall summary odds ratio. In other words:

$$\text{OR}_i = \frac{d_{1i} \times h_{0i}}{d_{0i} \times h_{1i}} = \text{OR}_{MH}$$

Multiplying both sides of the equation by $d_{0i} \times h_{1i}$ and rearranging shows that, under the null hypothesis of no effect modification, the following set of differences would be zero:

$$(d_{1i} \times h_{0i} - \text{OR}_{MH} \times d_{0i} \times h_{1i}) = 0$$

The **χ^2 test of heterogeneity** is based on a weighted sum of the squares of these differences:

$$\chi^2 = \Sigma \frac{(d_{1i} \times h_{0i} - \text{OR}_{MH} \times d_{0i} \times h_{1i})^2}{\text{OR}_{MH} \times V_i \times n_i^2}, \text{ d.f.} = c - 1$$

where V_i is as defined in Section 18.4, and c is the number of strata. The greater the differences between the stratum-specific odds ratios and OR_{MH}, the larger will be the heterogeneity statistic.

Example 18.1 (continued)

In our example, the odds ratios were 2.57 (95% CI 1.21 to 5.45) in males and 1.71 (95% CI 0.778 to 3.78) in females. Given that the confidence intervals easily overlapped, we would not expect to find evidence of effect modification (i.e. that the OR in males is different to the OR in females). The calculations needed to

Table 18.6 Calculations required for the χ^2 test of heterogeneity.

Stratum (i)	$(d_{1i} \times h_{0i} - \text{OR}_{MH} \times d_{0i} \times h_{1i})^2$	$\text{OR}_{MH} \times V_i \times n_i^2$	$\dfrac{(d_{1i} \times h_{0i} - \text{OR}_{MH} \times d_{0i} \times h_{1i})^2}{\text{OR}_{MH} \times V_i \times n_i^2}$
Males ($i = 1$)	$(36 \times 50 - 2.13 \times 50 \times 14)^2$ $= 97056.2$	$2.13 \times 8.21 \times 150^2$ $= 392737$	$\dfrac{97056.2}{392737} = 0.247$
Females ($i = 2$)	$(24 \times 90 - 2.13 \times 10 \times 126)^2$ $= 269601$	$2.13 \times 7.08 \times 150^2$ $= 940728$	$\dfrac{269601}{940728} = 0.287$
Total			0.534

apply the formula above are given in Table 18.6. The resulting value of the χ^2 test of heterogeneity is:

$$\chi^2 = 0.534, \text{ d.f.} = 1, \; P = 0.470$$

There is thus no evidence that gender modifies the association between rural/urban residence and leptospirosis antibodies.

When does effect modification matter?

As discussed above, Mantel–Haenszel methods assume that the true E–D odds ratio is the same in each stratum, and that the only reason for differences in the observed odds ratios between strata is sampling variation. We should therefore check this assumption, by applying the χ^2 test for heterogeneity, before reporting Mantel–Haenszel odds ratios, confidence intervals and P-values. This test has low *power* (see Chapter 35): it is unlikely to yield evidence for effect modification unless there are large differences between strata. A large P-value does not therefore establish the absence of effect modification. In fact, as the true odds ratios are never likely to be *exactly* the same in each stratum, effect modification is always present to some degree. Most researchers would accept, however, that minor effect modification should be ignored in order to simplify the presentation of the data.

The following box summarizes a practical approach to examining for effect modification, and recommends how analyses should be presented when evidence for effect modification is found. These issues are also discussed in Section 29.5 and Chapter 38, which describes strategies for data analysis.

BOX 18.1 A PRACTICAL APPROACH TO EXAMINING FOR EFFECT MODIFICATION

1 Always examine the pattern of odds ratios in the different strata: how different do they look, and is there any trend across strata?
2 If there is clear evidence of effect modification, and substantial differences in the E–D association between strata, report this and report the E–D association separately in each stratum.
3 If there is moderate evidence of effect modification, use Mantel–Haenszel methods but in addition report stratum-specific estimates of the E–D association.
4 If there is no evidence of effect modification, report this and use Mantel–Haenszel methods.

18.6 STRATIFICATION ON MORE THAN ONE CONFOUNDING VARIABLE

It is possible to apply the Mantel–Haenszel methods to control simultaneously for the effects of two or more confounders. For example, we can control additionally for differences in age distribution between the urban and rural areas by grouping our population into four age groups and forming the $2 \times 4 = 8$ strata corresponding to all combinations of gender and age group. The drawback to this approach is that the number of strata increases rapidly as we attempt to control for the effects of more confounding variables, so that it becomes impossible to estimate the stratum-specific odds ratios (although the Mantel-Haenszel OR can still be derived).

The alternative is to use regression models. The use of logistic regression models to control for confounding is considered in detail in Chapter 20.

CHAPTER 19

Logistic regression: comparing two or more exposure groups

19.1 INTRODUCTION

In this chapter we introduce **logistic regression**, the method most commonly used for the analysis of *binary* outcome variables. We show how it can be used to examine the effect of a *single* exposure variable, and in particular, how it can be used to:

- Compare a binary outcome variable between two exposure (or treatment) groups.
- Compare more than two exposure groups.
- Examine the effect of an ordered or continuous exposure variable.

We will see that it gives *very similar results* to the methods for analysing *odds ratios* described in Chapters 16, 17 and 18, and is an alternative to them. We will also see how logistic regression provides a flexible means of analysing the association between a binary outcome and a *number* of exposure variables. In the next chapter, we will explain how it is used to control for confounding. We will also briefly describe the regression analysis of risk ratios, and methods for the analysis of categorical outcomes with more than two levels.

We will explain the principles of logistic regression modelling in detail in the next section, in the simple context of comparing two exposure groups. In particular, we will show how it is based on modelling odds ratios, and explain how to interpret the computer output from a logistic regression analysis. We will then introduce the general form of the logistic regression equation, and explain where the name ‘logistic’ comes from. Finally we will explain how to fit logistic regression models for categorical, ordered or continuous exposure variables.

Links between multiple regression models for the analysis of numerical outcomes, the logistic regression models introduced here, and other types of regression model introduced later in the book, are discussed in detail in Chapter 29.

19.2 LOGISTIC REGRESSION FOR COMPARING TWO EXPOSURE GROUPS

Introducing the logistic regression model

We will start by showing, in the simple case of two exposure groups, how logistic regression models the association between binary outcomes and exposure variables in terms of odds ratios. Recall from Chapter 16 that the *exposure odds ratio* (OR) is defined as:

$$\text{Exposure odds ratio} = \frac{\text{Odds in exposed group}}{\text{Odds in unexposed group}}$$

If we re-express this as:

$$\text{Odds in exposed} = \text{Odds in unexposed} \times \text{Exposure odds ratio}$$

then we have the basis for a simple model for the odds of the outcome, which expresses the odds in each group in terms of two **model parameters**. These are:

1 The **baseline** odds. We use the term **baseline** to refer to the exposure group against which all the other groups will be compared. When there are just two exposure groups as here, then the baseline odds are the odds in the unexposed group. We will use the parameter name 'Baseline' to refer to the odds in the baseline group.

2 The **exposure odds ratio**. This expresses the effect of the exposure on the odds of disease. We will use the parameter name 'Exposure' to refer to the exposure odds ratio.

Table 19.1 shows the odds in each of the two exposure groups, in terms of the parameters of the logistic regression model.

Table 19.1 Odds of the outcome in terms of the parameters of a logistic regression model comparing two exposure groups.

Exposure group	Odds of outcome	Odds of outcome, in terms of the parameter names
Exposed (group 1)	Baseline odds × exposure odds ratio	Baseline × Exposure
Unexposed (group 0)	Baseline odds	Baseline

The logistic regression model defined by the two equations for the odds of the outcome shown in Table 19.1 can be abbreviated to:

$$\boxed{\text{Odds} = \text{Baseline} \times \text{Exposure}}$$

Since the two parameters in this model *multiply* together, the model is said to be **multiplicative**. This is in contrast to the multiple regression models described in Chapter 11, in which the effects of different exposures were *additive*. If there were two exposures (A and B), the model would be:

$$\text{Odds} = \text{Baseline} \times \text{Exposure(A)} \times \text{Exposure(B)}$$

Thus if, for example, exposure A doubled the odds of disease and exposure B trebled it, a person exposed to both would have a six times greater odds of disease than a person in the baseline group exposed to neither. We describe such models in detail in the next chapter.

Example 19.1

All our examples of logistic regression models are based on data from a study of onchocerciasis ('river blindness') in Sierra Leone (McMahon *et al*. 1988, *Trans Roy Soc Trop Med Hyg* **82**; 595–600), in which subjects were classified according to whether they lived in villages in savannah (grassland) or rainforest areas. In addition, subjects were classified as infected if microfilariae (*mf*) of *Onchocerciasis volvulus* were found in skin snips taken from the iliac crest. The study included persons aged 5 years and above. Table 19.2 shows that the prevalence of microfilarial infection appears to be greater for individuals living in rainforest areas compared to those living in the savannah; the associated odds ratio is $2.540/1.052 = 2.413$.

We will now show how to use logistic regression to examine the association between area of residence and microfilarial infection in these data. To use a **computer package to fit a logistic regression model**, it is necessary to specify just two items:

1 The *name of the outcome* variable, which in this case is *mf*. The **required convention for coding** is to code the outcome event (D) as 1, and the absence of the outcome event (H) as 0. The variable *mf* was therefore coded as 0 for uninfected subjects and 1 for infected subjects.

2 The *name of the exposure* variable(s). In this example, we have just one exposure variable, which is called *area*. The required *convention for coding* is that used throughout this book; thus *area* was coded as 0 for subjects living in savannah areas (the *baseline* or '*unexposed*' group) and 1 for subjects living in rainforest areas (the '*exposed*' group).

Table 19.2 Numbers and percentages of individuals infected with onchocerciasis according to their area of residence, in a study of 1302 individuals in Sierra Leone.

	Microfilarial infection			
Area of residence	Yes	No	Total	Odds of infection
Rainforest	$d_1 = 541$ (71.7%)	$h_1 = 213$ (28.3%)	754	$541/213 = 2.540$
Savannah (baseline group)	$d_0 = 281$ (51.3%)	$h_0 = 267$ (48.7%)	548	$281/267 = 1.052$
Total	822	480	1302	

Table 19.3 First ten lines of the computer dataset from the study of onchocerciasis.

id	*mf*	Area
1	1	0
2	1	1
3	1	0
4	0	1
5	0	0
6	0	1
7	1	0
8	1	1
9	1	1
10	1	1

The first ten lines of the dataset, when entered on the computer, are shown in Table 19.3. For example, subject number 1 lived in a savannah area and was infected, number 2 lived in a rainforest area and was also infected, whereas subject number 4 lived in a rainforest area but was not infected.

The **logistic regression model** that will be fitted is:

$$\text{Odds of } mf \text{ infection} = \text{Baseline} \times \text{Area}$$

Its two parameters are:

1 baseline: the odds of infection in the baseline group (subjects living in savannah areas); and

2 area: the odds ratio comparing odds of infection among subjects living in rainforest areas with that among those living in savannah areas.

Table 19.4 shows the computer output obtained from fitting this model. The two *rows* in the output correspond to the two *parameters* of the logistic regression model; area is our exposure of interest, and the **constant** term refers to the baseline group. The same format is used for both parameters, and is based on what makes sense for interpretation of the effect of exposure. This means that some of the information presented for the constant (baseline) parameter is not of interest.

Table 19.4 Logistic regression output for the model relating odds of infection to area of residence, in 1302 subjects participating in a study of onchocerciasis in Sierra Leone.

	Odds ratio	z	$P > \|z\|$	95% CI
Area	2.413	7.487	0.000	1.916 to 3.039
Constant	1.052	0.598	0.550	0.890 to 1.244

The column labelled 'Odds ratio' contains the **parameter estimates**:

1 For the first row, labelled 'area', this is the *odds ratio* (2.413) comparing rainforest (area 1) with savannah (area 0). This is identical to the odds ratio which was calculated directly from the raw data (see Table 19.3).

2 For the second row, labelled 'constant', this is the *odds of infection in the baseline group* (1.052 = odds of infection in the savannah area, see Table 19.3). As we will see, this apparently inconsistent labelling is because output from regression models is labelled in a uniform way.

The remaining columns present *z* statistics, *P*-values and 95% confidence intervals corresponding to the model parameters. The values for *area* are exactly the same as those that would be obtained by following the procedures described in Section 16.7 for the calculation of a 95% confidence interval for an odds ratio, and the associated Wald test. They will be explained in more detail in the explanation of Table 19.5 below.

The logistic regression model on a log scale

As described in Chapter 16, confidence intervals for odds ratios are derived by using the standard error of the *log* odds ratio to calculate a confidence interval for the *log* odds ratio. The results are then *antilogged* to express them in terms of the original scale. The same is true for logistic regression models; they are *fitted on a log scale*. Table 19.5 shows the two equations that define the logistic regression model for the comparison of two exposure groups. The middle column shows the model for the odds of the outcome, as described above. Using the rules of logarithms (see p. 156, Section 16.5), it follows that corresponding equations on the log scale for the log of the odds of the outcome are as shown in the right-hand column. Note that as in the rest of the book all logs are to the base *e* (natural logarithms) unless they are explicitly denoted as logs to the base 10 by $\log_{10}$ (see Section 13.2).

Table 19.5 Equations defining the logistic regression model for the comparison of two exposure groups.

Exposure group	Odds of outcome	Log odds of outcome
Exposed (group 1)	Baseline odds × exposure OR	Log(baseline odds) + log(exposure OR)
Unexposed (group 0)	Baseline odds	Log(baseline odds)

Using the parameter names introduced earlier in this section, the logistic regression model on the log scale can be written:

$$\log(\text{Odds}) = \log(\text{Baseline}) + \log(\text{Exposure odds ratio})$$

In practice, we abbreviate it to:

$$\log(\text{Odds}) = \text{Baseline} + \text{Exposure}$$

since it is clear from the context that *output on the log scale refers to log odds and log odds ratios*. Note that whereas the exposure effect on the odds ratio scale is *multiplicative*, the exposure effect on the log scale is *additive*.

Example 19.1 (continued)

In this example, the model on the log scale is:

$$\log(\text{Odds of } mf \text{ infection}) = \text{Baseline} + \text{Area}$$

where

1 *baseline* is the *log odds* of infection in the savannah areas; and

2 *area* is the *log odds ratio* comparing the odds of infection in rainforest areas with that in savannah areas.

Table 19.6 shows the results obtained on the log scale, for this model. We will explain each item in the table, and then discuss how the results relate to those on the odds ratio scale, shown in Table 19.4.

Table 19.6 Logistic regression output (log scale) for the association between microfilarial infection and area of residence.

	Coefficient	s.e.	z	$P > \|z\|$	95% CI
Area	0.881	0.118	7.487	0.000	0.650 to 1.112
Constant	0.0511	0.0854	0.598	0.550	−0.116 to 0.219

1 The two *rows* in the output correspond to the terms in the model; area is our exposure of interest, and as before the **constant term** corresponds to the baseline group.

2 The *first* column gives the results for the **regression coefficients** (corresponding to the parameter estimates on a log scale):

(a) For the row labelled 'area', this is the **log odds ratio** comparing rainforest with savannah. It agrees with what would be obtained if it were calculated directly from Table 19.3, and with the value in Table 19.4:

$$\log \text{OR} = \log(2.540/1.052) = \log(2.413) = 0.881$$

(b) For the row labelled 'constant', this is the **log odds in the baseline group** (the group with exposure level 0), i.e. the log odds of microfilarial infection in the savannah:

$$\log \text{odds} = \log(281/267) = \log(1.052) = 0.0511.$$

3 The *second* column gives the standard error(s) of the regression coefficient(s). In the simple example of a binary exposure variable, as we have here, the standard errors of the regression coefficients are exactly the same as those derived using the formulae given in Chapter 16. Thus:

(a) s.e.(log OR comparing rainforest with savannah) is:

$$\begin{aligned}\sqrt{(1/d_1 + 1/h_1 + 1/d_0 + 1/h_0)} &= \sqrt{(1/541 + 1/213 + 1/281 + 1/267)}\\ &= 0.118\end{aligned}$$

(b) s.e.(log odds in savannah) is:

$$\sqrt{(1/d_0 + 1/h_0)} = \sqrt{(1/281 + 1/267)} = 0.0854$$

4 The 95% confidence intervals for the regression coefficients in the *last* column are derived in the usual way.

(a) For the log OR comparing rainforest with savannah, the 95% CI is:

$$0.881 - (1.96 \times 0.118) \text{ to } 0.881 + (1.96 \times 0.118) = 0.650 \text{ to } 1.112$$

(b) For the log odds in the savannah, the 95% CI is:

$$0.0511 - (1.96 \times 0.0854) \text{ to } 0.0511 + (1.96 \times 0.0854) = -0.116 \text{ to } 0.219$$

5 The z statistic in the *area* row of the third column is used to derive a **Wald test** (see Chapter 28) of the null hypothesis that the *area* coefficient = 0, i.e. that the exposure has no effect (since if log OR = 0, then OR must be equal to 1). This z statistic is simply the regression coefficient divided by its standard error:

$$z = 0.881/0.118 = 7.487$$

6 The P-value in the *fourth* column is derived from the z statistic in the usual manner (see Table A1 and Chapter 8), and can be used to assess the strength of the evidence against the null hypothesis that the true (population) exposure effect is zero. Thus, the P-value of 0.000 (which should be interpreted as < 0.001) for the log OR comparing rainforest with savannah indicates that there is strong evidence against the null hypothesis that the odds of microfilarial infection are the same in the two areas.

7 We are usually not interested in in the third and fourth columns (the z statistic and its P-value) for the *constant* row. However, for completeness, we will explain their meanings:

(a) The z statistic is the result of testing the null hypothesis that the log odds of infection in the savannah areas are 0 (or, equivalently, that the odds of infection are 1). This would happen if the risk of infection in the savannah areas was 0.5; in other words if people living in the savannah areas were equally likely to be infected as they were to be not infected.

(b) The P-value of 0.550 for the log odds in savannah areas indicates that there is no evidence against this null hypothesis.

Relation between outputs on the ratio and log scales

We will now explain the relationship between the two sets of outputs, since the results in Table 19.4 (output on the original, or ratio, scale) are derived from the results in Table 19.6 (output on the log scale). Once this is understood, it is rarely necessary to refer to the output displayed on the log scale: the most useful results are the odds ratios, confidence intervals and P-values displayed on the original scale, as in Table 19.4.

1 In Table 19.4, the column labelled 'Odds Ratio' contains the *exponentials* (antilogs) of the logistic regression coefficients shown in Table 19.6. Thus the OR comparing rainforest with savannah $= \exp(0.881) = 2.413$.

2 The z statistics and P-values are derived from the log odds ratio and its standard error, and so are identical in the two tables.

3 The 95% confidence intervals in Table 19.4 are derived by antilogging (exponentiating) the confidence intervals on the log scale presented in Table 19.6. Thus the 95% CI for the OR comparing rainforest with savannah is:

$$95\%\ \text{CI} = \exp(0.650) \text{ to } \exp(1.112) = 1.916 \text{ to } 3.039$$

This is identical to the 95% CI calculated using the methods described in Section 16.7:

$$95\%\ \text{CI (OR)} = \text{OR}/\text{EF to OR} \times \text{EF}, \text{ where EF} = \exp[1.96 \times \text{s.e.}(\log\ \text{OR})]$$

Note that since the calculations are multiplicative:

$$\frac{\text{Odds ratio}}{\text{Lower confidence limit}} = \frac{\text{Upper confidence limit}}{\text{Odds ratio}}$$

This can be a useful check on confidence limits presented in tables in published papers.

19.3 GENERAL FORM OF THE LOGISTIC REGRESSION EQUATION

We will now introduce the general form of the logistic regression model with several exposure variables, and explain how it corresponds to what we used above in the simple case when we are comparing two exposure groups, and therefore have a single exposure variable in our model. The general form of the logistic regression model is similar to that for multiple regression (see Chapter 11):

$$\text{log odds of outcome} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

The difference is that we are modelling a transformation of the outcome variable, namely the *log of the odds of the outcome*. The quantity on the right-hand side of the equation is known as the **linear predictor** of the log odds of the outcome, given the particular value of the p exposure variables x_1 to x_p. The β's are the **regression coefficients** associated with the p exposure variables.

The transformation of the probability, or risk, π of the outcome into the log odds is known as the **logit function**:

$$\text{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right)$$

and the name **logistic** is derived from this. Recall from Section 14.6 (Table 14.2) that while probabilities must lie between 0 and 1, odds can take any value between 0 and infinity (∞). The log odds are not constrained at all; they can take any value between $-\infty$ and ∞.

We will now show how the general form of the logistic regression model corresponds to the logistic regression model we used in Section 19.2 for comparing two exposure groups. The general form for comparing two exposure groups is:

$$\text{log odds of outcome} = \beta_0 + \beta_1 x_1$$

where x_1 (the exposure variable) equals 1 for those in the *exposed* group and 0 for those in the *unexposed* group. Table 19.7 shows the value of the log odds predicted

Table 19.7 Log odds of the outcome according to exposure group, as calculated from the linear predictor in the logistic regression equation.

Exposure group	Log odds of outcome, predicted from model	Log odds of outcome, in terms of the parameter names
Exposed ($x_1 = 1$)	$\beta_0 + \beta_1 \times 1 = \beta_0 + \beta_1$	log(Baseline odds) + log(Exposure odds ratio)
Unexposed ($x_1 = 0$)	$\beta_0 + \beta_1 \times 0 = \beta_0$	log(Baseline odds)

from this model in each of the two exposure groups, together with the log odds expressed in terms of the parameter names, as in Section 19.2.

We can see that the first regression coefficient, β_0, corresponds to the log odds in the unexposed (baseline) group. We will now show how the other regression coefficient, β_1, corresponds to the log of the exposure odds ratio. Since:

$$\text{Exposure OR} = \frac{\text{odds in exposed group}}{\text{odds in unexposed group}}$$

it follows from the rules of logarithms (see p. 156) that:

$$\text{log OR} = \text{log(odds in exposed group)} - \text{log(odds in unexposed group)}$$

Putting the values predicted from the logistic regression equation (shown in Table 19.7) into this equation gives:

$$\text{log OR} = \beta_0 + \beta_1 - \beta_0 = \beta_1$$

The equivalent model on the ratio scale is:

$$\text{Odds of disease} = \exp(\beta_0 + \beta_1 x_1) = \exp(\beta_0) \times \exp(\beta_1 x_1)$$

In this *multiplicative model* $\exp(\beta_0)$ corresponds to the odds of disease in the baseline group, and $\exp(\beta_1)$ to the exposure odds ratio. Table 19.8 shows how this model corresponds to the model shown in Table 19.1.

Table 19.8 Odds of outcome according to exposure group, as calculated from the linear predictor in the logistic regression equation.

Exposure group	Odds of outcome, predicted from model	Odds of outcome, in terms of the parameter names
Exposed ($x_1 = 1$)	$\exp(\beta_0) \times \exp(\beta_1)$	Baseline odds × Exposure odds ratio
Unexposed ($x_1 = 0$)	$\exp(\beta_0)$	Baseline odds

19.4 LOGISTIC REGRESSION FOR COMPARING MORE THAN TWO EXPOSURE GROUPS

We now consider logistic regression models for **categorical exposure variables** with more than two levels. To examine the effect of categorical variables in logistic and other regression models, we look at the effect of each level compared to a **baseline** group. When the exposure is an *ordered* categorical variable, it may also be useful to examine the average change in the log odds per exposure group, as described in Section 19.5.

Table 19.9 Association between age group and microfilarial infection in the onchocerciasis study.

Age group (years)	Coded value in dataset	Microfilarial infection		Odds of infection	Odds ratio compared to the baseline group
		Yes	No		
5–9	0	46	156	46/156 = 0.295	1
10–19	1	99	119	99/119 = 0.832	0.832/0.295 = 2.821
20–39	2	299	125	299/125 = 2.392	2.392/0.295 = 8.112
≥ 40	3	378	80	378/80 = 4.725	4.725/0.295 = 16.02
Total		822	480		

Example 19.2

In the onchocerciasis study, introduced in Example 19.1, subjects were classified into four age groups: 5–9, 10–19, 20–39 and ≥ 40 years. Table 19.9 shows the association between age group and microfilarial infection. The odds of infection increased markedly with increasing age. A chi-squared test for association in this table gives $P < 0.001$, so there is clear evidence of an association between age group and infection. We chose the 5–9 year age group as the **baseline** exposure group, because its coded value in the dataset is zero, and calculated odds ratios for each non-baseline group relative to the baseline group.

The corresponding logistic regression model uses this same approach; the effect of each non-baseline age group is expressed in terms of the odds ratio comparing it with the baseline. The parameters of the model, on both the odds and log odds scales, are shown in Table 19.10.

Table 19.10 Odds and log odds of the outcome in terms of the parameters of a logistic regression model comparing four age groups.

Age group	Odds of infection	Log odds of infection
0 (5–9 years)	Baseline	Log(Baseline)
1 (10–19 years)	Baseline × Agegrp(1)	Log(Baseline) + Log(Agegrp(1))
2 (20–39 years)	Baseline × Agegrp(2)	Log(Baseline) + Log(Agegrp(2))
3 (≥ 40 years)	Baseline × Agegrp(3)	Log(Baseline) + Log(Agegrp(3))

Here, Agegrp(1) is the odds ratio (or, on the log scale, the log odds ratio) comparing group 1 (10–19 years) with group 0 (5–9 years, the baseline group), and so on. This regression model has four parameters:

1 the odds of infection in the 5–9 year group (the baseline group); and

2 the *three* odds ratios comparing the non-baseline groups with the baseline.

Using the notation introduced in Section 19.2, the four equations for the odds that define the model in Table 19.10 can be written in abbreviated form as:

$$\text{Odds} = \text{Baseline} \times \text{Agegrp}$$

or on a log scale, as:

$$\log(\text{Odds}) = \text{Baseline} + \text{Agegrp}$$

The effect of categorical variables is modelled in logistic and other regression models by using **indicator variables**, which are created automatically by most statistical packages when an exposure variable is defined as categorical. This is explained further in Box 19.1. Output from this model (expressed on the odds ratio scale, with the constant term omitted) is shown in Table 19.11.

Table 19.11 Logistic regression output (odds ratio scale) for the association between microfilarial infection and age group.

	Odds ratio	z	$P > \lvert z \rvert$	95% CI
agegrp(1)	2.821	4.802	0.000	1.848 to 4.308
agegrp(2)	8.112	10.534	0.000	5.495 to 11.98
agegrp(3)	16.024	13.332	0.000	10.658 to 24.09

BOX 19.1 USE OF INDICATOR VARIABLES IN REGRESSION MODELS

To model the effect of an exposure with more than two categories, we estimate the odds ratio for each non-baseline group compared to the baseline. In the logistic regression equation, we represent the exposure by a set of **indicator variables** (variables which take only the values 0 and 1) representing each non-baseline value of the exposure variable. The regression coefficients for these indicator variables are the corresponding (log) odds ratios. For example, to estimate the odds ratios comparing the 10–19, 20–39 and ≥ 40 year groups with the 5–9 year group, we create three indicator variables which we will call ageind$_1$, ageind$_2$ and ageind$_3$ (the name is not important). The table below shows the value of these indicator variables according to age group.

Value of indicator variables for use in logistic regression of the association between microfilarial infection and age group.

Age group	ageind$_1$	ageind$_2$	ageind$_3$
0 (5–9 years)	0	0	0
1 (10–19 years)	1	0	0
2 (20–29 years)	0	1	0
3 (≥ 40 years)	0	0	1

All three of these indicator variables (but not the original variable) are then included in a logistic regression model. Most statistical packages create the indicator variables automatically when the original variable is declared as categorical.

The *P*-values for the three indicator variables (corresponding to the non-baseline age groups) can be used to test the null hypotheses that there is no difference in odds of the outcome between the individual non-baseline exposure groups and the baseline group. However, these are not usually of interest: we need a test, analogous to the χ^2 test for a table with four rows and two columns, of the general null hypothesis that there is no association between age group and infection. We will see how to test such null hypotheses in regression models in Chapter 29, and in the next section we address the special case when the categorical variable is ordered, as is the case here. It is usually a mistake to conclude that there is a difference between one exposure group and the rest based on a particular (small) *P*-value corresponding to one of a set of indicator variables.

19.5 LOGISTIC REGRESSION FOR ORDERED AND CONTINUOUS EXPOSURE VARIABLES

Until now, we have considered logistic regression models for binary or categorical exposure variables. For binary variables, logistic regression estimates the odds ratio comparing the two exposure groups, while for categorical variables we have seen how to estimate odds ratios for each non-baseline group compared to the baseline. This approach does not take account of ordering of the exposure variable. For example, we did not use the fact that subjects aged ≥ 40 years are older than those aged 20–39 years, who in turn are older than those aged 10–19 years and so on.

Example 19.3

The odds of microfilarial infection in each age group in the onchocerciasis dataset are shown in Table 19.9 in Section 19.4, and are displayed in Figure 19.1. We do not have a straight line; the slope of the line increases with increasing age group. In other words, this increase in the odds of infection with increasing age does not appear to be constant.

However, Figure 19.2 shows that there *is* an approximately linear increase in the **log odds** of infection with increasing age group. This log-linear increase means that we are able to express the association between age and the log odds of microfilarial infection by a single linear term (as described below) rather than by a series of indicator variables representing the different groups.

Relation with linear regression models

Logistic regression models can be used to estimate the *most likely* value of the increase in log odds per age group, assuming that the increase is the same in each age group. (We will define the meaning of ‘most likely’ more precisely in Chapter

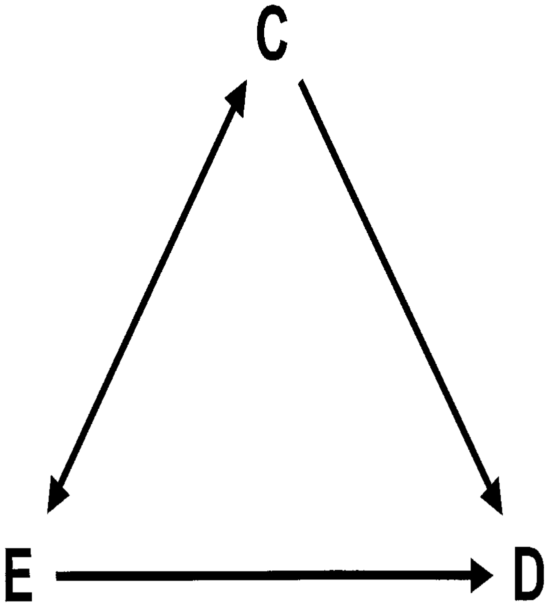

Fig. 19.1 Odds of microfilarial infection according to age group for the onchocerciasis data.

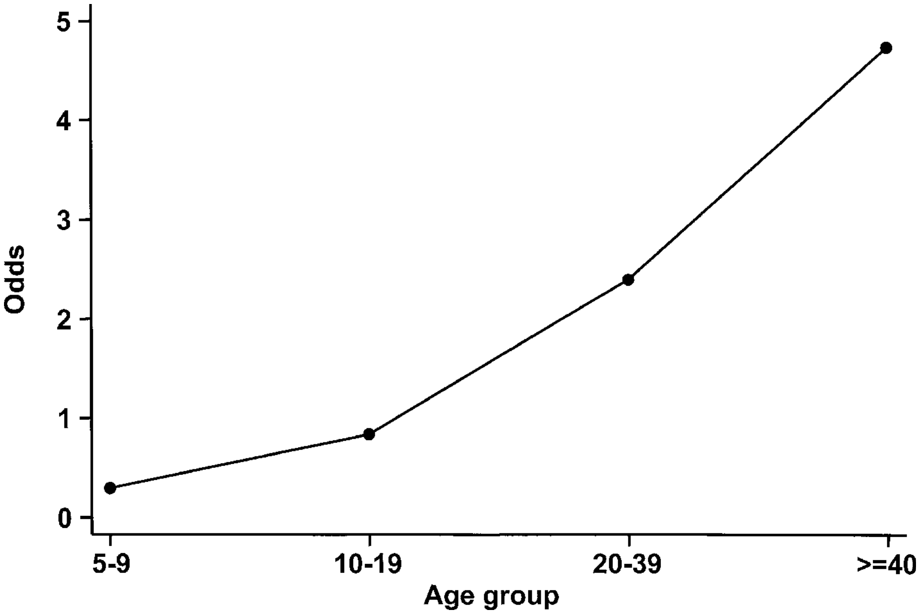

Fig. 19.2 Log odds of microfilarial infection according to age group for the onchocerciasis data.

28.) The model is analogous to the simple linear regression model described in Chapter 11. If we assume that:

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

then the intercept β_0 is the value of y when $x = 0$, and the slope β_1 represents the increase in y when x increases by 1. Logistic regression models assume that:

$$\text{log odds} = \beta_0 + \beta_1 x$$

so that the intercept β_0 is the value of the log odds when $x = 0$, and the slope β_1 represents the increase in log odds when x increases by 1. We will use the notation

$$\text{log odds} = \text{Baseline} + [\text{X}]$$

where the square brackets indicate our assumption that variable X has a linear effect on the log odds of the outcome. For the onchocerciasis data, our model is

$$\text{log odds} = \text{Baseline} + [\text{Agegrp}]$$

Example 19.3 (continued)

Table 19.12(a) shows logistic regression output for the model assuming a linear effect of logistic regression on the log odds of microfilarial infection. The estimated increase in log odds for every unit increase in age group is 0.930 (95% CI = 0.805 to 1.055). This corresponds to an odds ratio per group of 2.534 (95% CI = 2.236 to 2.871; see output in Table 19.12b). The constant term corresponds to the estimated log odds of microfilarial infection in age group 0 (5–9 years, log odds = −1.115), *assuming a linear relation* between age group and the log odds of infection. It does not therefore numerically equal the baseline term in the

Table 19.12 Logistic regression output for the linear association between the log odds of microfilarial infection and age group (data in Table 19.9).

(a) Output on log scale.

	Coefficient	s.e.	z	$P > \lvert z\rvert$	95% CI
Age group	0.930	0.0638	14.587	0.000	0.805 to 1.055
Constant	−1.115	0.127	−8.782	0.000	−1.364 to −0.866

(b) Output on ratio scale.

	Odds ratio	z	$P > \lvert z\rvert$	95% CI
Age group	2.534	14.587	0.000	2.236 to 2.871

Table 19.13 Predicted log odds in each age group, derived from a logistic regression model assuming a linear relationship between the log odds of microfilarial infection and age group.

Age group	Logistic regression equation	Predicted log odds
0	log odds = constant + 0 × age group	−1.115 + 0.930 × 0 = −1.115
1	log odds = constant + 1 × age group	−1.115 + 0.930 × 1 = −0.185
2	log odds = constant + 2 × age group	−1.115 + 0.930 × 2 = 0.745
3	log odds = constant + 3 × age group	−1.115 + 0.930 × 3 = 1.674

regression equation when age is included as a categorical variable, as described in Section 19.4.

Substitution of the estimated regression coefficients into the logistic regression equation gives the **predicted log odds** in each age group. These are shown in Table 19.13. Figure 19.3 compares these predicted log odds from logistic regression with the observed log odds in each group. This shows that the linear assumption gives a good approximation to the observed log odds in each group. Section 29.6 describes how to test such linear assumptions.

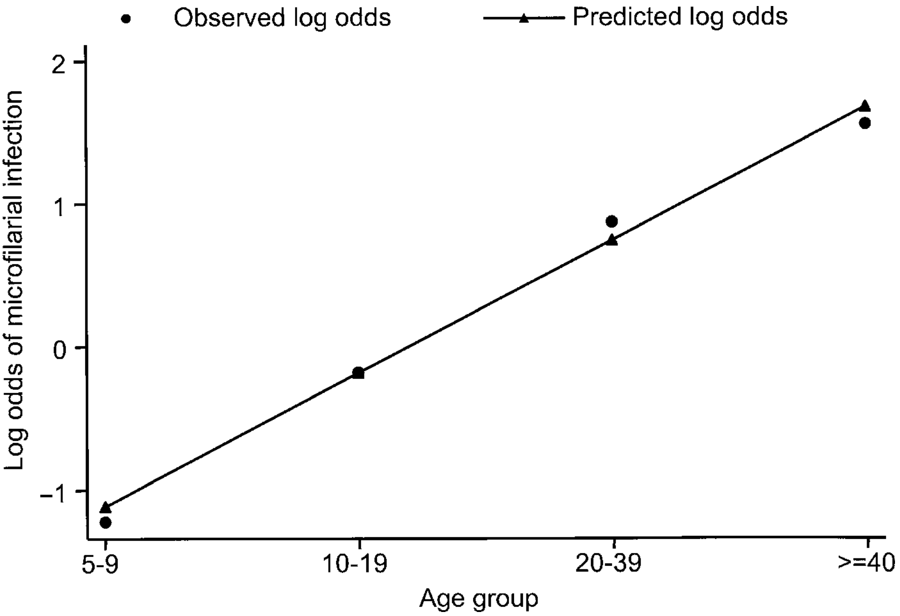

Fig. 19.3 Observed log odds in each age group (circles) and predicted log odds from logistic regression (triangles, connected by line).

CHAPTER 20

Logistic regression: controlling for confounding and other extensions

20.1 INTRODUCTION

In the last chapter we introduced the principles of logistic regression models, and described how to use logistic regression to examine the effect of a single exposure variable. We now describe how these models can be extended to control for the confounding effects of one or more additional variables. In addition, we briefly cover regression modelling for risk ratios, rather than odds ratios, and for outcomes with more than two levels.

20.2 CONTROLLING FOR CONFOUNDING USING LOGISTIC REGRESSION

In Chapter 18 we saw how to control for a **confounding** variable by dividing the sample into strata defined by levels of the confounder, and examining the effect of the exposure in each stratum. We then used the Mantel–Haenszel method to combine the odds ratios from each stratum into an overall summary odds ratio. We also explained how this approach assumes that effect modification (interaction) is not present, i.e. that the true odds ratio comparing exposed with unexposed individuals is the same in each stratum. We now see how making the same assumption allows us to control for confounding using logistic regression.

We will explain this in the context of the onchocerciasis dataset used throughout Chapter 19. Recall that we found strong associations of both area of residence and of age group with the odds of microfilarial (*mf*) infection. If the age distributions differ in the two types of area, then it is possible that age is a confounding variable for the association between area and *mf* infection. We will control for this possible confounding by fitting a logistic regression model, which includes the effects of both area and age group. We will start with *hypothetical* data, constructed so that it is easy to see how this logistic regression model works. We will then explain how to interpret the output when we apply the model to the real data.

Example 20.1 (hypothetical data)

Table 20.1 shows *hypothetical* data for the odds of *mf* infection according to area of residence (exposure) and age group. You can see that:

1 Table 20.1(a) shows that the exposure effect is *exactly* the same in each of the age groups; the age-specific odds ratios comparing exposed with unexposed individuals are all equal to 3.0. (Note also that when the age groups are combined, the crude odds ratio is $1.86/0.92 = 2.02$, which is considerably less than the individual age-specific odds ratios of 3, confirming that age group confounds the association between *mf* infection and area.)

2 Table 20.1(b) shows that the age group effect is *exactly* the same in each area of residence. For example, the odds ratio comparing age group 1 with age group 0 in the savannah areas is $0.5/0.2 = 2.5$, the same as the odds ratio in the forest areas ($1.5/0.6 = 2.5$). Similarly, the odds ratio comparing age group 2 with age group 0 are 10 in each area, and the odds ratios comparing age group 3 with age group 0 are 15 in each area.

Table 20.1 *Hypothetical* data for the odds of *mf* infection, according to area of residence and age group.

(a) Crude data, and odds of disease in each group (d = number infected and h = number uninfected), plus odds ratios for area in each age-group and overall.

	Savannah areas (Unexposed)		Rainforest areas (Exposed)		Odds ratio for area effect
Age group	*d/h*	Odds	*d/h*	Odds	
0	20/100	0.2	30/50	0.6	3.0
1	40/80	0.5	60/40	1.5	3.0
2	80/40	2.0	60/10	6.0	3.0
3	90/30	3.0	45/5	9.0	3.0
All age groups combined	230/250	0.92	195/105	1.86	2.02

(b) Age group odds ratios (comparing age groups 1, 2 and 3 with age group 0), in each type of area of residence.

	Odds ratios for age group effects	
Age group	Savannah areas	Rainforest areas
0	1.0	1.0
1	2.5 (= 0.5/0.2)	2.5 (= 1.5/0.6)
2	10.0 (= 2.0/0.2)	10.0 (= 6.0/0.6)
3	15.0 (= 3.0/0.2)	15.0 (= 9.0/0.6)

These two facts mean that we can *exactly* express the odds of *mf* infection in the eight area–age subgroups in terms of the following five *parameters*, as shown in Table 20.2(a):

1 0.2: the odds of mf infection at the *baseline* values of *both* area and age group;

2 3.0: the *area* odds ratio comparing the odds of infection in rainforest areas compared to savannah areas; and

3 2.5, 10.0 and 15.0: the three *age* odds ratios comparing age groups 1, 2 and 3 with age group 0 (respectively).

Table 20.2(b) shows the corresponding equations in terms of the parameter names; these follow the convention we introduced in Chapter 19. These equations define the *logistic regression model* for the effects of area and age group on the odds of *mf* infection. As described in Chapter 19, such a logistic regression model can be abbreviated to:

$$\text{Odds} = \text{Baseline} \times \text{Area} \times \text{Agegrp}$$

As explained in Section 19.2, it is a *multiplicative* model for the joint effects of area and age group. Note that the Baseline parameter now refers to the odds of the disease at the baseline of *both* variables. This model *assumes* that the odds ratio for area is the same in each age group and that the odds ratios for age group are the same in each area, i.e. that there is *no interaction* between the effects of area and age group.

Table 20.2 Odds of *mf* infection by area and age group, expressed in terms of the parameters of the logistic regression model: Odds = Baseline × Area × Age group.

(a) Expressed in terms of the parameter values.

	Odds of *mf* infection	
Age group	Savannah areas (Unexposed)	Rainforest areas (Exposed)
0	$0.2 = 0.2$	$0.6 = 0.2 \times 3.0$
1	$0.5 = 0.2 \times 2.5$	$1.5 = 0.2 \times 3.0 \times 2.5$
2	$2.0 = 0.2 \times 10.0$	$6.0 = 0.2 \times 3.0 \times 10.0$
3	$3.0 = 0.2 \times 15.0$	$9.0 = 0.2 \times 3.0 \times 15.0$

(b) Expressed in terms of the parameter names.

	Odds of *mf* infection	
Age group	Savannah areas (Unexposed)	Rainforest areas (Exposed)
0	Baseline	Baseline × Area
1	Baseline × Agegrp(1)	Baseline × Area × Agegrp(1)
2	Baseline × Agegrp(2)	Baseline × Area × Agegrp(2)
3	Baseline × Agegrp(3)	Baseline × Area × Agegrp(3)

(c) Expressed on a *log* scale, in terms of the parameter names.

	Log odds of *mf* infection	
Age group	Savannah areas (Unexposed)	Rainforest areas (Exposed)
0	log(Baseline)	log(Baseline) + log(Area)
1	log(Baseline) + log(Agegrp(1))	log(Baseline) + log(Area) + log(Agegrp(1))
2	log(Baseline) + log(Agegrp(2))	log(Baseline) + log(Area) + log(Agegrp(2))
3	log(Baseline) + log(Agegrp(3))	log(Baseline) + log(Area) + log(Agegrp(3))

As explained in Chapter 19, the calculations to derive confidence intervals and P-values for the parameters of logistic regression models are done on the log scale, in which case the baseline parameter refers to the *log odds* in the baseline group, and the other parameters refer to *log odds ratios*. The effects of the exposure variables are additive on the log scale (as described in Section 19.2). Table 20.2(c) shows the equations for the log odds in each of the area–age subgroups. The corresponding logistic regression model, defined by these eight equations, is:

$$\log(\text{Odds}) = \log(\text{Baseline}) + \log(\text{Exposure}) + \log(\text{Age})$$

Example 20.2 (real data)

In our hypothetical example, we were able to precisely express the odds in the *eight* sub-groups in the table in terms of *five* parameters, because we created the data so that the effect of area was exactly the same in each age group, and the effect of age exactly the same in savannah and rainforest areas. Of course, sampling variation means that real data is never this neat, even if the model proposed is correct. Table 20.3 shows the odds of *mf* infection in the eight area–age subgroups, using the data that were actually observed in the onchocerciasis study.

Table 20.3 Odds of microfilarial infection and odds ratios comparing individuals living in forest areas with those living in savannah areas, separately for each age group.

	Area of residence		
Age group	Savannah	Rainforest	Odds ratio for area
0 (5–9 years)	16/77 = 0.208	30/79 = 0.380	1.828
1 (10–19 years)	22/50 = 0.440	77/69 = 1.116	2.536
2 (20–39 years)	123/85 = 1.447	176/40 = 4.400	3.041
3 ($\geq$ 40 years)	120/55 = 2.182	258/25 = 10.32	4.730

From the previous chapter (Table 19.4) we know that the crude odds ratio for area is 2.413 (the odds ratio which does not take into account the effects of age group, or any other variables). We can see in Table 20.3 that in three out of the four age groups the stratum-specific odds ratios for the effect of area of residence are larger than this. If we use Mantel–Haenszel methods (*see* Chapter 18) to estimate the effect of area of residence controlling for age group, we obtain an estimated odds ratio of 3.039 (95% CI = 2.310 to 3.999). This is noticeably larger than the crude odds ratio of 2.413.

As in the hypothetical example above, we can express the odds of *mf* infection in the rainforest areas in terms of the odds ratios for the effect of area of residence in each age group (Table 20.4a). Alternatively, we can express the odds of *mf* infection in terms of the odds ratios for each of the three age groups compared to age group 0 (Table 20.4b). Note that (in contrast to the hypothetical example above) these sets of odds ratios are not exactly the same in each area. This means that we cannot calculate the parameter estimates directly from the raw data, as we

Table 20.4 Odds of *mf* infection, according to area of residence and age group, for the data observed in the onchocerciasis study.

(a) With the odds in the rainforest areas expressed in terms of the age-specific odds ratios for the association between area and infection.

	Area	
Age group	Savannah	Rainforest
0 (5–9 years)	0.208	0.208 × 1.828
1 (10–19 years)	0.440	0.440 × 2.536
2 (20–39 years)	1.447	1.447 × 3.041
3 ($\geq$ 40 years)	2.182	2.182 × 4.730

(b) With the odds of infection in age groups 2 to 4 expressed in terms of the area-specific odds ratios for the association between age group and infection.

	Area	
Age group	Savannah	Rainforest
0 (5–9 years)	0.208	0.380
1 (10–19 years)	0.208 × 2.118	0.380 × 2.939
2 (20–39 years)	0.208 × 6.964	0.380 × 11.59
3 ($\geq$ 40 years)	0.208 × 10.50	0.380 × 27.18

could for the simpler examples in Chapter 19. Instead we use a computer package to fit the model and to estimate the *most likely* values for the effect of area controlling for age group, and the effect of age group controlling for area, on the basis of the assumption that there is no interaction between the effects of the two variables. The meaning of 'most likely' is explained more precisely in Chapter 28.

The computer output from this model (on the odds ratio scale) is shown in Table 20.5. The estimated odds ratio of 3.083 (95% CI = 2.354 to 4.038) for area controlling for age group is very close to that derived using the Mantel–Haenszel method (OR 3.039, 95% CI = 2.310 to 3.999), and again is noticeably larger than

Table 20.5 Logistic regression output for the model for *mf* infection, including both area of residence and age group.

	Odds ratio	z	$P > \|z\|$	95% CI
Area	3.083	8.181	0.000	2.354 to 4.038
Agegrp(1)	2.599	4.301	0.000	1.682 to 4.016
Agegrp(2)	9.765	10.944	0.000	6.493 to 14.69
Agegrp(3)	17.64	13.295	0.000	11.56 to 26.93
Constant*	0.147	−9.741	0.000	0.100 to 0.217

*Constant (baseline odds) = estimated odds of *mf* infection for 5–9 year olds living in the savannah areas, assuming no interaction between the effects of area and age group.

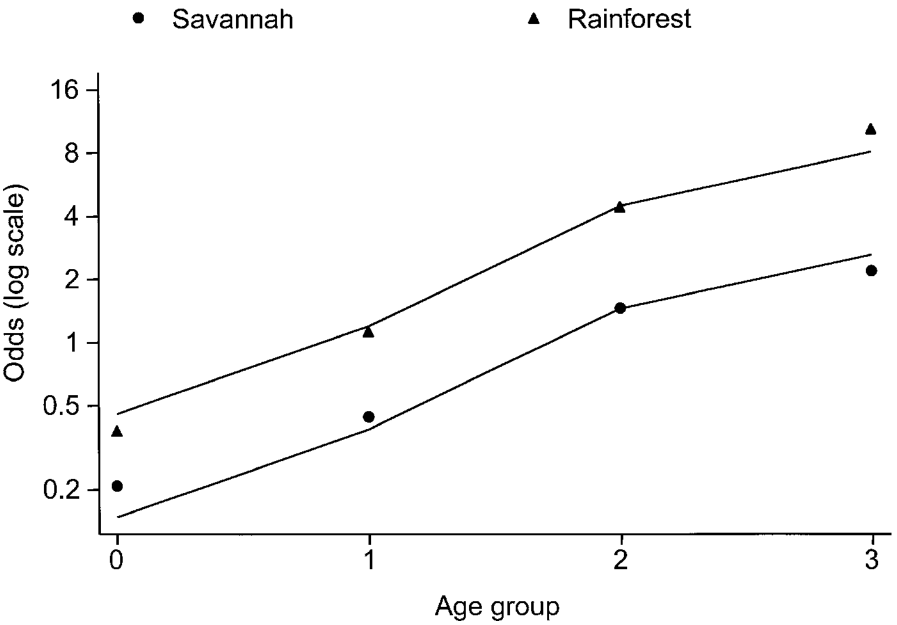

Fig. 20.1 Observed odds of *mf* infection in the eight area–age subgroups, together with lines showing the predicted odds from the logistic regression model defined in Table 20.2(b).

the crude odds ratio of 2.413. Thus the confounding effect of age meant that the crude odds ratio for area was too small.

We can use the parameter estimates shown in Table 20.5 to calculate the *predicted odds* in each group, using the equations for the odds in this logistic regression model, shown in Table 20.2(b). These calculations are shown in Table 20.6. Figure 20.1 compares the *observed* odds of *mf* infection in the eight area–age subgroups (shown in Table 20.3) with the *predicted* odds from the logistic regression model (shown by separate lines for the savannah and rainforest). The odds are plotted on a log scale; this means that, since the model *assumes* that the area odds ratios are the same in each age group, the two lines showing the predicted odds are *parallel*.

Table 20.6 Odds of *mf* infection by area and age group, as estimated from the logistic regression model.

	Odds of *mf* infection	
Age group	Savannah areas	Rainforest areas
0 (5–9 years)	0.147	0.147 × 3.083 = 0.453
1 (10–19 years)	0.147 × 2.599 = 0.382	0.147 × 3.083 × 2.599 = 1.178
2 (20–39 years)	0.147 × 9.765 = 1.435	0.147 × 3.083 × 9.765 = 4.426
3 ($\geq$ 40 years)	0.147 × 17.64 = 2.593	0.147 × 3.083 × 17.64 = 7.993

20.3 TESTING FOR INTERACTION, AND MORE COMPLEX LOGISTIC REGRESSION MODELS

We have explained the interpretation of logistic regression models for one and two variables in great detail. The extension to models for more than two variables is straightforward, and the interpretation of results follows the same principles. Regression modelling, including hypothesis testing, examining interaction between variables and modelling dose–response relationships, is described in more detail in Chapter 29. For now we note two important points:

1 In the logistic regression model for two variables (area and age group) described above, we assumed that the effect of each was the same regardless of the level of the other. In other words, we assumed that there was no **interaction** between the effects of the two variables. Interaction (also known as **effect modification**) was described in Chapter 18. It is straightforward to use regression modelling to examine this; see Section 29.5 for details.

2 Similarly, when we include three or more variables in a logistic regression model, we assume that there is no interaction between any of them. On the basis of this assumption, we estimate the effect of each, controlling for the effect of all the others.

More information about logistic regression models may be found in Hosmer and Lemeshow (2000).

20.4 REGRESSION ANALYSIS OF RISK RATIOS

Most regression analyses of binary outcomes are conducted using odds ratios: partly because of the mathematical advantages of analyses based on odds ratios (see Section 16.6) and partly because computer software to do logistic regression analyses is so widely available. However, it is straightforward to do regression analyses of risk ratios, if it is considered important to express exposure effects in that way.

This is carried out by relating the effect of the exposure variable(s) to the log of the risk of the outcome rather than the log of the odds, using a statistical software package that allows the user to fit **generalized linear models** (see Chapter 29) for a range of outcome distributions and a range of what are known as **link functions**. For logistic regression the outcome variable is assumed to have a binomial distribution (see Chapter 15) and the link function is the logit function $\text{logit}(\pi) = \log[\pi/(1-\pi)]$ (see Section 19.3). To model exposure effects as risk ratios instead of odds ratios, we simply specify a log link function ($\log \pi$) instead of a logit link function. The outcome distribution is still binomial. The model is:

$$\log(\text{risk of outcome}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

If the outcome is rare then odds ratios are approximately the same as risk ratios (see Section 16.6) and so the choice of odds ratio or risk ratio as the measure of exposure effect is unimportant. When the outcome is common, the two measures are different, and as stated in Section 16.6, it is important that odds ratios are not misinterpreted as risk ratios. The problem with the regression analysis of risk ratios is that when the outcome is common, it can prove difficult to fit models based on risk ratios, because they are constrained (see Section 16.6); this means that computer model-fitting routines often fail to produce results. Furthermore, exposure effects will differ depending on whether the presence *or* absence of the outcome event is considered as the outcome. For these reasons, it is likely that logistic regression will continue to be the method of choice for the regression analysis of binary outcome variables.

20.5 OUTCOMES WITH MORE THAN TWO LEVELS

Finally, we briefly describe extensions to logistic regression that may be used for categorical outcomes with more than two categories. In Chapter 2 we distinguished between categorical variables such as ethnic group, for which there is no natural ordering of the categories, and **ordered categorical** variables such as social class, in which the different categories, though non-numerical, have a natural ordering. We will briefly introduce the regression models appropriate for each of these types of outcome variable. We will denote the outcome variable by y, and assume that y has k possible categories.

Multinomial logistic regression

Multinomial logistic regression, also known as **polychotomous logistic regression**, extends logistic regression by estimating the effect of one or more exposure variables on the probability that the outcome is in a particular category. For example, in a study of risk factors for asthma the outcome might be defined as no asthma, allergic asthma and non-allergic asthma. One of the outcome levels is chosen as the comparison level, and $(k - 1)$ regression coefficients, corresponding to each other outcome level, are estimated for each exposure variable in the regression model. If there are only two outcome levels the model is identical to standard logistic regression. However, when the outcome has *more than* two levels, interpretation of the regression coefficients is less straightforward than for logistic regression, because the estimated effect of an exposure variable is measured by the combined effects of $(k - 1)$ regression coefficients.

Ordinal logistic regression

Ordinal logistic regression is an extension of logistic regression which is appropriate when the outcome variable is *ordered categorical*. For example, in a study of risk factors for malnutrition the outcome might be classified as severe, moderate, mild, or no malnutrition. The most commonly used type of model is the **proportional odds** model, whose parameters represent the exposure odds ratios for being in the highest j categories compared to the lowest $(k - j)$ categories. For example, if there were four outcome categories and a single exposure variable, then the exposure odds ratio would represent the combined comparison of outcome: category 4 with categories 3, 2 and 1, categories 4 and 3 with categories 2 and 1, and categories 4, 3 and 2 with category 1. It is assumed that the effect of exposure is the same for all such splits of the categories of the outcome variable. Some statistical software packages provide tests of this assumption, others do not.

Other, less commonly used models for ordered categorical outcome variables include the **continuation ratio model** and the **stereotype model**.

Further reading

Regression models for categorical variables with more than two levels are described by Agresti (1996). Models for ordered categorical outcome variables have been reviewed by Armstrong and Sloan (1989), and Ananth and Kleinbaum (1997).

CHAPTER 21

Matched studies

21.1 INTRODUCTION

In this chapter we introduce methods for studies in which we have binary outcome observations that are **matched** or **paired** in some way. The two main reasons why matching occurs are:

1 When the outcome is observed on the *same* individual on two separate occasions, under different exposure (or treatment) circumstances, or using two different methods.

2 The study has used a **matched design** in selecting individuals. This mainly occurs with **case–control studies**; each case (subjects with the disease) is matched with one or more controls (subjects without the disease), deliberately chosen to have the same values for major confounding variables. For example, controls might be selected because they are of similar age to a case, or because they live in the same neighbourhood as the case. We will discuss case–control studies in more detail in Chapter 34, where we will see that matched designs often have few advantages, and may have serious disadvantages, compared to unmatched designs. It is also *very occasionally* used in **clinical trials**, for example in a trial comparing two treatments for an eye condition, the two treatments may be randomly assigned to the left and right eyes of each patient.

It is essential that the matching be allowed for in the analysis of such studies.

21.2 COMPARISON OF TWO PROPORTIONS: PAIRED CASE

Example 21.1

Consider the results of an experiment to compare the Bell and Kato–Katz methods for detecting *Schistosoma mansoni* eggs in faeces in which two subsamples from each of 315 specimens were analysed, one by each method. Here, the exposure is the type of method, and the outcome is the test result. The correct way to analyse such data is to consider the results of each *pair* of subsamples. For any pair there are four

Table 21.1 Possible results when a pair of subsamples is tested using two methods for detecting *Schistosoma mansoni* eggs.

	Notation	Description
Both tests positive		Concordant pairs
Both tests negative		
Bell positive, Kato–Katz negative	*r*	Discordant pairs
Kato–Katz positive, Bell negative	*s*	

possible outcomes, as shown in Table 21.1. The results for each of the 315 specimens (pairs of subsamples) are shown in Table 21.2(a). Note that it would be incorrect to arrange the data as in Table 21.2(b) and to apply the standard chi-squared test, as this would take no account of the *paired* nature of the data, namely that it was the *same* 315 specimens examined with each method, and not 630 different ones.

One hundred and eighty-four specimens were positive with both methods and 63 were negative with both. These 247 specimens (the **concordant pairs**; see Table 21.1) therefore give us no information about which of the two methods is better at detecting *S. mansoni* eggs. The information we require is entirely contained in the 68 specimens for which the methods did not agree (the **discordant pairs**). Of these, 54 were positive with the Bell method only, compared to 14 positive with the Kato–Katz method only.

Table 21.2 Comparison of Bell and Kato–Katz methods for detecting *Schistosoma mansoni* eggs in faeces. The same 315 specimens were examined using each method. Data from Sleigh *et al*. (1982) *Transactions of the Royal Society of Tropical Medicine and Hygiene* **76**: 403–6 (with permission).

(a) Correct layout.

		Kato–Katz		
		+	−	Total
Bell	+	184	54(*r*)	238
	−	14(*s*)	63	77
Total		198	117	315

(b) Incorrect layout.

	Results		
	+	−	Total
Bell	238	77	315
Kato–Katz	198	117	315
Total	436	194	630

The proportions of specimens found positive with the two methods were 238/315 (0.756) using the Bell method and 198/315 (0.629) using the Kato–Katz method. The difference between the proportions was therefore 0.1270. This difference can also be calculated from the numbers of discordant pairs, *r* and *s*, and the total number of pairs, *n*:

$$\text{Difference between paired proportions} = \frac{r-s}{n},$$

$$\text{s.e.(difference)} = \frac{\sqrt{(r+s)}}{n}$$

In this example, the difference between the paired proportions is $(r - s)/n = (54 - 14)/315 = 0.1270$, the same as calculated above. Its standard error equals $[\sqrt{(r + s)}]/n = \sqrt{68}/315 = 0.0262$. An approximate **95% confidence interval** can be derived in the usual way:

$$\begin{aligned}95\%\ \text{CI} &= 0.1270 - (1.96 \times 0.0262) \text{ to } 0.1270 + (1.96 \times 0.0262)\\ &= 0.0756 \text{ to } 0.1784\end{aligned}$$

With 95% confidence, the positivity rate is between 7.6% and 17.8% higher if the Bell method is used to detect *S. mansoni* eggs than if the Kato–Katz method is used.

z-test for difference between proportions

If there was no difference in the abilities of the methods to detect *S. mansoni* eggs, we would not of course expect complete agreement since different subsamples were examined, but we would expect on average half the disagreements to be positive with the Bell method only and half to be positive with the Kato–Katz method only. Thus an appropriate test of the null hypothesis that there is no difference between the methods is to compare the proportion found positive with the Bell method only, namely 54/68, with the hypothetical value of 0.5. This may be done using the *z test*, as described in Section 15.6. As usual, we construct the test by dividing the difference by its standard error assuming the null hypothesis to be true, which gives:

$$z = \frac{54/68 - 0.5}{\sqrt{(0.5 \times 0.5/68)}} = 4.85,\ P < 0.001$$

There is strong evidence that the Bell method is more likely to detect *S. mansoni* eggs than the Kato–Katz method. (Note that other than for the sign of the *z* statistic exactly the same result would have been obtained had the proportion positive with the Kato–Katz method only, namely 14/68, been compared with 0.5.)

21.3 USING ODDS RATIOS FOR PAIRED DATA

An alternative approach to the analysis of matched pairs is to estimate the odds ratio comparing the Bell and Kato–Katz methods. Again, our analysis must take the pairing into account. This can be done using **Mantel–Haenszel** methods (see Section 18.4), with the data stratified into the individual pairs. Using the same notation as in Chapter 18, the notation for the *i*th pair is shown in Table 21.3. The Mantel–Haenszel estimate of the odds ratio (see Chapter 18) is given by:

$$\text{OR}_{MH} = \frac{\Sigma \dfrac{d_{1i} \times h_{0i}}{n_i}}{\Sigma \dfrac{d_{0i} \times h_{1i}}{n_i}}$$

Table 21.3 Notation for the 'stratified' 2 × 2 table giving the results for pair *i*.

	Outcome		
	+	−	Total
Bell method	d_{1i}	h_{1i}	1
Kato–Katz method	d_{0i}	h_{0i}	1
Total	d_i	h_i	2

As in the last section, the analysis can be simplified if we note that there are only four possible outcomes for each pair, and therefore only four possible types of 2 × 2 table. These are shown in Table 21.4, together with their contributions to the numerator and denominator in the formula for the Mantel–Haenszel OR. This shows that, again, only the discordant pairs contribute to the Mantel–Haenszel estimate of the odds ratio. The total for the numerator is *r*/2, while the total for the denominator is *s*/2. The estimated **odds ratio** is therefore:

$$\text{OR}_{MH} = \frac{r/2}{s/2} = \frac{r}{s}, \text{the ratio of the numbers of discordant pairs}$$

Table 21.4 Possible outcomes for each pair, together with their contributions to the numerator and denominator in the formula for the Mantel–Haenszel estimate of the odds ratio.

	Concordant pairs				Discordant pairs			
	+	−	+	−	+	−	+	−
Bell	1	0	0	1	1	0	0	1
Kato–Katz	1	0	0	1	0	1	1	0
Number of pairs					*r*		*s*	
$\frac{d_{1i} \times h_{0i}}{n_i}$	0		0		½		0	
$\frac{d_{0i} \times h_{1i}}{n_i}$	0		0		0		½	

An approximate **95% error factor** for the odds ratio is given by:

$$\text{EF} = \exp[1.96 \times \surd(1/\text{r} + 1/\text{s})]$$

In the example, the estimated odds ratio is given by $54/14 = 3.857$, while the error factor is $\exp[1.96 \times \surd(1/54 + 1/14)] = 1.800$. The approximate **95% confidence interval** is therefore given by:

$$95\%\ \text{CI} = \text{OR}/\text{EF} \text{ to OR} \times \text{EF} = 3.857/1.800 \text{ to } 3.857 \times 1.800 = 2.143 \text{ to } 6.943$$

McNemar's chi-squared test

A chi-squared test, based on the numbers of discordant pairs, can also be derived from the formula for the Mantel–Haenszel statistic presented in Chapter 18 and is given by:

$$\chi^2_{\text{paired}} = \frac{(r-s)^2}{r+s}, \text{d.f.} = 1$$

This is known as **McNemar's chi-squared test**. In the example $\chi^2 = (54 - 14)^2 / (54 + 14) = 40^2/68 = 23.53$, d.f. $= 1, P < 0.001$. Apart from rounding error, this χ^2 value is the same as the square of the z value obtained above ($4.85^2 = 23.52$), the two tests being mathematically equivalent.

Validity

The use of McNemar's chi-squared test or the equivalent z test is valid provided that the total number of discordant pairs is at least 10. The approximate error factor for the 95% CI for the odds ratio is valid providing that the total number of pairs is greater than 50. If these conditions are not met then methods based on exact binomial probabilities should be used (these are described by Alman *et al.* 2000).

21.4 ANALYSING MATCHED CASE–CONTROL STUDIES

The methods described above can also be used for the analysis of case–control studies and clinical trials which have employed a matched design, as described in the introduction. The rationale for this and the design issues are discussed in more detail in Chapter 34.

Example 21.2

Table 21.5 shows data from a study to investigate the association between use of oral contraceptives and thromboembolism. The cases were 175 women aged 15–44 discharged alive from 43 hospitals after initial attacks of thromboembolism. For each case a female patient suffering from some other disease (thought to be unrelated to the use of oral contraceptives) was selected from the same hospital to act as a control. She was chosen to have the same residence, time of hospitalisation, race, age, marital status, parity, and income status as the case. Participants were questioned about their past contraceptive history, and in particular

Table 21.5 Results of a *matched* case–control study, showing the association between use of oral contraceptives (OC) and thromboembolism. With permission from Sartwell *et al*. (1969) *American Journal of Epidemiology* **90**: 365–80.

		Controls		
		OC used	OC not used	Total
Cases	OC used	10	57	67
	OC not used	13	95	108
	Total	23	152	175
		OR = 57/13 = 4.38		

about whether they had used oral contraceptives during the month before they were admitted to hospital.

The pairing of the cases and controls is preserved in the analysis by comparing oral contraceptive use of each case against oral contraceptive use of their matched control. There were ten case–control pairs in which both case and control had used oral contraceptives and 95 pairs in which neither had. These 105 concordant pairs give no information about the association. This information is entirely contained in the 70 discordant pairs in which the case and control differed. There were 57 case–control pairs in which only the case had used oral contraceptives within the previous month compared to 13 in which only the control had done so. The odds ratio is measured by the ratio of these **discordant pairs** and equals 4.38, which suggests oral contraceptive use leads to a substantial increase in the risk of thromboembolism.

$$\begin{aligned}\text{OR} &= \text{ratio of discordant pairs}\\ &= \frac{\text{no. of pairs in which case exposed, control not exposed}}{\text{no. of pairs in which control exposed, case not exposed}}\end{aligned}$$

The *error factor* is $\exp[1.96 \times \sqrt{(1/57 + 1/13)}] = 1.827$. The 95% CI for the odds ratio is therefore $4.38/1.827$ to 4.38×1.827, which is 2.40 to 8.01. McNemar's χ^2 test gives: $\chi^2 = (57 - 13)^2/(57 + 13) = 27.7, P < 0.001$, corresponding to strong evidence against the null hypothesis that there is no association.

If *several* controls rather than a single matched control are selected for each case, the odds ratio can still be estimated by using Mantel–Haenszel methods. However, these methods are severely limited because they do not allow for further stratification on confounding variables which were not also matching variables. The solution to this problem is to use **conditional logistic regression**, which we describe next.

21.5 CONDITIONAL LOGISTIC REGRESSION

In general when analysing individually matched case–control studies we may wish to control for confounding variables, additional to those matched for in the design. This is done using **conditional logistic regression**, a variant of logistic regression in which cases are only compared to controls in the same matched set. In the simple case of individually-matched case–control studies with one control per case and no further confounders, conditional logistic regression will give identical results to the methods for paired data described earlier in the chapter. However, additional confounders may be included in the model, and there is no restriction on the numbers of cases and controls in each matched set.

Once the reader is familiar with the use of logistic regression, then conditional logistic regression should present no additional difficulties. The only difference is that in addition to the outcome and exposure variables, the computer software requires a variable that specifies which case (or cases) matches which control (or controls). Exposure effects are estimated by considering possible combinations of exposures, conditional on the observed exposures *within each matched set*. For example, if the set consists of one case and two controls, with only one of the set exposed and the other two unexposed, then the three possible combinations are:

	Case	Control 1	Control 2
1	Exposed	Unexposed	Unexposed
2	Unexposed	Exposed	Unexposed
3	Unexposed	Unexposed	Exposed

It is because the possible combinations are conditional on the total number of exposed and unexposed individuals in each matched set that the method is called *conditional* logistic regression. This argument extends in a straightforward manner to numeric exposure variables and to more than one exposure variable.

Example 21.3

Table 21.6 shows data from a matched case–control study of risk factors for infant death from diarrhoea in Brazil [Victora *et al.* (1987) *Lancet* **ii**: 319–322], in which an attempt was made to ascertain all infant deaths from diarrhoea occurring over a one-year period in two cities in southern Brazil, by means of weekly visits to all hospitals, coroners' services and death registries in the cities. Whenever the underlying cause of death was considered to be diarrhoea, a physician visited the parents or guardians to collect further information about the terminal illness, and data on possible risk factors. The same data were collected for two 'control' infants. Those chosen were the nearest neighbour aged less than 1 year, and the next nearest neighbour aged less than 6 months. This procedure was designed to provide a control group with a similar socio-economic distribution to that of the cases. The selection also ensures

Table 21.6 First 24 lines (eight case–control sets) of the dataset for the matched case–control study of risk factors for infant death from diarrhoea in southern Brazil. Reproduced with kind permission of C.G. Victora.

Observation number	*case*	*set*	*water*	*agegp*	*bwtgp*	*social*	*income*
1	1	1	0	2	3	1	3
2	0	1	1	3	4	2	2
3	0	1	1	2	3	1	3
4	1	2	1	1	2	1	2
5	0	2	1	3	4	2	3
6	0	2	1	2	4	1	2
7	1	3	1	2	3	2	2
8	0	3	1	5	3	2	4
9	0	3	1	1	3	2	4
10	1	4	1	3	3	1	2
11	0	4	1	4	3	1	3
12	0	4	1	2	4	1	2
13	1	5	1	2	2	2	2
14	0	5	1	4	2	2	2
15	0	5	1	1	2	2	3
16	1	6	1	2	3	2	2
17	0	6	1	4	4	1	2
18	0	6	1	2	3	1	2
19	1	7	1	2	1	1	2
20	0	7	1	4	3	1	2
21	0	7	1	2	4	1	2
22	1	8	1	3	3	1	3
23	0	8	1	5	2	1	2
24	0	8	1	2	4	1	1

that there are approximately twice as many controls less than 6 months old, as between 6–11 months; this matches what was known concerning the age distribution of the cases. During the one-year study period, data were collected on 170 cases together with their 340 controls. In addition to variable *case* ($1 =$ case, $0 =$ control), the dataset contains a variable *set* which gives the number (from 1 to 170) of the set to which each case and its two matched controls belong. Table 21.6 contains the first 24 lines (eight case–control sets) of this dataset.

Variable *water* denotes whether the child's household had access to water in their house or plot ($water = 1$) or not ($water = 0$). Variable *agegp* (age group) is coded as $1 = 0{-}1$ months, $2 = 2{-}3$ months, $3 = 4{-}5$ months, $4 = 6-8$ months and $5 = 9-11$ months. Variable *bwtgp* (birth weight group, kg) has values $1 = 1.50{-}2.49$, $2 = 2.50{-}2.99$, $3 = 3.00{-}3.49, 4 = \geq\ 3.50$ kg. The final two variables are *social* (household social group) from 1 (most deprived) to 3 (least deprived), and *income* (household income group) from 1 (least monthly income) to 4 (most monthly income).

Examining the effect of a single exposure variable

A total of 111 (65.3%) cases and 259 (76.2%) controls had access to water, suggesting that access to water might be protective against infant death from diarrhoea. Since this is a *matched* case–control study, the calculation of the odds ratio for this exposure and all other analyses must take into account the matching. Using Mantel–Haenszel methods stratified by set (170 strata, each containing 1 case and 2 controls) gives an estimated odds ratio of 0.275 (95% CI = 0.136 to 0.555). Access to water thus appears to be strongly protective against infant diarrhoea death. Table 21.7 shows corresponding output from a conditional logistic regression model (also stratifying on set for the effect of household water supply). The estimated odds ratio is similar to that derived using Mantel–Haenszel methods.

Table 21.7 Conditional logistic regression output (odds ratio scale) for the association between household water supply and infant diarrhoea death in southern Brazil.

	Odds ratio	z	$P > \|z\|$	95% CI
Water	0.2887	−3.67	0.000	0.1487 to 0.5606

A possible alternative approach to the analysis of such data is to fit a standard logistic regression model, incorporating an indicator variable in the model corresponding to each case–control set, as a way of controlling for the matching. It is important to note, however, that for finely matched data *this will give the wrong answer*, and that the odds ratios obtained will be further away from the null value of 1 than they should be. For data in which the sets consist of exactly one case and one control, the estimated odds ratio from such a model will be exactly the square of the odds ratio estimated using Mantel–Haenszel methods stratified by set, or using conditional logistic regression.

Controlling for confounders, additional to those used for matching

Since access to water may be associated with a household's social status, we may wish to control additionally for the effects of variables such as *social* and *income*. Because there are only three subjects in each stratum, further stratification using Mantel–Haenszel methods is not feasible. However, conditional logistic regression allows us to control for the effects of confounding variables in addition to those used in the matching. Table 21.8 shows output from a conditional logistic regression model, controlling for the effects of all the variables in Table 21.6. Here, *agegp*(2) is an indicator variable (see Section 19.4) which takes the value 1 for infants in age group 2 and 0 for infants in other age groups. However, the corresponding odds ratio of 2.6766 cannot be interpreted as the odds of death in age group 2 compared to age group 1, because age was used in the matching of cases to controls. The odds ratio for the effect of water is only slightly increased (closer to the null value of 1), so we would conclude that the additional variables

Table 21.8 Conditional logistic regression output (odds ratio scale) for the association between household water supply and infant diarrhoea death in southern Brazil, controlling for the effects of potentially confounding variables.

	Odds Ratio	z	$P > \|z\|$	95% CI
water	0.2991	−3.20	0.001	0.1427 to 0.6269
agegp(2)	2.6766	2.89	0.004	1.3719 to 5.2222
agegp(3)	2.4420	2.50	0.012	1.2121 to 4.9199
agegp(4)	3.2060	3.27	0.001	1.5940 to 6.4482
agegp(5)	0.8250	−0.43	0.666	0.3444 to 1.9758
bwtgp(2)	0.4814	−2.00	0.045	0.2354 to 0.9844
bwtgp(3)	0.4111	−2.52	0.012	0.2061 to 0.8199
bwtgp(4)	0.3031	−3.12	0.002	0.1431 to 0.6422
social(2)	0.9517	−0.21	0.830	0.6058 to 1.4951
social(3)	0.1527	−1.78	0.075	0.0192 to 1.2128
income(2)	0.7648	−0.85	0.394	0.4128 to 1.4170
income(3)	0.6970	−1.01	0.312	0.3459 to 1.4043
income(4)	0.6991	−0.86	0.389	0.3098 to 1.5774

included in the model had only a slight confounding effect, and that there is still a clear protective effect of having a water supply in a household.

This page intentionally left blank