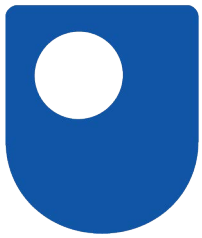

M249

Practical modern statistics

Medical statistics

About this module

M249 *Practical modern statistics* uses the software packages *IBM SPSS Statistics* (SPSS Inc.) and *WinBUGS*, and other software. This software is provided as part of the module, and its use is covered in the *Introduction to statistical modelling* and in the four computer books associated with *Books 1* to *4*.

This publication forms part of an Open University module. Details of this and other Open University modules can be obtained from the Student Registration and Enquiry Service, The Open University, PO Box 197, Milton Keynes MK7 6BJ, United Kingdom (tel. +44 (0)845 300 60 90; email general-enquiries@open.ac.uk).
Alternatively, you may visit the Open University website at www.open.ac.uk where you can learn more about the wide range of modules and packs offered at all levels by The Open University.
To purchase a selection of Open University materials visit www.ouw.co.uk, or contact Open University Worldwide, Walton Hall, Milton Keynes MK7 6AA, United Kingdom for a brochure (tel. +44 (0)1908 858779; fax +44 (0)1908 858787; email ouw-customer-services@open.ac.uk).

Note to reader
Mathematical/statistical content at the Open University is usually provided to students in printed books, with PDFs of the same online. This format ensures that mathematical notation is presented accurately and clearly. The PDF of this extract thus shows the content exactly as it would be seen by an Open University student. Please note that the PDF may contain references to other parts of the module and/or to software or audio-visual components of the module. Regrettably mathematical and statistical content in PDF files is unlikely to be accessible using a screenreader, and some OpenLearn units may have PDF files that are not searchable. You may need additional help to read these documents.

The Open University, Walton Hall, Milton Keynes MK7 6AA.

First published 2007.

Printed in the United Kingdom by The Charlesworth Group, Wakefield.

ISBN 978 0 7492 1366 4
1.1

Contents

Cohort studies and case-control studies

Introduction

The aim of an epidemiological investigation is usually to study the impact on health of one or more potential risk factors. Most often, studies focus on a particular health outcome. This could be a disease, for example colon cancer, or a consequence of disease, such as death from colon cancer, or, more widely, any event with a bearing on health, such as infection or accident. The risk factor is any variable that might be associated with the outcome. This could be age, socio-economic status, exposure to a toxic chemical, diet, and so on. The purpose of the investigation is to determine whether exposure to the potential risk factor is associated with an increase or a decrease in the frequency of the health outcome of interest.

An epidemiological study can help quantify the evidence for or against an association between exposure to a risk factor and occurrence of a disease. However, it cannot generally determine whether an association is causal. The issue of causality is discussed in more detail in Parts II and III, but it is important to be clear from the outset about the limitations of epidemiological investigations.

Association does not imply causation.

The two most commonly used types of studies in epidemiology are **cohort studies** and **case-control studies**. In Section 1, you will learn about cohort studies, and how to quantify the strength of an association. In Section 2, the binomial model for cohort studies is described, and used to calculate confidence intervals for different measures of association. A different type of study is introduced in Section 3 — the case-control study. In Section 4, the chi-squared test for no association is discussed.

1 Cohort studies

In this section, you will learn about cohort studies. In Subsection 1.1, cohort studies are described with several illustrative examples. In Subsection 1.2, two commonly used measures of association are introduced — the relative risk and the odds ratio; their use is discussed in Subsection 1.3.

1.1 What is a cohort study?

Epidemiological investigations aim to study the association, if any, between an exposure to a risk factor and a health outcome or disease. For simplicity, throughout this book we refer to the exposure E and the disease D. The key feature of a cohort study is that it comprises one or several groups of individuals, who are followed over time. The most frequently used type of cohort study is the *controlled cohort study*. This will be introduced using the study described in Example 1.1.

Example 1.1 Hypertension in later life

Some pregnant women are affected by severe gestational hypertension (that is, high blood pressure) known as pre-eclampsia and eclampsia. In a study to investigate the long-term effects of these problems, the proportions of women with hypertension later in life were compared in two groups. One group included 542 women who suffered from gestational hypertension during their first pregnancy. This is the **exposed group**. The other group included 277 women of similar ages to those in the first group, but who did not suffer from these conditions during their first pregnancy. This is the **control group**. ♦

Wilson, B.J., Watson, M.S., Prescott, G.J. *et al.* (2003) Hypertensive diseases of pregnancy and risk of hypertension and stroke in later life: results from cohort study. *British Medical Journal*, **326**, 845–849.

In Example 1.1, the exposure E is pre-eclampsia or eclampsia during the first pregnancy, and the disease D is hypertension in later life. If experiencing pre-eclampsia or eclampsia during the first pregnancy increases the chance of hypertension later in life then, provided the two groups are similar in other respects, a higher proportion of women would be expected to develop hypertension in the exposed group than in the control group.

This study is a typical example of a **cohort study** of the association, if any, between an exposure E and a disease D. The key features of a cohort study are as follows.

Cohort studies

A **cohort study** to investigate the association between an exposure E and a disease D has the following features.

- ◇ It includes one group with the exposure E and a comparable control group without exposure E.
- ◇ The groups are followed over time and the occurrences of disease D in each group are identified.

The word 'cohort' conveys the idea of a group of individuals marching forward. In the context of epidemiology, the group marches forward through time. In Example 1.1 the cohort comprises the $542 + 277 = 819$ women in the study. An important aspect of a cohort study from a statistical perspective is that the exposure status (E or not E) of each individual is treated as fixed, whereas the disease outcome (D or not D) is not known in advance and is regarded as random.

In Roman times a cohort was a military unit, typically comprising 480 soldiers.

This type of cohort study is more precisely a **controlled** cohort study, since it includes a control group comprising individuals who do not have exposure E but are in other respects comparable to the exposed group.

Cohort studies without a separate control group are appropriate in some special settings, but will not be considered here.

There are many variants on this basic design. For example, a cohort study may include more than one exposed group. Thus, in a study of the effect of environmental pollutants on asthma it might be appropriate to study exposures to several different pollutants. Such studies will be discussed in Subsection 3.3. Similarly, a single cohort may be used to study several outcomes.

The data from a cohort study can be arranged in a simple table. The hypertension data for the study described in Example 1.1 are shown in Table 1.1. Note that the table contains not just the numbers, but also concise labelling.

Table 1.1 Gestational pre-eclampsia or eclampsia and hypertension in later life

Exposure category	Hypertension	No hypertension	Total
Pre-eclampsia or eclampsia	327	215	542
No pre-eclampsia or eclampsia	76	201	277

Activity 1.1 Serious self-inflicted injury and compulsory redundancy

A study in New Zealand investigated the relationship between involuntary redundancy and serious self-inflicted injury (such as suicide attempts) leading to hospitalization or death. Two groups of workers were compared. One group comprised 1945 workers made compulsorily redundant from the Whakatu meat-processing plant in the Hawkes Bay region of New Zealand. This plant was closed down in 1986. The other group comprised 1767 workers from the neighbouring Tomoana meat-processing plant that remained open until 1994. The workforces of the two plants were similar in terms of age, sex, ethnicity, and duration of employment. The numbers of workers who were hospitalized or died due to self-inflicted injury between 1986 and 1994 were obtained. There were 14 such cases among the Whakatu workers and 4 among the Tomoana workers.

Keefe, V., Reid, P., Ormsby, C. *et al.* (2002) Serious health events following involuntary job loss in New Zealand meat processing workers. *International Journal of Epidemiology*, **31**, 1155–1161.

(a) Identify the exposure E and the 'disease' D in this study.

(b) Identify the exposed group and the control group.

(c) Set out the data from the study in a table similar to Table 1.1, taking care to provide clear labelling in the title, and in the row and column headings.

The term 'disease' is used here to mean an adverse state of health.

The cohort studies considered here are *comparative* studies, in that they are designed to provide a comparison between the exposure group and an unexposed control group. If there is no association between exposure E and disease D, then the underlying probabilities of disease D in the exposed group and the control group are the same, and the sample proportions, that is the proportions actually observed, will differ by an amount consistent with chance variation. If, on the other hand, there is a positive association between E and D, then the underlying probability of disease D will be higher in exposed individuals than in unexposed individuals, and the sample proportions are likely to reflect this difference. Similarly, if there is a negative association between E and D, then the underlying probability of disease D will be lower in exposed individuals than in unexposed individuals.

The probability of disease in exposed individuals is denoted $P(D|E)$. The vertical bar indicates that this is a *conditional* probability, in this case, the underlying probability of disease, *given* that the individual is exposed. Similarly, the probability of disease in unexposed individuals is written $P(D|\text{not } E)$, where 'not E' means 'not exposed'.

In general, data from a cohort study may be presented as in Table 1.2. The column corresponding to 'no disease' has been labelled 'not D'.

Table 1.2 A general data table for a cohort study

Exposure category	Disease outcome D	not D	Total
E (exposed group)	a	b	$n_1 = a + b$
not E (control group)	c	d	$n_2 = c + d$

The probabilities of disease in the exposed group and the control group are estimated by the sample proportions:

$$\widehat{P}(D|E) = \frac{a}{n_1}, \quad \widehat{P}(D|\text{not } E) = \frac{c}{n_2}.$$

The hat symbol is used to distinguish the sample estimates from the underlying probabilities.

Example 1.2 Hypertension in later life: results from cohort study

For the hypertension data of Table 1.1, the proportion of the exposed group with hypertension in later life is

$$\widehat{P}(D|E) = \frac{a}{n_1} = \frac{327}{542} \simeq 0.60.$$

This is an estimate of $P(D|E)$, the underlying probability of disease (in this case, hypertension in later life), given exposure (which in this case is pre-eclampsia or eclampsia during the first pregnancy). The proportion of the control group with hypertension in later life is

$$\widehat{P}(D|\text{not } E) = \frac{c}{n_2} = \frac{76}{277} \simeq 0.27.$$

This is an estimate of $P(D|\text{not } E)$, the underlying probability of disease, given no exposure. Since the proportion $\widehat{P}(D|E)$ is greater than the proportion $\widehat{P}(D|\text{not } E)$, this may be evidence that hypertension later in life is associated with pre-eclampsia or eclampsia during the first pregnancy. However, the difference may be due to random variation. Further analysis is required to establish whether there is a link. ♦

Activity 1.2 Seat belts and children's safety in car accidents

In many countries, babies and small children must be secured in specially designed seats when travelling by car. Older children, however, must use the same belts as adults. Concern has been expressed that these belts might injure children owing to the immature anatomy of a child's pelvis. In a study from Canada, data were obtained on 85 children aged between 4 and 14 years sitting in the left-hand back seat (behind the driver) of cars involved in serious accidents. The numbers sustaining at least moderately severe injury among children who were wearing seat belts at the time of the accident and among children who were not wearing seat belts were counted. The data are shown in Table 1.3.

Halman, I., Chipman, M., Parkin, P.C. and Wright, J.G. (2002) Are seat belt restraints as effective in school age children as in adults? A prospective crash study. *British Medical Journal*, **324**, 1123–1125.

Table 1.3 Seat belt use and injury sustained by children aged 4–14

Exposure category	D: sustained at least moderately severe injury Yes	No	Total
Not wearing a seat belt (E)	14	19	33
Wearing a seat belt (not E)	13	39	52

(a) Estimate $P(D|E)$ and $P(D|\text{not } E)$ from Table 1.3, where E is 'not wearing a seat belt'.

(b) Is it correct to conclude from these estimates alone that $P(D|E)$ is greater than $P(D|\text{not } E)$, and hence that not wearing a seat belt is associated with sustaining severe injury in the event of a car crash?

It may not be immediately clear to you that the study in Activity 1.2 is a cohort study, since the population studied consists of children aged 4–14 who were involved in a serious car accident, rather than 'followed over time'. One way to think of it is that the two groups to be compared are specified according to the exposure status of the children (wearing a seat belt or not wearing a seat belt) just before being involved in a serious accident. At this point their exposure status is known, but the outcome is not. The data for this study were assembled after the accidents had occurred, but are analysed so as to recreate the follow-up, like a video being wound back and replayed. This type of study is called a **retrospective** cohort study and is commonly used in epidemiology.

1.2 Measures of association

Intuitively, it is clear that the stronger the association between an exposure E and a disease D, the larger will be the difference between the probabilities of disease with and without the exposure. A measure of association that is commonly used in epidemiology is the **relative risk** RR, which is defined as follows:

$$RR = \frac{P(D|E)}{P(D|\text{not } E)}.$$

In medical statistics, the terms 'risk' and 'probability' are to a large extent interchangeable. Thus, for example, the probability $P(D|E)$ is commonly referred to as the **risk** of disease given exposure. However, note that RR is called the relative *risk* rather than the relative probability. Since the value of RR can be greater than 1, the relative risk RR is *not* a probability.

If there is no association between E and D, then $P(D|E) = P(D|\text{not } E)$, and hence $RR = 1$. Values of RR greater than 1 correspond to **positive** associations, in which presence of the exposure E is associated with an *increase* in the disease risk. Values of RR less than 1 correspond to **negative** associations, in which the presence of the exposure E is associated with a *decrease* in the disease risk.

The relative risk RR is estimated by substituting the sample proportions for the probabilities. Using the notation in Table 1.4, which was introduced in Table 1.2, the estimated relative risk is

$$\widehat{RR} = \frac{\widehat{P}(D|E)}{\widehat{P}(D|\text{not } E)} = \frac{a/n_1}{c/n_2}. \qquad (1.1)$$

Table 1.4 A general data table for a cohort study

	D	not D	Total
E	a	b	n_1
not E	c	d	n_2

Example 1.3 Seat belts and relative risk of injury

See Activity 1.2.

For the seat belt data of Table 1.3, the estimated relative risk is

$$\widehat{RR} = \frac{\widehat{P}(D|E)}{\widehat{P}(D|\text{not } E)} = \frac{a/n_1}{c/n_2} = \frac{14/33}{13/52} \simeq 1.70.$$

Thus the sample relative risk is 1.70. This value is greater than 1, suggesting a positive association between not wearing a seat belt and increased risk of at least moderately severe injury. ♦

The interpretation of the relative risk of 1.70 in Example 1.3 is that, in the event of a serious car accident, not wearing a seat belt multiplies the risk of at least moderately severe injury by 1.70. Another way of saying the same thing is that the risk of at least moderately severe injury is increased by 70%.

The relative risk is used frequently in medical statistics, especially when the underlying risks are low. However, it has some drawbacks as a measure of strength of association.

First, the interpretation of the relative risk RR depends not only on the strength of association, but also on the magnitude of the risks involved. To see this, consider, for instance, a disease outcome that occurs in the control group with probability 0.5; that is, $P(D|\text{not } E) = 0.5$. Since the probability of the outcome in the exposed group cannot be greater than 1, that is $P(D|E) \leq 1$, it follows that

$$RR = \frac{P(D|E)}{P(D|\text{not } E)} \leq \frac{1}{0.5} = 2.$$

So, however strongly the exposure is associated with the outcome, the relative risk cannot be greater than 2. In this case, $RR = 2$ must correspond to the strongest possible association. In contrast, if the probability of the disease outcome in the control group were 0.05 (say) then, by a similar argument, $RR \leq 20$, and $RR = 20$ would correspond to the strongest possible association. So, in this case, a relative risk of 2 would certainly not be regarded as the strongest possible association. Thus the interpretation of the relative risk depends on the magnitude of the risks involved.

A second reason why the relative risk is not an ideal measure of strength of association is provided by the results of Activity 1.3.

Activity 1.3 Protective effect of wearing a seat belt

In Activity 1.2 and Example 1.3, the disease D was defined as 'sustained at least moderately severe injury', and the exposure E was defined as 'not wearing a seat belt'. These definitions of D and E are, to some extent, arbitrary choices. Equally reasonably, the disease D^* could be defined as 'avoided moderate or worse injury' in the event of a serious car accident, and the exposure E^* could be defined as 'wearing a seat belt'. Thus D^* and E^* have been obtained from D and E simply by changing the labels of the disease outcomes and the exposure categories: not D is relabelled D^*, and not E is relabelled E^*. In this case, the data would be presented as in Table 1.5.

Table 1.5 Seat belt use and injury avoided by children aged 4–14

	D^*: avoided moderate or worse injury		
Exposure category	Yes	No	Total
Wearing a seat belt (E^*)	39	13	52
Not wearing a seat belt (not E^*)	19	14	33

(a) Use the data in Table 1.5 to estimate the relative risk RR as

$$\widehat{RR} = \frac{\widehat{P}(D^*|E^*)}{\widehat{P}(D^*|\text{not } E^*)}.$$

(b) Compare this estimate with the estimate of 1.70 for the relative risk that was obtained in Example 1.3.

(c) Interpret the relative risk in terms of the chance of avoiding moderate or worse injury. In your view, has the strength of association between seat belt use and moderate or worse injury changed simply by relabelling the disease outcomes and the exposure categories?

Activity 1.3 shows that the relative risk depends on the labelling of the exposure categories and the disease outcomes. If the labels of the exposure categories are switched and the labels of the disease outcomes are switched, then in general the relative risk will change, even though the association it relates to is the same. The only exception to this is when $RR = 1$: in this case, relabelling the disease outcomes and the exposure categories will still produce $RR = 1$.

Ideally, a measure of strength of association should not depend on the way the exposure categories and the disease outcomes are labelled. Such a measure does exist, and is based on the *odds* of an event. For an event A with probability $P(A)$, the **odds of event A** is written $OD(A)$ and is defined by

$$OD(A) = \frac{P(A)}{1 - P(A)}.$$

For example, if you roll a fair die with faces labelled $1, 2, \ldots, 6$, the odds of obtaining a 6, that is the odds of the die coming to rest with the face labelled 6 uppermost, is

$$OD(6) = \frac{P(6)}{1 - P(6)} = \frac{1/6}{1 - 1/6} = 0.2.$$

Odds are also commonly used for betting. In this context, the odds of obtaining a 6 would be reported as '5 to 1 against'.

Note that odds can take any non-negative value. For example, the odds of not obtaining a 6 is

$$OD(\text{not } 6) = \frac{P(\text{not } 6)}{1 - P(\text{not } 6)} = \frac{5/6}{1 - 5/6} = 5.$$

The odds of disease D given exposure E and the odds of D given no exposure E are calculated in exactly the same way:

$$OD(D|E) = \frac{P(D|E)}{1 - P(D|E)} = \frac{P(D|E)}{P(\text{not } D|E)},$$
$$OD(D|\text{not } E) = \frac{P(D|\text{not } E)}{1 - P(D|\text{not } E)} = \frac{P(D|\text{not } E)}{P(\text{not } D|\text{not } E)}.$$

A second measure of strength of association between an exposure E and a disease D is the **odds ratio**, which is denoted OR and defined by

$$OR = \frac{OD(D|E)}{OD(D|\text{not } E)} = \frac{P(D|E) \times P(\text{not } D|\text{not } E)}{P(\text{not } D|E) \times P(D|\text{not } E)}.$$

To calculate the sample odds ratio, note that, in the notation of Table 1.2,

$$\widehat{OD}(D|E) = \frac{\widehat{P}(D|E)}{\widehat{P}(\text{not } D|E)} = \frac{a/n_1}{b/n_1} = \frac{a}{b}$$

and

$$\widehat{OD}(D|\text{not } E) = \frac{\widehat{P}(D|\text{not } E)}{\widehat{P}(\text{not } D|\text{not } E)} = \frac{c/n_2}{d/n_2} = \frac{c}{d}.$$

Hence

$$\widehat{OR} = \frac{\widehat{OD}(D|E)}{\widehat{OD}(D|\text{not } E)} = \frac{a/b}{c/d} = \frac{a \times d}{b \times c}. \tag{1.2}$$

As for the relative risk, if the odds ratio is equal to 1, then there is no association between the exposure E and the disease D. Also as for the relative risk, an odds ratio greater than 1 indicates a positive association, and an odds ratio less than 1 indicates a negative association.

Example 1.4 Odds ratio of injury with and without seat belts

Taking D to denote 'sustained at least moderately severe injury' and E to denote 'not wearing a seat belt', as in Example 1.3, the sample odds are

The data are in Table 1.3.

$$\widehat{OD}(D|E) = \frac{14}{19} \simeq 0.7368$$

and

$$\widehat{OD}(D|\text{not } E) = \frac{13}{39} \simeq 0.3333.$$

So the sample odds ratio is

$$\widehat{OR} = \frac{\widehat{OD}(D|E)}{\widehat{OD}(D|\text{not } E)} \simeq \frac{0.7368}{0.3333} \simeq 2.21.$$

Alternatively, and more directly, using Formula (1.2),

$$\widehat{OR} = \frac{a \times d}{b \times c} = \frac{14 \times 39}{19 \times 13} \simeq 2.21.$$

The estimated odds ratio is greater than 1, indicating a positive association between not wearing a seat belt and sustaining at least moderately severe injury in the event of a serious accident. The corresponding relative risk was 1.70. In general, the odds ratio is further away from 1 than the relative risk. ♦

In Example 1.4 the odds were estimated to four decimal places, whereas in all previous calculations in this section (for example, in Example 1.2) only two decimal places were retained. The reason for this is that the odds in Example 1.4 were then used to obtain the odds ratio. In general, if a quantity is to be used in a later calculation, four decimal places will be retained. Usually, however, final results will be quoted to two decimal places, as for the estimated odds ratio in Example 1.4.

Activity 1.4 Odds ratio for the protective effect of seat belts

Using Table 1.5, calculate the sample odds ratio for avoidance of moderate or worse injury (D^*) for seat belt use (E^*) in children aged 4–14 involved in a serious car accident. Verify that the value is the same as that obtained in Example 1.4.

In Activity 1.3, you saw that, in most instances, relabelling the exposure categories and the disease categories changes the relative risk. Activity 1.4 shows that this is not the case for the odds ratio: switching the labels for both the exposure categories and the disease outcomes does not change the odds ratio. This suggests that of the two measures of strength of association, the odds ratio is the better.

The two measures of association are summarized in the following box.

Measures of association

The relative risk RR and the odds ratio OR are measures of association between an exposure E and a disease D. They are defined by

$$RR = \frac{P(D|E)}{P(D|\text{not } E)}, \quad OR = \frac{P(D|E) \times P(\text{not } D|\text{not } E)}{P(\text{not } D|E) \times P(D|\text{not } E)}.$$

Data from a cohort study may be presented conveniently as in the following table.

	Disease outcome		
Exposure category	D	not D	Total
E (exposed group)	a	b	$n_1 = a + b$
not E (control group)	c	d	$n_2 = c + d$

The relative risk RR and the odds ratio OR may be estimated in a cohort study by

$$\widehat{RR} = \frac{a/n_1}{c/n_2}, \quad \widehat{OR} = \frac{a \times d}{b \times c}.$$

1.3 Which measure of association should be used?

In Subsection 1.2, two measures of association in cohort studies were defined: the relative risk RR and the odds ratio OR. It was argued that the odds ratio represents the better measure of strength of association. This is one reason for preferring OR over RR.

However, measuring the strength of an association in a satisfactory manner is just one aspect of summarizing epidemiological data: communicating the results is also important. The most appropriate measure to use depends on the context. For example, a doctor who needs to convey to her patients the impact of smoking might find it more convenient to quote relative risks than odds ratios. Thus a statement such as 'smoking is associated with an x-fold increase in the risk of lung cancer' is another way of saying that the relative risk of lung cancer in smokers compared to non-smokers is $RR = x$.

For uncommon diseases, the odds ratio and the relative risk are virtually identical. This is because, in the notation of Table 1.4, a is then very much less than n_1, and c is very much less than n_2. Thus $b = n_1 - a \simeq n_1$ and $d = n_2 - c \simeq n_2$, and hence

$$\widehat{OR} = \frac{a \times d}{b \times c} \simeq \frac{a \times n_2}{n_1 \times c} = \frac{a/n_1}{c/n_2} = \widehat{RR}.$$

Activity 1.5 Measures of association for redundancy data

In Activity 1.1 data were presented on serious self-inflicted injury (SSII) in two groups of workers, one made compulsorily redundant and the other not. The data are reproduced in Table 1.6.

Table 1.6 Serious self-inflicted injury (SSII) and compulsory redundancy in meat-processing workers in New Zealand, 1986–94

Exposure category	SSII	No SSII	Total
Made compulsorily redundant (Whakatu workers)	14	1931	1945
Not made compulsorily redundant (Tomoana workers)	4	1763	1767

(a) Use these data to obtain estimates of RR and OR for the association between compulsory redundancy and SSII. What do these indicate about a possible association between compulsory redundancy and SSII?

(b) Comment on the relative sizes of the estimates of RR and OR.

(c) Express the information you obtained in part (a) in one or two sentences, using language appropriate for a non-statistical audience.

Summary of Section 1

Cohort studies for investigating the association between an exposure E and a disease D have been described. A controlled cohort study involves two groups: one group with exposure E and a control group without this exposure. Both groups are followed over time and occurrences of the disease D are recorded. Two measures of association have been introduced: the relative risk RR, and the odds ratio OR. These are used to quantify the strength of association, if any, between an exposure E and a disease D. If there is no association, then $RR = 1$ and $OR = 1$. If there is a positive association between E and D, then $RR > 1$ and $OR > 1$. Similarly, if there is a negative association, then $RR < 1$ and $OR < 1$. If the disease is uncommon, then the relative risk and the odds ratio are similar.

Exercise on Section 1

Exercise 1.1 Pre-eclampsia or eclampsia and hypertension in later life

(a) Use the data from Table 1.1 to estimate the relative risk RR and the odds ratio OR for the association between hypertension in later life and pre-eclampsia or eclampsia.

(b) What do these estimates suggest about a possible association between pre-eclampsia or eclampsia, and hypertension later in life?

2 Models for cohort studies

The relative risk RR and the odds ratio OR may be estimated in a cohort study as described in Section 1. However, these estimates are subject to sampling variability: if the cohort study were repeated with different individuals from the same population, the estimates would almost certainly be different. Indeed, if the original study were small, the difference could be large, possibly even suggesting an association in the opposite direction.

Example 2.1 Post-traumatic stress disorder in Gulf War veterans

Within months of returning from the 1991 Gulf War, some veterans began to report various symptoms and illnesses. One study from the USA collected data on post-traumatic stress disorder (PTSD) in 6617 veterans who were deployed to the Persian Gulf and 2963 veterans who were deployed to areas other than the Persian Gulf. There were 893 cases of PTSD among veterans deployed to the Gulf, and 180 cases of PTSD among those deployed to other areas. The data are shown in Table 2.1.

Kang, H.K., Natelson, B.H., Mahan, C.M., Lee, K.Y. and Murphy, F.M. (2003) Post-traumatic stress disorder and chronic fatigue syndrome-like illness among Gulf War veterans: a population-based survey of 30 000 veterans. *American Journal of Epidemiology*, **157**, 141–148.

Table 2.1 Deployment to the Gulf and post-traumatic stress disorder (PTSD) in US veterans

Exposure category	PTSD	No PTSD	Total
Deployed to the Gulf	893	5724	6617
Deployed to other areas	180	2783	2963

The estimated odds ratio for association between deployment to the Gulf and PTSD is

$$\widehat{OR} = \frac{893 \times 2783}{5724 \times 180} \simeq 2.41.$$

This is greater than 1, indicating a positive association in this particular sample: among these particular veterans deployment to the Gulf is associated with a higher rate of post-traumatic stress disorder than deployment to other areas. However, to make inferences beyond this particular sample about veterans in general, the uncertainty in the estimate of OR resulting from sampling variability must be quantified. One way to do this is to calculate a confidence interval for OR. ♦

To calculate confidence intervals, a statistical model to represent the random variation in a cohort study must be specified. This is done in Subsection 2.1. Then, in Subsections 2.2 and 2.3, this model is used to derive confidence intervals for the relative risk RR and the odds ratio OR, respectively.

2.1 The binomial model

A controlled cohort study involves two groups: a group of individuals with the exposure E, of size n_1, and a control group of individuals without exposure E, of size n_2. These groups are then followed for a specified period and observed for the occurrence of disease D. For individuals in the exposed group the probability of disease during the course of the study is $p_1 = P(D|E)$, and for individuals in the unexposed control group, the probability is $p_2 = P(D|\text{not } E)$.

Let X be a random variable denoting the number of individuals who develop disease D in an exposed group of size n_1, and let Y denote the corresponding number in a control group of size n_2. These variables are displayed in Table 2.2.

Table 2.2 Random variables for a cohort study

	Disease outcome		
Exposure category	D	not D	Total
E (exposed group)	X	$n_1 - X$	n_1
not E (control group)	Y	$n_2 - Y$	n_2

Table 2.2 is similar to Table 1.2, except that the observed frequencies a and c have been replaced by the random variables X and Y.

Provided that the disease outcomes for the individuals in the exposed group and the control group are independent, and that the probability of disease is the same for each individual within each group, then the natural probability models for X and Y are binomial:

$$X \sim B(n_1, p_1), \quad Y \sim B(n_2, p_2).$$

In addition, it will be assumed that X and Y are independent. These assumptions should be checked in each specific application. Activity 2.1 gives an example where some of the assumptions are violated.

Activity 2.1 Air bags and the risk of dying in a car accident

To evaluate the effectiveness of air bags in reducing the probability of dying in a car crash, a study was undertaken based on records of car accidents in the United States. Records were selected of serious accidents involving cars with, in addition to the driver, a single person in the front passenger seat.

Cummings, P., McKnight, B., Rivara, F.P. and Grossman, D.C. (2002) Association of driver air bags with driver fatality: a matched cohort study. *British Medical Journal*, **324**, 1119–1122.

There were 8517 cars in which the driver had an air bag and the passenger did not have an air bag. Table 2.3 shows the numbers of fatalities among drivers and passengers.

Table 2.3 Fatalities among drivers with air bags and passengers without air bags

	Died		
Exposure category	Yes	No	Total
Driver with air bag	4474	4043	8517
Passenger without air bag	5496	3021	8517

(a) Comment on the relation between seat position and having an air bag.

(b) Are the outcome variables X and Y independent? Explain your reasoning.

Comment

In the published analysis, additional data were used to separate out the effects of seat position and air bag use, and the method of analysis allowed for the pairing of drivers and passengers. The conclusion of the study was that air bags reduced the risk of death by about 8%, compared to a reduction of 65% associated with seat belts. Using both reduced the risk of death by 68%.

2.2 Confidence intervals for the relative risk

In this subsection approximate confidence intervals for the relative risk are obtained. Since, in practice, it is simpler to work with the logarithm of the relative risk than with the relative risk itself, the method used involves first finding a z-interval for the logarithm of the relative risk, and then calculating the corresponding confidence interval for the relative risk.

The notation introduced in Table 1.2 for the entries in a data table for a cohort study will be used throughout this subsection. It is reproduced in Table 2.4 for ease of reference.

Table 2.4 A general data table for a cohort study

	D	not D	Total
E	a	b	n_1
not E	c	d	n_2

In general, for a sufficiently large sample, an approximate $100(1-\alpha)\%$ confidence interval (or z-interval) for a parameter θ, which is denoted (θ^-, θ^+), is given by

$$(\theta^-, \theta^+) = (\widehat{\theta} - z\widehat{\sigma}, \widehat{\theta} + z\widehat{\sigma}),$$

where $\widehat{\theta}$ is the sample estimate of θ, $\widehat{\sigma}$ is the estimated standard error of the estimator $\widehat{\theta}$ and z is the $(1-\alpha/2)$-quantile of the standard normal distribution.

Let $\theta = \log(RR)$. The estimate of the relative risk is given by Formula (1.1):

$$\widehat{RR} = \frac{a/n_1}{c/n_2}.$$

So the estimate $\widehat{\theta}$ of θ is $\log(\widehat{RR})$.

If σ denotes the standard error of the estimator $\widehat{\theta} = \log(\widehat{RR})$ then, for n_1 and n_2 sufficiently large, it can be shown that σ can be estimated by

$$\widehat{\sigma} = \sqrt{\frac{1}{a} - \frac{1}{n_1} + \frac{1}{c} - \frac{1}{n_2}}. \tag{2.1}$$

The proof uses a mathematical technique known as Taylor series expansion and, while it is not difficult, it will be omitted.

Thus an approximate $100(1-\alpha)\%$ confidence interval for $\theta = \log(RR)$ is given by

$$\left(\theta^-, \theta^+\right) = \left(\log(\widehat{RR}) - z \times \widehat{\sigma}, \log(\widehat{RR}) + z \times \widehat{\sigma}\right),$$

where $\widehat{\sigma}$ is given by Formula (2.1) and z is the $(1-\alpha/2)$-quantile of the standard normal distribution.

Since $\theta = \log(RR)$, it follows that $RR = \exp(\theta)$, and hence an approximate $100(1-\alpha)\%$ confidence interval (RR^-, RR^+) for the relative risk RR is $(\exp(\theta^-),\ \exp(\theta^+))$. Thus

$$RR^- = \exp\Big(\log(\widehat{RR}) - z \times \widehat{\sigma}\Big) = \widehat{RR} \times \exp(-z \times \widehat{\sigma}),$$

$$RR^+ = \exp\Big(\log(\widehat{RR}) + z \times \widehat{\sigma}\Big) = \widehat{RR} \times \exp(z \times \widehat{\sigma}).$$

So an approximate $100(1-\alpha)\%$ confidence interval for the relative risk RR is given by

$$(RR^-, RR^+) = \left(\widehat{RR} \times \exp(-z \times \widehat{\sigma}), \widehat{RR} \times \exp(z \times \widehat{\sigma})\right),$$

where $\widehat{\sigma}$ is given by (2.1) and z is the $(1-\alpha/2)$-quantile of $N(0,1)$.

The details are summarized in the following box.

Confidence intervals for the relative risk

The sample estimate $\widehat{RR}$ of the relative risk RR derived from a cohort study is, using the notation of Table 2.4, given by

$$\widehat{RR} = \frac{a/n_1}{c/n_2}. \tag{2.2}$$

For sufficiently large n_1 and n_2, an approximate $100(1-\alpha)\%$ confidence interval for the relative risk RR is

$$(RR^-, RR^+) = \left(\widehat{RR} \times \exp(-z \times \widehat{\sigma}), \widehat{RR} \times \exp(z \times \widehat{\sigma})\right), \tag{2.3}$$

where z is the $(1-\alpha/2)$-quantile of the standard normal distribution and

$$\widehat{\sigma} = \sqrt{\frac{1}{a} - \frac{1}{n_1} + \frac{1}{c} - \frac{1}{n_2}}. \tag{2.4}$$

Example 2.2 Breast cancer and hormone-replacement therapy

Concerns over a possible link between breast cancer and hormone-replacement therapy (HRT) led to a very large cohort study being undertaken in the UK. Between 1996 and 2001, the study, known as the Million Women Study, recruited 1 084 110 women aged between 50 and 64 who were followed up for cancer. Table 2.5 contains data from this study on two groups of women: women who were using combined oestrogen-progestagen HRT at the time of recruitment (the exposed group), and women who had never used HRT at the time of recruitment (the control group). In each group, the number of new cases of invasive breast cancer occurring during the study follow-up period was counted.

Million Women Study Collaborators (2003) Breast cancer and hormone-replacement therapy in the Million Women Study. *Lancet*, **362**, 419–427.

Table 2.5 Invasive breast cancer and use of oestrogen-progestagen HRT

	Invasive breast cancer		
Exposure category	Yes	No	Total
Currently using combined oestrogen-progestagen HRT	1934	140 936	142 870
Never used HRT	2894	389 863	392 757

The estimated relative risk of invasive breast cancer for use of oestrogen-progestagen HRT is

$$\widehat{RR} = \frac{a/n_1}{c/n_2} = \frac{1934/142\,870}{2894/392\,757} \simeq 1.8371.$$

For a 99% confidence interval, the 0.995-quantile of the standard normal distribution is required. From the table of quantiles of the standard normal distribution in the *Handbook*, this is $z = 2.576$.

The estimated standard error $\widehat{\sigma}$ is given by (2.4):

$$\begin{aligned}\widehat{\sigma} &= \sqrt{\frac{1}{a} - \frac{1}{n_1} + \frac{1}{c} - \frac{1}{n_2}} \\ &= \sqrt{\frac{1}{1934} - \frac{1}{142\,870} + \frac{1}{2894} - \frac{1}{392\,757}} \\ &\simeq 0.02921.\end{aligned}$$

So, using (2.3), the 99% confidence limits for the relative risk are

$$\begin{aligned}RR^- &= \widehat{RR} \times \exp(-z \times \widehat{\sigma}) \\ &\simeq 1.8371 \times \exp(-2.576 \times 0.02921) \\ &\simeq 1.70,\end{aligned}$$

$$\begin{aligned}RR^+ &= \widehat{RR} \times \exp(z \times \widehat{\sigma}) \\ &\simeq 1.8371 \times \exp(2.576 \times 0.02921) \\ &\simeq 1.98.\end{aligned}$$

Thus the risk of invasive breast cancer is 1.84 times higher in women taking oestrogen-progestagen HRT than in women who never took any HRT, with 99% confidence interval $(1.70, 1.98)$. The confidence interval is quite narrow, reflecting the large size of the study, and is located entirely above 1. This indicates a positive association between HRT use and breast cancer. ♦

Activity 2.2 *Efficacy of measles vaccines*

Prior to the introduction of routine childhood vaccination against measles in 1968, there were hundreds of thousands of cases of measles every year in the UK. In 1964, a cohort study was undertaken to evaluate the efficacy of measles vaccines. Children aged between 10 months and 2 years were enrolled into one of three groups: an unvaccinated group who received no vaccine, a vaccinated group who received live measles vaccine, and a second vaccinated group who received killed measles vaccine followed by live measles vaccine. Allocation to the groups was based on day of birth. Table 2.6 shows the numbers of children in the unvaccinated and live vaccine groups (the live vaccine was the one chosen subsequently for routine immunization), and the numbers of measles cases arising within six months after vaccination.

Medical Research Council (1966) Vaccination against measles: a clinical trial of live measles vaccine given alone and live vaccine preceded by killed vaccine. *British Medical Journal*, 19 February, 441–446.

Table 2.6 Measles vaccination and measles infection

	Measles within six months		
Exposure category	Yes	No	Total
Received live measles vaccine	156	9 421	9 577
Unvaccinated	1531	14 797	16 328

(a) Estimate the relative risk of measles after vaccination.

(b) Obtain a 99% confidence interval for the relative risk RR.

(c) Interpret your results. Does the live vaccine protect against measles?

2.3 Confidence intervals for the odds ratio

In this subsection, approximate confidence intervals for the odds ratio are obtained. The method for calculating confidence intervals for the odds ratio is very similar to that for the relative risk. The derivation of the formula for a confidence interval also involves using logarithms, and the formula for an approximate confidence interval for the odds ratio is similar in form to that for the relative risk. The main difference is in the formula for the estimated standard error $\widehat{\sigma}$ that is required to calculate a confidence interval. And, of course, the formula for the estimate $\widehat{OR}$ is different from that for $\widehat{RR}$. The details are given in the following box.

Confidence intervals for the odds ratio

The sample estimate $\widehat{OR}$ of the odds ratio derived from a cohort study is, using the notation of Table 2.4, given by

$$\widehat{OR} = \frac{a \times d}{b \times c}. \tag{2.5}$$

For sufficiently large $n_1 = a + b$ and $n_2 = c + d$, an approximate $100(1 - \alpha)\%$ confidence interval for the odds ratio OR is

$$(OR^-, OR^+) = \left(\widehat{OR} \times \exp(-z \times \widehat{\sigma}), \widehat{OR} \times \exp(z \times \widehat{\sigma})\right), \tag{2.6}$$

where z is the $(1 - \alpha/2)$-quantile of the standard normal distribution and

$$\widehat{\sigma} = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}. \tag{2.7}$$

Example 2.3 Cannabis use and mental health

Recreational use of cannabis is now widespread among young people in many countries. Uncertainty persists about its consequences for health. Some research has suggested that heavy use of cannabis is associated with depression and anxiety. A cohort study was undertaken in Australia to investigate this hypothesis. Teenagers were recruited from 44 schools in Victoria, Australia, and followed to age 20 or 21. Cannabis use was monitored using self-administered questionnaires.

Patton, G.C., Coffey, C., Carlin, J.B., Degenhardt, L., Lynskey, M. and Hall, W. (2002) Cannabis use and mental health in young people: cohort study. *British Medical Journal*, **325**, 1195–1198.

Table 2.7 shows data for young women derived from this study. Frequency of cannabis use was grouped into two categories: less than weekly (the control group), and weekly or more (the exposed group). The outcome of interest (D) is depression or anxiety, assessed at interviews with specialists.

Table 2.7 Depression and anxiety in young women according to cannabis use

Frequency of cannabis use	Depression or anxiety		
	Yes	No	Total
Weekly or more	35	34	69
Less than weekly	153	637	790

The estimated odds ratio of depression or anxiety for frequent use of cannabis compared to infrequent use among young women is given by (2.5):

$$\widehat{OR} = \frac{a \times d}{b \times c} = \frac{35 \times 637}{34 \times 153} \simeq 4.2859.$$

For a 95% confidence interval, the 0.975-quantile of the standard normal distribution is required; this is $z = 1.96$.

The estimated standard error $\widehat{\sigma}$ is given by (2.7):

$$\begin{aligned}\widehat{\sigma} &= \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}} \\ &= \sqrt{\frac{1}{35}+\frac{1}{34}+\frac{1}{153}+\frac{1}{637}} \\ &\simeq 0.2571.\end{aligned}$$

So, using (2.6), the 95% confidence limits for the odds ratio are

$$\begin{aligned}OR^{-} &= \widehat{OR} \times \exp\left(-z \times \widehat{\sigma}\right) \\ &\simeq 4.2859 \times \exp(-1.96 \times 0.2571) \\ &\simeq 2.59,\end{aligned}$$

$$\begin{aligned}OR^{+} &= \widehat{OR} \times \exp\left(z \times \widehat{\sigma}\right) \\ &\simeq 4.2859 \times \exp(1.96 \times 0.2571) \\ &\simeq 7.09.\end{aligned}$$

Thus the odds ratio is 4.29, with 95% confidence interval $(2.59, 7.09)$. The confidence interval for OR is located well above 1, suggesting that heavy cannabis use is associated with a big increase in the odds of depression or anxiety in young women. ♦

Activity 2.3 Cannabis use and depression or anxiety in young men

The data in Example 2.3 relate only to young women. The data on young men from the same study are shown in Table 2.8.

Table 2.8 Depression and anxiety in young men according to cannabis use

Frequency of cannabis use	Depression or anxiety		
	Yes	No	Total
Weekly or more	20	126	146
Less than weekly	51	534	585

(a) Estimate the odds ratio of anxiety or depression for frequent use of cannabis relative to infrequent use among young men.

(b) Obtain a 95% confidence interval for the odds ratio.

(c) Interpret your findings.

Summary of Section 2

The binomial model for cohort studies has been introduced. The assumptions of this model are that outcomes for individuals within each group are independent and occur with the same probability. To calculate confidence intervals, it is also assumed that outcomes are independent between groups. Approximate confidence intervals for RR and OR have been presented, based on the logarithm of the relative risk and odds ratio.

Exercise on Section 2

Exercise 2.1 Post-traumatic stress disorder

(a) In Example 2.1 data were presented on post-traumatic stress disorder (PTSD) in US veterans and the odds ratio OR was estimated. Use the data in Table 2.1 to obtain a 99% confidence interval for OR.

(b) Is it plausible that deployment to the Gulf is not associated with increased rates of PTSD in US veterans? Explain carefully why you reach your conclusion.

3 Case-control studies

One drawback of cohort studies is that, when the disease of interest is uncommon, a cohort study may need to be very large or involve very lengthy follow-up in order to obtain sufficient numbers of individuals with the disease. For example, if a health event occurs on average in one per thousand individuals over a given period, then to obtain ten cases would require about 10 000 individuals to be followed over that period. Such large studies are usually time-consuming and costly to undertake.

In this section, a different study design, the **case-control** study, is considered. This study design makes it possible to study uncommon health outcomes without the need for very large samples or very lengthy studies. In Subsection 3.1, case-control studies are described, and in Subsection 3.2, measures of association for case-control studies are discussed. In Subsection 3.3, studies with several exposure categories are considered.

3.1 What is a case-control study?

You have seen that, in a typical controlled cohort study, two groups of individuals — an exposed group with exposure E and a control group without exposure E — are followed over time and occurrences of disease D are counted within each group. The data from such a study may be summarized as in Table 3.1.

Table 3.1 A general data table for a cohort study

	Disease outcome		
Exposure category	D	not D	Total
E (exposed group)	a	b	$n_1 = a + b$
not E (control group)	c	d	$n_2 = c + d$

A key aspect of the cohort study design is the difference between the status of the exposure E and the disease D. For each individual in the study, the exposure category is regarded as *fixed* and the disease outcome is regarded as *random*.

In a case-control study, a sample of cases — that is, individuals who have the disease D (for example, women with breast cancer) — is selected. Then a second group of individuals who are not cases (for example, women who do not have breast cancer) is selected. These individuals are the controls. Thus in these two groups, cases and controls, the disease outcome is known. However, the exposure category of the cases and the controls is treated as random. After the cases and

controls have been selected, their previous exposures (for example, whether they ever used hormone-replacement therapy) are ascertained. This results in a data table as set out in Table 3.2.

Table 3.2 A general data table for a case-control study

	Disease outcome	
Exposure category	D (cases)	not D (controls)
E	a	b
not E	c	d
Total	$m_1 = a + c$	$m_2 = b + d$

Note the differences between Table 3.2 and Table 3.1. In a case-control study (Table 3.2), the numbers with disease D and without disease D are fixed in advance: the study includes m_1 cases and m_2 controls. In contrast, in a cohort study, the exposure group and the control group, of sizes n_1 and n_2, are fixed in advance. In a case-control study (Table 3.2), the presence or absence of exposure E is determined by looking back in time at the histories of the cases and controls. In contrast, in a cohort study the disease outcome is determined by following the exposed group and the unexposed group forward in time. The key features of a case-control study are set out in the following box.

Case-control studies

A case-control study of the association between an exposure E and a disease D has the following features.

- ◇ It includes a group of cases with the disease D and a group of controls without the disease D who are otherwise comparable to the cases.
- ◇ The past exposures of cases and controls are determined and occurrences of exposure E are identified.

An important issue in case-control studies is how to select the controls. As a general rule, they should be selected from the population that gave rise to the cases, and should have had the same opportunity as the cases to become exposed.

Once exposures in cases and controls have been determined, the proportions of cases and controls with exposure E are compared. If there is no association between E and D, then the proportions exposed should be similar in cases and controls. On the other hand, if D and E are positively associated, a greater proportion of cases than controls would be expected to have exposure E.

Example 3.1 Smoking and lung cancer

The 1950 case-control study of smoking and lung cancer by Richard Doll and Austin Bradford Hill is a classic example. An unexplained increase in lung cancer deaths had taken place over the previous decades in the UK and several other countries. The increase was spectacular: lung cancer deaths in the UK had increased fifteen-fold between 1922 and 1947. Several exposures had been suggested as possible causes for the increase, including industrial pollution, tarred roads, exhaust fumes from cars, as well as smoking.

Doll, R. and Hill, A.B. (1950) Smoking and carcinoma of the lung: preliminary report. *British Medical Journal*, 30 September, 739–748.

Doll and Hill investigated the association by means of a case-control study. The cases were patients admitted to hospital with lung cancer. For each case, a control of the same sex and similar age admitted to the hospital for a disease other than cancer was selected.

Smoking histories were then obtained for each of the 709 cases and 709 controls. Table 3.3 shows the data for the males (649 cases and 649 controls), with exposure defined as having been a smoker at any time in the past.

Table 3.3 Smoking and lung cancer in males

Exposure category	Cases of lung cancer	Controls
Smoked	647	622
Never smoked	2	27
Total	649	649

An important feature of the case-control design is that it involves only 649 cases and 649 controls. A cohort study would have had to be very large to result in 649 lung cancer cases.

The proportion of smokers among cases is $647/649 \simeq 0.9969$ compared to $622/649 \simeq 0.9584$ among controls. Thus it appears that smoking was extremely widespread, but more common among lung cancer cases than among controls. In Subsection 3.2, a suitable measure of association will be discussed. ♦

In Example 3.1 the numbers of controls and cases were the same. This is not a requirement: the numbers of cases and controls are often different, as in the study described in Activity 3.1.

Activity 3.1 Political activity and homicide in Karachi

In 2001 it was estimated that every year half a million people are murdered in the world. In Karachi, Pakistan, homicide rates vary substantially between neighbourhoods. A case-control study was undertaken in one area to identify whether political activity was associated with death by homicide.

Mian, A., Mahmood, S.F., Chotani, H. and Luby, S. (2002) Vulnerability to homicide in Karachi: political activity as a risk factor. *International Journal of Epidemiology*, **31**, 581–585.

Altogether 35 victims of homicide were included in the study, and 85 controls with similar age and sex distribution as the victims. Household members were questioned about the political activities of the study subjects. Of the 35 victims, eleven had attended political meetings, compared to two of the controls.

(a) Arrange these data in a table similar to Table 3.3, indicating clearly what the exposure is.

(b) Informally compare the proportions of exposed cases (victims of homicide) and exposed controls. What does this suggest?

3.2 Measures of association in case-control studies

In a cohort study, the numbers exposed and not exposed are regarded as fixed, and the measures of association, namely the relative risk RR and the odds ratio OR, are based on $P(D|E)$, the probability of disease given exposure, and $P(D|\text{not } E)$, the probability of disease given no exposure.

In a case-control study, however, the numbers of cases (with disease D) and controls (without disease D) included are decided in advance by the investigator. In consequence, $P(D|E)$ and $P(D|\text{not } E)$ cannot be estimated in a case-control study. So the relative risk RR, which is the ratio of these probabilities, cannot be estimated in a case-control study. However, the odds ratio can be estimated. This is illustrated in Example 3.2.

Example 3.2 Measures of association in case-control studies

In Example 3.1, data from the famous Doll and Hill case-control study of smoking and lung cancer were presented. These data are reproduced in Table 3.4. Also included in the table are the row totals that would be presented if this were a cohort study.

Table 3.4 Smoking and lung cancer in males

Exposure category	Cases of lung cancer	Controls	Total
Smoked	647	622	1269
Never smoked	2	27	29
Total	649	649	1298

Suppose that the relative risk and odds ratio were to be estimated as if this were a cohort study. Then the estimates would be as follows:

$$\widehat{RR} = \frac{647/1269}{2/29} \simeq 7.39, \quad \widehat{OR} = \frac{647 \times 27}{622 \times 2} \simeq 14.04.$$

Now suppose that Doll and Hill had, in fact, decided to obtain ten times as many controls as cases. This is perfectly admissible: the numbers of cases and controls chosen is entirely within the control of the investigators. Then, assuming that the proportion of smokers among these controls was the same as among those actually selected, the data would have been as below.

Exposure category	Cases of lung cancer	Controls $\times 10$	Total
Smoked	647	6220	6867
Never smoked	2	270	272
Total	649	6490	7139

This would produce the following estimates of the relative risk and odds ratio:

$$\widehat{RR} = \frac{647/6867}{2/272} \simeq 12.81, \quad \widehat{OR} = \frac{647 \times 270}{6220 \times 2} \simeq 14.04.$$

The relative risk has increased, but the odds ratio is unchanged. Similarly, if Doll and Hill had used ten times as many cases as they did, the data would have been as follows.

Exposure category	Cases of lung cancer $\times 10$	Controls	Total
Smoked	6470	622	7092
Never smoked	20	27	47
Total	6490	649	7139

The estimated relative risk and odds ratio in this case would be as follows:

$$\widehat{RR} = \frac{6470/7092}{20/47} \simeq 2.14, \quad \widehat{OR} = \frac{6470 \times 27}{622 \times 20} \simeq 14.04.$$

The relative risk is now much lower, but again the odds ratio is unchanged at 14.04. In fact, the odds ratio will be the same whatever the ratio of cases to controls, whereas the relative risk will vary. ♦

Example 3.2 shows that the relative risk RR *cannot* be estimated in a case-control study, since its value varies according to how many cases and controls are selected. However, the odds ratio OR does not depend on how many cases and controls are selected, only on the proportions of cases and controls with the exposure. In particular, its estimate from a case-control study would have the same value as if a cohort study had been undertaken. Thus the odds ratio *can* be estimated in a case-control study. This may be demonstrated more generally as follows.

In a case-control study it is possible to estimate $OD(E|D)$, the odds of exposure in cases, and $OD(E|\text{not } D)$, the odds of exposure in controls. The estimates, using the notation of Table 3.2, are as follows:

$$\widehat{OD}(E\,|D\,) = \frac{a/m_1}{c/m_1} = \frac{a}{c},$$

$$\widehat{OD}(E\,|\text{not } D\,) = \frac{b/m_2}{d/m_2} = \frac{b}{d}.$$

Clearly, it is also possible to estimate the ratio of these odds:

$$\frac{\widehat{OD}(E\,|D)}{\widehat{OD}(E\,|\text{not } D)} = \frac{a/c}{b/d} = \frac{a \times d}{b \times c}.$$

Notice that this expression is the same as that obtained for the estimate of the odds ratio OR in a cohort study (see Result (1.2)). It follows that

$$\widehat{OR} = \frac{\widehat{OD}(D|E)}{\widehat{OD}(D|\text{not } E)} = \frac{\widehat{OD}(E|D)}{\widehat{OD}(E|\text{not } D)}.$$

This identity also applies to the underlying parameters, that is, without the 'hat' symbols.

Hence the odds ratio OR can be estimated in a case-control study, even though the relative risk RR cannot. Confidence intervals for OR are also calculated in the same way in a case-control study as in a cohort study. These facts are summarized in the following box.

Measures of association in case-control studies

In a case-control study, the odds ratio OR can be estimated but the relative risk RR cannot. Using the notation of Table 3.2, the odds ratio is estimated by

$$\widehat{OR} = \frac{a \times d}{b \times c}.$$

Approximate $100(1-\alpha)\%$ confidence intervals for OR are calculated in the same way in a case-control study as in a cohort study (see Formulas (2.6) and (2.7)).

In Subsection 1.3 you saw that, for uncommon diseases, the odds ratio and the relative risk are in fact very close. Thus, in case-control studies, although the relative risk cannot be estimated directly, when the disease is uncommon, it can be approximated by the odds ratio.

Example 3.3 Alcohol consumption and fatal car accidents

The twentieth century saw the emergence of a new and deadly epidemic: injury from car accidents. Alcohol consumption was soon identified as a likely cause of accidents. This example is based on the first controlled study of the role of alcohol consumption in causing fatal car accidents. The study design chosen was the case-control design.

McCarroll, J.R. and Haddon, W. (1962) A controlled study of fatal automobile accidents in New York City. *Journal of Chronic Diseases*, **15**, 811–826.

Details were obtained of all fatalities from car accidents in New York between June and October in 1959 and in 1960. The fatalities were classified according to whether or not the dead person was considered to be responsible for the accident. This example includes 24 drivers who were killed in car accidents for which they were considered to be responsible. This group of 24 constitutes the cases. Their blood alcohol levels were obtained from post-mortem examinations.

Controls were obtained by selecting drivers passing the locations where the accidents of the cases occurred, at the same time of day and on the same day of the week. A total of 154 controls were selected in this way. The controls were breathalyzed. Exposure is defined as a high blood alcohol level, namely a concentration greater than or equal to 100 mg% (1 mg% is 1 mg of alcohol per 100 ml of blood). The data are in Table 3.5 (overleaf).

The study authors, who were accompanied by police, report encountering 'occasional hostility, and in one case an initial plea of diplomatic immunity'.

The estimated odds ratio for the association between alcohol level and dying in a car accident for which one is responsible is given by

$$\widehat{OR} = \frac{a \times d}{b \times c} = \frac{14 \times 146}{8 \times 10} = 25.55.$$

Thus the odds of causing a car accident and dying in it are 25.55 times higher for drivers with high blood alcohol levels than for other drivers.

Table 3.5 Alcohol levels and fatal car accidents

Alcohol level	Cases	Controls
≥ 100 mg%	14	8
< 100 mg%	10	146
Total	24	154

For a 95% confidence interval, the 0.975-quantile of the standard normal distribution is required: $z = 1.96$. The estimated standard error $\widehat{\sigma}$ is given by (2.7):

$$\begin{aligned}\widehat{\sigma} &= \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}} \\ &= \sqrt{\frac{1}{14} + \frac{1}{8} + \frac{1}{10} + \frac{1}{146}} \\ &\simeq 0.5507.\end{aligned}$$

So the confidence limits are

$$\begin{aligned}OR^{-} &= \widehat{OR} \times \exp(-z \times \widehat{\sigma}) \\ &\simeq 25.55 \times \exp(-1.96 \times 0.5507) \\ &\simeq 8.68,\end{aligned}$$

$$\begin{aligned}OR^{+} &= \widehat{OR} \times \exp(z \times \widehat{\sigma}) \\ &\simeq 25.55 \times \exp(1.96 \times 0.5507) \\ &\simeq 75.19.\end{aligned}$$

The confidence interval for the odds ratio is $(8.68, 75.19)$, indicating a positive (and rather strong) association between alcohol consumption and fatal car accidents. ♦

Activity 3.2 Cot deaths and sleeping position

Sudden unexplained deaths in apparently normal babies under one year of age are known as sudden infant deaths, or cot deaths. In the UK, they are the leading cause of death in babies aged between one month and one year. The causes of Sudden Infant Death Syndrome (SIDS) are not known. In 1990 a case-control study was published that suggested that babies who were put down to sleep on their front and who were too heavily wrapped were more likely to die of SIDS. Following this study, the 'Back to Sleep' campaign was launched in several countries to encourage parents to place their babies to sleep on their back and to avoid overheating and smoky environments. In subsequent years, deaths from SIDS dropped by over 50%. Table 3.6 shows data from the 1990 study.

Fleming, P.J., Gilbert, R., Azaz, Y. *et al.* (1990) Interaction between bedding and sleeping position in the sudden infant death syndrome: a population based case-control study. *British Medical Journal*, **301**, 85–89.

Table 3.6 Sleeping position and deaths from SIDS

Position baby last placed down to sleep	Cases	Controls
On its front	62	76
In another position	5	55
Total	67	131

A total of 67 babies who died of SIDS were included. The controls are live babies of similar ages and from the same localities as the babies that died. The exposure is last placing the baby down to sleep on its front.

(a) Estimate the odds ratio of SIDS associated with the front sleeping position.

(b) Calculate a 95% confidence interval for the odds ratio.

(c) Interpret your results.

3.3 Studies with more than two exposure categories

So far, all the studies we have considered (whether cohort or case-control) have involved just two exposure categories: exposed and unexposed. In some cases, however, there may be more than two exposure categories. In this subsection, you will learn how to analyse data from such studies. The methods described apply to cohort studies as well as case-control studies.

For studies involving more than two exposure categories, it is common practice to identify one category as a **reference** exposure category, and calculate odds ratios for the other exposure categories relative to this reference category. The reference category is usually chosen to represent lack of exposure, or 'normal' exposure in some sense, though the choice is to some extent arbitrary. The choice of reference category depends on the context of each study. The procedure is illustrated in Example 3.4.

Example 3.4 Sleeping position and SIDS

After the success of the 'Back to Sleep' campaign (see Activity 3.2), further research was undertaken to examine the relationship between sleeping position and sudden infant death syndrome (SIDS). Between 1993 and 1995, a case-control study, similar to that described in Activity 3.2, was undertaken with the aim of investigating further the causes of SIDS.

Fleming, P.J., Blair, P.S., Bacon, C. *et al.* (1996) Environment of infants during sleep and risk of the sudden infant death syndrome: results of 1993–5 case-control study for confidential inquiry into stillbirths and deaths in infancy. *British Medical Journal*, **313**, 191–195.

Data for 188 cases and 774 controls were obtained on the position the baby was placed down to sleep: back, side or front. Thus there are three exposure categories, rather than two. The data are in Table 3.7.

Table 3.7 Sleeping position and deaths from SIDS

Position baby last placed down to sleep	Cases	Controls
On its front	30	24
On its side	76	241
On its back	82	509
Total	188	774

To investigate the association between sleeping position and SIDS in a table such as this, one of the exposure categories is chosen as the reference category. In this example, the back position will be chosen (rather arbitrarily) as reference.

Odds ratios for the other two sleeping positions are then calculated, relative to the reference category. For example, for the front position, the odds ratio is calculated using rows 1 and 3 of Table 3.7. Thus

$$\widehat{OR}_{\text{front}} = \frac{30 \times 509}{24 \times 82} \simeq 7.7591.$$

A 95% confidence interval for this odds ratio is calculated in the same way as for a 2×2 table, ignoring the data in the row corresponding to the side sleeping position. Thus the estimated standard error $\widehat{\sigma}$ is given by

$$\widehat{\sigma} = \sqrt{\frac{1}{30} + \frac{1}{24} + \frac{1}{82} + \frac{1}{509}} \simeq 0.2986.$$

It follows that the 95% confidence interval for OR_{front} is $(4.32, 13.93)$. This confirms the findings described in Activity 3.2: the front sleeping position is associated with higher rates of SIDS.

This confidence interval was calculated using (2.6).

The calculation is repeated for the side sleeping position, again relative to the reference exposure category. In this case, the data in the second and third rows of Table 3.7 are used. Thus

$$\widehat{OR}_{\text{side}} = \frac{76 \times 509}{241 \times 82} \simeq 1.9575, \quad \widehat{\sigma} = \sqrt{\frac{1}{76} + \frac{1}{241} + \frac{1}{82} + \frac{1}{509}} \simeq 0.1774.$$

The 95% confidence interval for OR_{side} is $(1.38, 2.77)$. Thus the side sleeping position is also associated with an increased risk of SIDS, though the association is not as strong as for the front position. ♦

Activity 3.3 *Ectopic pregnancy and genital infection*

A pregnancy is called ectopic if the baby develops outside the womb. An ectopic pregnancy is a life-threatening condition requiring emergency treatment. A large case-control study was undertaken in the Auvergne, France, to study the factors associated with ectopic pregnancy.

Bouyer, J., Coste, J., Shojaei, T. *et al.* (2003) Risk factors for ectopic pregnancy: a comprehensive analysis based on a large case-control, population-based study in France. *American Journal of Epidemiology*, **157**, 185–194.

The data in Table 3.8 include 780 women who experienced an ectopic pregnancy, and 1673 control women who gave birth normally. The exposure considered here is history of genital infections. These exposures are classified in three categories: PID (standing for Pelvic Inflammatory Disease, a particular type of genital infection), Non-PID (genital infections other than PID), and None (no history of genital infections).

Table 3.8 Ectopic pregnancy and history of genital infections

History of genital infections	Cases	Controls
PID	212	112
Non-PID	157	407
None	411	1154
Total	780	1673

(a) Identify a suitable reference exposure category.

(b) Estimate the odds ratios and 95% confidence intervals for the other two exposure categories, relative to this reference category.

(c) Interpret your findings.

The procedure for larger tables described in this subsection applies also to cohort studies. For cohort studies, relative risks can be estimated relative to a reference category as well as odds ratios.

Summary of Section 3

Case-control studies have been described and contrasted with cohort studies. In a case-control study, a group of cases with the disease D is compared with a group of controls without the disease. The previous exposures E for cases and controls are then documented. For uncommon diseases, cohort studies may need to be very large and in such circumstances a case-control study may be more practical. You have seen that the relative risk RR cannot be estimated in a case-control study, but the odds ratio OR can. The odds ratio is estimated and confidence intervals are calculated in the same way as for cohort studies. If the disease is uncommon, the relative risk can be approximated by the odds ratio. Studies involving more than two exposure categories are analysed by identifying a reference exposure category and calculating odds ratios relative to this reference category.

Exercise on Section 3

Exercise 3.1 Hib meningitis

Haemophilus Influenzae type b (Hib) causes meningitis. Hib meningitis is a rare but serious disease, occurring most frequently in young children. In the UK, most but not all children are vaccinated against Hib in the first few months of life. In 2003, concerns were expressed about the efficacy of the vaccine used in the UK.

(a) Describe briefly the design of a cohort study to investigate the association between receipt of Hib vaccine (the exposure E) and subsequent Hib meningitis (the disease D).

(b) Describe briefly the design of a case-control study to investigate this association.

(c) State one advantage of the case-control method compared to the cohort method.

(d) An estimate of the relative risk RR is required. How might this be approximated in a case-control study?

4 Testing for no association in cohort studies and case-control studies

In Sections 1 to 3, you have seen that an estimate of a suitable measure of association and a confidence interval for the measure can be used to summarize the strength of association between an exposure E and a disease D. In this section, the question addressed is: 'Is the exposure E associated with the disease D?' This is done by testing the null hypothesis of no association between E and D. Two significance tests are described — the chi-squared test for no association and Fisher's exact test. In Subsection 4.1, the test statistic for the chi-squared test is developed; and in Subsection 4.2, the test is described. Fisher's exact test is discussed briefly in Subsection 4.3.

4.1 The chi-squared test statistic

In medical statistics, it is common to report the results of a significance test for no association as well as quoting the estimated odds ratio or the relative risk and a confidence interval.

In a significance test, the evidence against a specified null hypothesis is quantified using a p value. This is the probability that data at least as extreme as those observed would have arisen if the null hypothesis were true. In the context of the cohort studies and case-control studies considered so far, the null hypothesis is that there is no association — that is, the odds ratio OR is equal to 1. Thus a significance test quantifies the evidence against the null hypothesis of no association.

If there are more than two exposure categories then, as you saw in Subsection 3.3, the strength of association cannot be summarized by a single odds ratio. One approach is to undertake an overall significance test for no association, before investigating the data in more detail.

Example 4.1 Childhood asthma and gestational age

The proportion of people with asthma is increasing in many parts of the world, though there are large geographical variations. The reasons for the increase, and for the variations in asthma rates, are not known. It has been suggested that environmental factors operating at any time after conception could be the cause. For example, it has been suggested that factors related to the development of the baby before birth may be involved. This hypothesis was investigated in a large cohort study in Denmark.

Yuan, W., Basso, O., Sorensen, H.T. and Olsen, J. (2002) Fetal growth and hospitalization with asthma during early childhood: a follow-up study in Denmark. *International Journal of Epidemiology*, **31**, 1240–1245.

Pregnant women in Odense and Aalborg were recruited and their babies were followed up for asthma up to age 12 years. The disease D was hospitalization for asthma. The investigators analysed several variables related to foetal growth. In this example, gestational age (that is, the duration of pregnancy) is considered. Births are classified as Pre-term (the baby was premature), Term (the baby was born close to its due date) or Post-term (the baby was overdue). The data are in Table 4.1.

Table 4.1 Gestational age and childhood asthma

Gestational age	Hospitalized for asthma	Not hospitalized for asthma	Total
Pre-term	18	266	284
Term	402	8565	8967
Post-term	45	1100	1145

One approach to analysing these data is to select one group as a reference category — for example, the Term group. Then relative risks or odds ratios relative to the reference category can be estimated, and 95% confidence intervals calculated. This is the approach that was described in Subsection 3.3. For the data in Table 4.1, estimates of the relative risks are

$$\widehat{RR}_{\text{pre}} = \frac{18/284}{402/8967} \simeq 1.41, \quad \widehat{RR}_{\text{post}} = \frac{45/1145}{402/8967} \simeq 0.88.$$

The 95% confidence intervals are $(0.89, 2.23)$ for RR_{pre} and $(0.65, 1.19)$ for RR_{post}. However, if the Post-term category had been chosen as reference, then the estimated relative risk $\widehat{RR}_{\text{pre}}$ would be given by

$$\widehat{RR}_{\text{pre}} = \frac{18/284}{45/1145} \simeq 1.61,$$

and the 95% confidence interval for RR_{pre} would be $(0.95, 2.74)$. Thus a different choice of reference category produces different results, as would be expected. One advantage of an overall test for no association is that it does not require a reference category to be chosen. ♦

In testing the null hypothesis of no association, the observed frequencies are compared with the frequencies that would be expected if there were no association. So the first step is to calculate these expected frequencies. The calculation is done conditional on the row totals and column totals for the observed frequencies. Table 4.2 shows the asthma data from Table 4.1 with the column totals added.

Table 4.2 Gestational age and childhood asthma with marginal totals

Gestational age	Hospitalized for asthma	Not hospitalized for for asthma	Total
Pre-term	18	266	284
Term	402	8 565	8 967
Post-term	45	1 100	1 145
Total	465	9 931	10 396

The row totals and column totals are collectively called **marginal totals**. Also shown in the bottom right-hand corner of Table 4.2 is the overall total: 10 396.

Under the null hypothesis of no association between gestational age and asthma, the probability of being hospitalized for asthma is the same in each group (Pre-term, Term and Post-term). Conditional on the marginal totals given in Table 4.2, this probability is 465/10 396. We would expect this proportion of the 284 children who were pre-term babies to be hospitalized for asthma. So the expected frequency of pre-term children hospitalized for asthma is

$$\frac{465}{10\,396} \times 284 \simeq 12.70.$$

Similarly, the expected frequency of term children not hospitalized for asthma is obtained by multiplying the overall proportion not hospitalized (9931/10 396) by the number in the term group (8967); this gives

$$\frac{9931}{10\,396} \times 8967 \simeq 8565.92.$$

In general, the expected frequency in any given cell under the hypothesis of no association is obtained using the following formula:

$$\text{expected frequency} = \frac{\text{row total} \times \text{column total}}{\text{overall total}}. \tag{4.1}$$

Note that the expected frequency is generally not an integer.

Example 4.2 Calculation of expected frequencies under the hypothesis of no association

Applying Formula (4.1) to the data in Table 4.2 leads to the expected frequencies shown in Table 4.3.

Table 4.3 Expected frequencies for childhood asthma data

Gestational age	Hospitalized for asthma	Not hospitalized for asthma	Total
Pre-term	12.70	271.30	284
Term	401.08	8565.92	8 967
Post-term	51.21	1093.79	1 145
Total	464.99	9931.01	10 396

Notice that the entries in the column labelled 'Not hospitalized for asthma' in Table 4.3 sum to 9931.01, not 9931 (as in Table 4.2); and the entries in the column 'Hospitalized for asthma' sum to 464.99, not 465 (as in Table 4.2). These discrepancies are due to rounding errors; they are not important and may be ignored.

Overall, the differences between the observed frequencies in Table 4.2 and the expected frequencies in Table 4.3 appear to be quite small. ♦

Examples 4.1 and 4.2 used data from a cohort study. The method for obtaining the expected frequencies under the null hypothesis of no association is exactly the same for case-control studies as for cohort studies.

Activity 4.1 Sleeping position and SIDS

In Example 3.4, odds ratios and confidence intervals were calculated for data on sleeping position and sudden infant death syndrome (SIDS). The data from Table 3.7 are reproduced in Table 4.4. The row totals and the overall total are also given.

Table 4.4 Sleeping position and deaths from SIDS

Position baby last placed down to sleep	Cases	Controls	Total
On its front	30	24	54
On its side	76	241	317
On its back	82	509	591
Total	188	774	962

(a) Obtain the expected frequencies under the null hypothesis of no association between sleeping position and SIDS.

(b) In which cells are the observed numbers of cases greater than expected? What does this suggest about a possible association between sleeping position and SIDS?

In Example 4.2, the observed frequencies were quite close to those expected. In Activity 4.1, the differences were larger; for example, 30 cases were last placed down on their front, compared to 10.55 expected if there were no association. This might suggest that the null hypothesis of no association should be rejected. However, since the differences might be due to random variation, a formal significance test is required before a conclusion can be drawn.

Consider data from a cohort study or a case-control study, arranged in a table with r rows (usually exposure groups) and c columns (usually disease outcome groups). Thus there are $r \times c$ cells in the table, not counting the margins. Let O_i denote the *observed* count in the ith cell, and let E_i denote the *expected* frequency for that cell, calculated under the null hypothesis of no association. The first step to constructing the significance test is to decide upon a test statistic. This should reflect the magnitudes of the differences $O_i - E_i$. A convenient test statistic is

The symbol χ is the Greek letter chi. It is pronounced 'kye'.

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}.$$

This statistic is called the **chi-squared test statistic**. It was devised by Karl Pearson, so it is sometimes called the Pearson chi-squared statistic. It is the basis of a commonly used significance test known as the **chi-squared test** for no association. If the value of χ^2 is 0, this indicates perfect agreement between the observed frequencies and the expected frequencies. In general, the greater the differences are between the observed frequencies and those expected under the null hypothesis of no association, the larger is the value of χ^2.

Example 4.3 The chi-squared test statistic for the childhood asthma data

In Example 4.1, data were presented from a cohort study on the relationship between gestational age (categorized as Pre-term, Term and Post-term) and hospitalization for asthma during childhood. The expected frequencies under the hypothesis of no association were calculated in Example 4.2. The observed frequencies and the expected frequencies are shown together in Table 4.5 (the expected frequencies are in brackets).

Table 4.5 Observed frequencies O_i and expected frequencies E_i (in brackets) for childhood asthma data

Gestational age	Hospitalized for asthma	Not hospitalized for asthma	Total
Pre-term	18 (12.70)	266 (271.30)	284
Term	402 (401.08)	8565 (8565.92)	8967
Post-term	45 (51.21)	1100 (1093.79)	1145

So the value of the chi-squared test statistic is given by

$$\begin{aligned}\chi^2 &= \sum_i \frac{(O_i - E_i)^2}{E_i} \\ &\simeq \frac{(18-12.70)^2}{12.70} + \frac{(402-401.08)^2}{401.08} + \frac{(45-51.21)^2}{51.21} \\ &\quad + \frac{(266-271.30)^2}{271.30} + \frac{(8565-8565.92)^2}{8565.92} + \frac{(1100-1093.79)^2}{1093.79} \\ &\simeq 2.2118 + 0.0021 + 0.7531 + 0.1035 + 0.0001 + 0.0353 \\ &\simeq 3.11.\end{aligned}$$

In this calculation the expected frequencies were rounded to two decimal places. If full accuracy is retained for the expected frequencies, then the value obtained for the chi-squared test statistic is 3.104. ♦

Activity 4.2 The chi-squared test statistic for the SIDS data

Using the expected frequencies that you calculated in Activity 4.1, obtain the value of the chi-squared test statistic for the data on sleeping position and SIDS in Table 4.4.

4.2 The chi-squared test for no association

In Subsection 4.1, the test statistic for the chi-squared test for no association was developed. In order to complete the test, the null distribution of the test statistic, that is its distribution under the null hypothesis, must be known. In this subsection, the null distribution is stated without proof and the test is described.

The null distribution of the chi-squared test statistic may be approximated by a standard continuous probability distribution known as a **chi-squared distribution**. This is a member of the chi-squared family of distributions, which is indexed by a parameter ν called the **degrees of freedom**; ν takes the values $1, 2, 3, \ldots$. The chi-squared distribution with ν degrees of freedom is denoted $\chi^2(\nu)$.

The probability density functions (p.d.f.s) for chi-squared distributions with 1, 2, 4 and 8 degrees of freedom are shown in Figure 4.1.

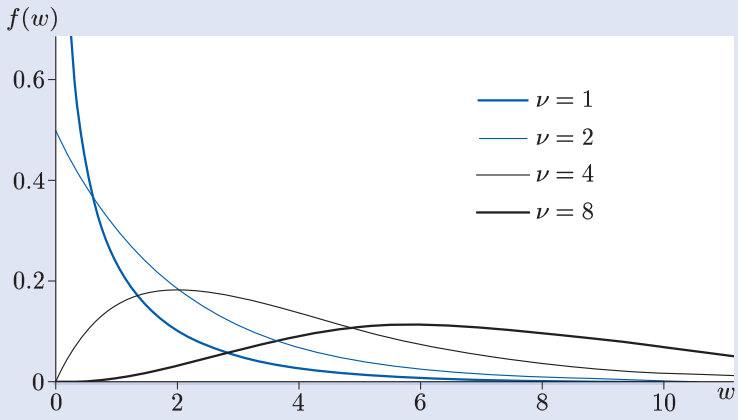

Figure 4.1 The probability density functions for four chi-squared distributions

The larger the value of the degrees of freedom parameter ν, the larger is the mean of the distribution and the more spread out the distribution. This is apparent for the four distributions represented in Figure 4.1. In fact, a random variable W with the chi-squared distribution on ν degrees of freedom has mean ν and variance 2ν. Note also that a chi-squared random variable W is defined for strictly positive values only.

The p.d.f. of the random variable $W \sim \chi^2(\nu)$ is rather complicated and is not given here. It is customary to use a computer or a table to obtain quantiles of $\chi^2(\nu)$. A table of quantiles for a range of values of ν is given in the *Handbook*. Part of the table is reproduced in Table 4.6.

Table 4.6 Selected quantiles for chi-squared distributions

ν	0.80	0.90	0.95	0.975	0.99	0.995	0.999
1	1.64	2.71	3.84	5.02	6.63	7.88	10.83
2	3.22	4.61	5.99	7.38	9.21	10.60	13.82
3	4.64	6.25	7.81	9.35	11.34	12.84	16.27
4	5.99	7.78	9.49	11.14	13.28	14.86	18.47

Example 4.4 Using the table

The 0.95-quantile of $\chi^2(1)$ is the number in Table 4.6 in the row labelled 1 (corresponding to $\nu = 1$) and in the column headed 0.95. This is 3.84. Similarly, the 0.99-quantile of $\chi^2(3)$ is 11.34. ♦

Activity 4.3 Tail probabilities for chi-squared distributions

Use the table of quantiles for chi-squared distributions in the *Handbook* to answer the following questions.

(a) Find the 0.95-quantile of $\chi^2(5)$.

(b) Find the value w such that $P(W > w) = 0.025$, where $W \sim \chi^2(8)$.

(c) Find the best possible lower bound and the best possible upper bound available from the tables for $P(W > 10.25)$, where $W \sim \chi^2(3)$.

A table with r rows and c columns (excluding row totals and column totals) is called an $r \times c$ table; this is read 'r by c'. For instance, Table 4.5 is a 3×2 table. The value of the degrees of freedom parameter ν for the chi-squared test for no association in such a table is $(r-1) \times (c-1)$.

For example, for a 3×2 table, that is, one with three rows and two columns (so that $r = 3$ and $c = 2$), the degrees of freedom are $\nu = (3-1) \times (2-1) = 2$.

The degrees of freedom have a direct interpretation in terms of the maximum number of cells that can be specified freely, subject to the constraints imposed by the marginal totals. This is illustrated in Example 4.5.

Example 4.5 Interpretation of the degrees of freedom

Table 4.4 has three rows and two columns: $r = 3$ and $c = 2$. The layout of the table, emptied of its contents except for the marginal totals, is shown in Figure 4.2(a).

Now try and enter numbers into the six cells of the table in such a way that the row totals and column totals are respected. For example, start with the number 50 in the top left-hand cell (you have some freedom of choice here). The value in the cell in row 1, column 2 must then be 4, since the values in row 1 must add up to 54. The number 50 was a free choice and is entered in bold in Figure 4.2(b); the number 4 was not, and is entered in italics.

Moving to row 2, enter a number in the first column. (You also have some freedom in selecting this value.) Suppose that you pick the number 100. Since the numbers in row 2 must sum to 317, the value in column 2 must be 217. These values are shown in Figure 4.2(c).

But now, note that the remaining two empty cells in the table are also determined: since the numbers in column 1 sum to 188, the remaining entry in column 1 must be 38; and, similarly, the remaining entry in column 2 must be 553. (See Figure 4.2(d).)

Thus you are able to exercise some choice in filling only two of the six cells of the table. Another way of describing this fact is that, conditional on the marginal totals, the table has two degrees of freedom.

Similarly, for a 3×3 table, there are $(3-1) \times (3-1) = 4$ degrees of freedom, meaning that you have some discretion in filling in four of the nine cells of the table, the values in the remaining five cells then being wholly determined by the marginal totals. ♦

		54
		317
		591
188	774	

(a)

50	*4*	54
		317
		591
188	774	

(b)

50	*4*	54
100	*217*	317
		591
188	774	

(c)

50	*4*	54
100	*217*	317
38	*553*	591
188	774	

(d)

Figure 4.2 Filling in the cells

Return now to the chi-squared test for no association. You have seen that the null distribution of the test statistic is approximately $\chi^2(\nu)$. This approximation is adequate provided that all the expected frequencies E_i are at least 5. If this is the case, then the chi-squared distribution may be used to calculate the significance probability, or p value, for the test. The chi-squared test statistic measures the extent to which observed frequencies differ from those expected under the hypothesis of no association: the higher the value of χ^2, the greater the discrepancy between the observed and expected frequencies. Thus the test is one-sided: only high values of χ^2 provide evidence against the hypothesis of no association. So only the upper tail of $\chi^2(\nu)$ is used in calculating the p value.

The p value can be calculated using a computer, or tables can be used to obtain an approximate value for p as in part (c) of Activity 4.3. Small p values suggest that data as extreme as those observed are unlikely to have arisen by chance if the null hypothesis of no association is true. So a small p value is interpreted as evidence against the null hypothesis of no association.

Example 4.6 Sleeping position and SIDS: calculating the p value

In Activity 4.2, you found that, for the data in Table 4.4 on SIDS and sleeping position, the value of the chi-squared test statistic χ^2 is 60.60. Since the data form a 3×2 table, the null distribution of the test statistic is approximately $\chi^2(2)$. The expected frequencies are all at least 5, so the approximation is adequate.

From Table 4.6, the 0.999-quantile of $\chi^2(2)$ is 13.82. Since the observed value of the test statistic, 60.60, is greater than this, it follows that the p value of the test for no association is less than $0.001 : p < 0.001$. In fact, using a computer gives a p value of 6.93×10^{-14}; such a tiny value would usually be reported as $p < 0.0001$.

This small p value indicates that data as extreme as those observed are very unlikely to have arisen by chance if the null hypothesis of no association were true. This is strong evidence that the null hypothesis is not true. The conclusion is that there is a statistically significant association between sleeping position and SIDS. ♦

In the context of cohort studies and case-control studies, the p value can be interpreted in terms of evidence against the null hypothesis of no association. For example, an estimated odds ratio greater than 1 or less than 1 provides *some* evidence of an association (only if it were exactly equal to 1 could it truly be said to provide *no* evidence of an association). However, if the p value of the test for no association is large (greater than 0.1, say), then there is little evidence against the null hypothesis of no association from this particular study, and hence little evidence of an association. A rough guide to the interpretation of p values in the context of cohort studies and case-control studies is provided in Table 4.7.

Table 4.7 Interpretation of p values for the test for no association

Significance probability p	Rough interpretation
$p > 0.10$	little evidence of an association
$0.10 \geq p > 0.05$	weak evidence of an association
$0.05 \geq p > 0.01$	moderate evidence of an association
$p \leq 0.01$	strong evidence of an association

The thresholds in Table 4.7 are to some extent arbitrary. For example, p values of 0.049 and 0.051 should lead to broadly similar conclusions.

Activity 4.4 Childhood asthma: calculating and interpreting the p value

In Example 4.3, the value of the chi-squared test statistic for the test for no association between gestational age and childhood asthma was found to be 3.11.

(a) Obtain a range of values for the p value for the test for no association between gestational age and asthma.

(b) Interpret your results.

The procedure for the chi-squared test for no association between the variables in an $r \times c$ table is summarized in the following box.

The chi-squared test for no association in an $r \times c$ table

1 Calculate a table of expected frequencies under the null hypothesis of no association: the expected frequency for a cell is given by

$$\text{expected frequency} = \frac{\text{row total} \times \text{column total}}{\text{overall total}}.$$

2 Calculate the value of the chi-squared test statistic using the formula

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i},$$

where O_i is the observed frequency and E_i is the expected frequency for the ith cell, and where the summation is over the $r \times c$ cells in the table.

3 Obtain the null distribution of the test statistic: under the null hypothesis of no association, the distribution of the test statistic is approximately chi-squared with $\nu = (r-1) \times (c-1)$ degrees of freedom.

The approximation is adequate provided that all the expected frequencies are at least 5.

4 Calculate the p value for the test and interpret your answer.

Activity 4.5 Ectopic pregnancy and history of genital infections

In Activity 3.3, you calculated odds ratios and confidence intervals using data on ectopic pregnancy and history of genital infections. These data are reproduced in Table 4.8. Carry out the chi-squared test for no association using these data.

Table 4.8 Ectopic pregnancy and history of genital infections

History	Cases	Controls
PID	212	112
Non-PID	157	407
None	411	1154
Total	780	1673

Note that the *strength of evidence* of an association (as quantified by a p value) is not the same thing as the *strength of the association* (as quantified by the odds ratio). For example, a large study can provide strong evidence of a weak association: the p value may be very small, though the odds ratio is close to unity. In medical statistics, both strength of evidence and strength of association are of interest, so it is usual to carry out a significance test and also to quote the odds ratio or relative risk and a confidence interval.

Activity 4.6 will give you some practice in carrying out the chi-squared test for no association, calculating confidence intervals, and reporting the results.

Activity 4.6 Smoking and lung cancer

In Example 3.1, data from the historic 1950 case-control study of smoking and lung cancer by Doll and Hill were presented. The data are reproduced in Table 4.9.

Table 4.9 Smoking and lung cancer in males

Exposure category	Cases of lung cancer	Controls
Smoked	647	622
Never smoked	2	27
Total	649	649

(a) Test the null hypothesis that smoking and lung cancer in males are not associated.

(b) Estimate the odds ratio for smoking and lung cancer in males, and obtain a 95% confidence interval for the odds ratio.

(c) Interpret your results.

4.3 Fisher's exact test

When no expected frequencies are less than 5, the null distribution of the chi-squared test statistic is well approximated by a chi-squared distribution, and the chi-squared test gives reliable results. However, when one or more expected frequencies are less than 5, the chi-squared test may give unreliable results. Such a situation is described in Example 4.7.

Example 4.7 Measles case fatality rate in South Africa

In South Africa, measles vaccination campaigns were conducted in 1996 and 1997 with the aim of stopping the circulation of the measles virus. To evaluate the campaign, data on measles were collected corresponding to the periods before and after the campaign.

Uzicanin, A., Eggers, R., Webb, E. *et al.* (2002) Impact of the 1996–1997 supplementary measles vaccination campaigns in South Africa. *International Journal of Epidemiology*, **31**, 968–976.

The mass campaign very substantially reduced the incidence of measles. In the Western Cape Province, there were 736 cases of measles hospitalized in the period 1992–97, prior to the campaign, of whom 23 died. Between the beginning of 1998 and July 1999, after the campaign, there were only 29 cases of measles sufficiently serious to be hospitalized, none of whom died. The data are shown in Table 4.10.

Table 4.10 Measles hospitalizations and deaths

Period	Died	Did not die	Total
After campaign	0	29	29
Before campaign	23	713	736
Total	23	742	765

The number expected to die after the campaign under the hypothesis of no association between the measles vaccination campaign and the proportion of serious cases who died is, using (4.1),

$$\frac{29 \times 23}{765} \simeq 0.87.$$

So the observed number, zero, is not very different from that expected: the vaccination campaign reduced the number of measles cases, but probably did not affect the proportion of cases who died. But how can this be tested formally? The chi-squared test for no association may not be valid in this instance as the expected frequency for the zero cell is less than 5. ♦

In fact, there is an exact test that applies in all circumstances. This is **Fisher's exact test**, named after the renowned statistician Ronald Aylmer Fisher. The test involves working through all possible tables with the same marginal totals as the table observed, and summing the probabilities of those that are as extreme or more extreme than the observed table. This is a time-consuming procedure, and is best done by a computer. Fisher originally developed the test for use with 2×2 tables, and it was later generalized to tables of arbitrary dimensions. For large tables, the computations can be extremely demanding, even with the use of a computer. So application of the test will be left to the computer book.

Summary of Section 4

In this section, the chi-squared test of the null hypothesis of no association has been described in the context of cohort studies and case-control studies. The chi-squared test statistic is a measure of discrepancy between the observed frequencies and the frequencies expected under the null hypothesis. The null distribution is approximately a chi-squared distribution. The approximation is good and the chi-squared test gives reliable results when none of the expected frequencies is less than 5. In other cases, Fisher's exact test may be used.

Exercise on Section 4

Exercise 4.1 Seat belts and children's safety in car accidents

In Activity 1.2, data on seat belt use and injuries sustained by children in car accidents were described. The data are reproduced in Table 4.11.

Table 4.11 Seat belt use and injury sustained by children aged 4–14

	D: sustained at least moderately severe injury		
Exposure category	Yes	No	Total
Not wearing a seat belt (E)	14	19	33
Wearing a seat belt (not E)	13	39	52

Halman, I., Chipman, M., Parkin, P.C. and Wright, J.G. (2002) Are seat belt restraints as effective in school age children as in adults? A prospective crash study. *British Medical Journal*, **324**, 1123–1125.

Carry out a chi-squared test for no association between children sustaining at least moderately severe injury and failing to wear a seat belt.

Solutions to Activities

Solution 1.1

(a) The exposure E is compulsory redundancy. The disease D is serious self-inflicted injury leading to hospitalization or death.

(b) The exposed group comprises the Whakatu workers. The control group comprises the Tomoana workers.

(c) The data from the study can be arranged as shown in Table S.1.

Table S.1 Serious self-inflicted injury (SSII) and compulsory redundancy in meat-processing workers in New Zealand, 1986–94

Exposure category	SSII	No SSII	Total
Made compulsorily redundant (Whakatu workers)	14	1931	1945
Not made compulsorily redundant (Tomoana workers)	4	1763	1767

Solution 1.2

(a) The estimated probabilities are

$$\widehat{P}(D|E) = \frac{14}{33} \simeq 0.42$$

and

$$\widehat{P}(D|\text{not } E) = \frac{13}{52} = 0.25.$$

(b) The estimated probability that a child sustains at least moderately severe injury is higher for children not wearing a seat belt than for children wearing a seat belt. However, it cannot be concluded that $P(D|E)$ is greater than $P(D|\text{not } E)$ in the population, since the observed difference might be due to random variation.

Solution 1.3

(a) The estimated relative risk is

$$\widehat{RR} = \frac{a/n_1}{c/n_2} = \frac{39/52}{19/33} \simeq 1.30.$$

(b) This estimate is less than the value 1.70 obtained in Example 1.3.

(c) Wearing a seat belt is associated with a 30% increase in the chance of avoiding moderate or worse injury. The strength of association between seat belt use and moderate or worse injury has not changed, but our measure of it has.

Solution 1.4

Using Formula (1.2),

$$\widehat{OR} = \frac{a \times d}{b \times c} = \frac{39 \times 14}{13 \times 19} \simeq 2.21.$$

This value is the same as that obtained for $\widehat{OR}$ in Example 1.4.

Solution 1.5

(a) The estimated relative risk is

$$\widehat{RR} = \frac{a/n_1}{c/n_2} = \frac{14/1945}{4/1767} \simeq 3.18.$$

The estimated odds ratio is

$$\widehat{OR} = \frac{a \times d}{b \times c} = \frac{14 \times 1763}{1931 \times 4} \simeq 3.20.$$

The estimated relative risk and the estimated odds ratio are both greater than 1. Thus both measures indicate that compulsory redundancy may be positively associated with SSII.

(b) The estimated relative risk and the estimated odds ratio are very similar. Serious self-inflicted injury is uncommon, so in this case the relative risk and the odds ratio nearly coincide.

(c) For a non-statistical audience the terms 'relative risk' and 'odds ratio' should be avoided. So an appropriate description of the results, which implicitly uses relative risks, is as follows.

'Workers who were made compulsorily redundant were three times more likely to suffer serious self-inflicted injury, compared to workers who were not made redundant.'

Solution 2.1

(a) The two groups are drivers with air bags, and passengers without air bags. Thus the effects of seat position and air bag use cannot be separated. The data suggest that the risk of death is lower for drivers with air bags compared to passengers without air bags (0.53 compared to 0.65). However, it is not possible, from these data alone, to infer whether this is due to an association with seat position, air bags, or both.

(b) The random variables X and Y denoting numbers of fatalities in the two groups are not independent, since the data are obtained from driver-passenger pairs. For example, in very severe crashes there is a higher probability that both driver and passenger will be killed than in less severe crashes.

Solution 2.2

(a) The estimated relative risk is given by (2.2):

$$\widehat{RR} = \frac{a/n_1}{c/n_2} = \frac{156/9577}{1531/16\,328} \simeq 0.1737.$$

(b) The estimated standard error $\widehat{\sigma}$ of RR is given by (2.4):

$$\begin{aligned}\widehat{\sigma} &= \sqrt{\frac{1}{a} - \frac{1}{n_1} + \frac{1}{c} - \frac{1}{n_2}} \\ &= \sqrt{\frac{1}{156} - \frac{1}{9577} + \frac{1}{1531} - \frac{1}{16\,328}} \\ &\simeq 0.08305.\end{aligned}$$

For a 99% confidence interval, the 0.995-quantile of the standard normal distribution is required; this is $z = 2.576$. The confidence limits are given by (2.3):

$$\begin{aligned} RR^- &= \widehat{RR} \times \exp(-z \times \widehat{\sigma}) \\ &\simeq 0.1737 \times \exp(-2.576 \times 0.08305) \\ &\simeq 0.14, \\ RR^+ &= \widehat{RR} \times \exp(z \times \widehat{\sigma}) \\ &\simeq 0.1737 \times \exp(2.576 \times 0.08305) \\ &\simeq 0.22. \end{aligned}$$

So a 99% confidence interval for RR is $(0.14, 0.22)$.

(c) The relative risk is 0.17, with 99% confidence interval $(0.14, 0.22)$. The estimate of RR and its confidence interval are located well below 1, indicating a negative association between measles vaccination and measles infection.

Solution 2.3

(a) The estimated odds ratio is

$$\widehat{OR} = \frac{a \times d}{b \times c} = \frac{20 \times 534}{126 \times 51} \simeq 1.6620.$$

(b) To calculate a confidence interval, the estimated standard error $\widehat{\sigma}$ is required:

$$\begin{aligned} \widehat{\sigma} &= \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}} \\ &= \sqrt{\frac{1}{20} + \frac{1}{126} + \frac{1}{51} + \frac{1}{534}} \\ &\simeq 0.2818. \end{aligned}$$

For a 95% confidence interval, the 0.975-quantile of the standard normal distribution is required, namely $z = 1.96$. The 95% confidence limits are

$$\begin{aligned} OR^- &= \widehat{OR} \times \exp(-z \times \widehat{\sigma}) \\ &\simeq 1.6620 \times \exp(-1.96 \times 0.2818) \\ &\simeq 0.96, \\ OR^+ &= \widehat{OR} \times \exp(z \times \widehat{\sigma}) \\ &\simeq 1.6620 \times \exp(1.96 \times 0.2818) \\ &\simeq 2.89. \end{aligned}$$

So a 95% confidence interval for OR is $(0.96, 2.89)$.

(c) The odds ratio is 1.66, with 95% confidence interval $(0.96, 2.89)$. The confidence interval includes 1. This means that we cannot conclude that there is an association. Note, however, that this study does not rule out such an association: the data are also consistent with values of OR that are greater than 1.

Solution 3.1

(a) The exposure in this study is attendance at political meetings. A table similar to Table S.2 is required.

Table S.2 Homicide and political activity in Karachi

Exposure category	Cases	Controls
Attended political meetings	11	2
Did not attend political meetings	24	83
Total	35	85

(b) The proportion exposed in cases is $11/35 \simeq 0.31$ and in controls is $2/85 \simeq 0.02$. Thus the proportion exposed (that is, who attended political meetings) is considerably higher in cases (homicide victims) than in controls. This suggests that attending political meetings might be positively associated with death by homicide.

Solution 3.2

(a) The estimated odds ratio is

$$\widehat{OR} = \frac{a \times d}{b \times c} = \frac{62 \times 55}{76 \times 5} \simeq 8.9737.$$

(b) For a 95% confidence interval, the 0.975-quantile of the standard normal distribution is required, namely $z = 1.96$. The estimated standard error $\widehat{\sigma}$ is given by

$$\begin{aligned} \widehat{\sigma} &= \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}} \\ &= \sqrt{\frac{1}{62} + \frac{1}{76} + \frac{1}{5} + \frac{1}{55}} \\ &\simeq 0.4975. \end{aligned}$$

So the confidence limits are

$$\begin{aligned} OR^- &= \widehat{OR} \times \exp(-z \times \widehat{\sigma}) \\ &\simeq 8.9737 \times \exp(-1.96 \times 0.4975) \\ &\simeq 3.38, \\ OR^+ &= \widehat{OR} \times \exp(z \times \widehat{\sigma}) \\ &\simeq 8.9737 \times \exp(1.96 \times 0.4975) \\ &\simeq 23.79. \end{aligned}$$

So a 95% confidence interval for the odds ratio is $(3.38, 23.79)$.

(c) The estimated odds ratio is 8.97. This means that the odds of SIDS for infants placed to sleep on their front is 8.97 times the odds of SIDS for infants laid down to sleep in other positions. The 95% confidence interval is $(3.38, 23.79)$. This lies well above 1. In conclusion, the data indicate that there exists a positive association between death from SIDS and putting the baby down to sleep on its front.

Solution 3.3

(a) A good choice of reference category is women with no history of genital infections, though the choice is to some extent arbitrary.

(b) Relative to this category, the estimated odds ratio for PID is

$$\widehat{OR} = \frac{212 \times 1154}{112 \times 411} \simeq 5.3147.$$

In this case,

$$\widehat{\sigma} = \sqrt{\frac{1}{212} + \frac{1}{112} + \frac{1}{411} + \frac{1}{1154}} \simeq 0.1302.$$

So the approximate 95% confidence limits are

$$OR^- \simeq 5.3147 \times \exp(-1.96 \times 0.1302) \simeq 4.12,$$

$$OR^+ \simeq 5.3147 \times \exp(1.96 \times 0.1302) \simeq 6.86.$$

Hence the 95% confidence interval is $(4.12, 6.86)$.

The odds ratio for Non-PID infections relative to None is

$$\widehat{OR} = \frac{157 \times 1154}{407 \times 411} \simeq 1.0831.$$

In this case,

$$\widehat{\sigma} = \sqrt{\frac{1}{157} + \frac{1}{407} + \frac{1}{411} + \frac{1}{1154}} \simeq 0.1101.$$

So the 95% confidence limits for this odds ratio are

$$OR^- \simeq 1.0831 \times \exp(-1.96 \times 0.1101) \simeq 0.87,$$

$$OR^+ \simeq 1.0831 \times \exp(1.96 \times 0.1101) \simeq 1.34.$$

Thus the 95% confidence interval is $(0.87, 1.34)$.

(c) There is a positive association between a pregnancy being ectopic and PID: the odds ratio is 5.31, and the 95% confidence interval $(4.12, 6.86)$ is located well above 1. However, for infections other than PID, there is little evidence of any association with ectopic pregnancy: the odds ratio is only 1.08, and the 95% confidence interval $(0.87, 1.34)$ contains 1.

Solution 4.1

(a) The frequencies expected under the null hypothesis of no association are given in Table S.3.

Table S.3 Expected frequencies for SIDS data

Position baby last placed down to sleep	Cases	Controls	Total
On its front	10.55	43.45	54
On its side	61.95	255.05	317
On its back	115.50	475.50	591
Total	188	774	962

For example, the expected frequency of cases among babies last placed on their front is given by

$$\frac{\text{row total} \times \text{column total}}{\text{overall total}} = \frac{54 \times 188}{962} \simeq 10.55.$$

(b) The numbers of cases observed are greater than expected for babies placed on their front or side. This suggests that there might be a positive association between SIDS and placing a baby down on its front or side. (However, this does not on its own prove the existence of such an association.)

Solution 4.2

The observed frequencies O_i are in Table 4.4 and the expected frequencies E_i are in Table S.3. The value of the chi-squared test statistic is given by

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

$$\simeq \frac{(30 - 10.55)^2}{10.55} + \frac{(76 - 61.95)^2}{61.95} + \frac{(82 - 115.50)^2}{115.50}$$

$$+ \frac{(24 - 43.45)^2}{43.45} + \frac{(241 - 255.05)^2}{255.05} + \frac{(509 - 475.50)^2}{475.50}$$

$$\simeq 35.8581 + 3.1865 + 9.7165 + 8.7066 + 0.7740 + 2.3601$$

$$\simeq 60.60.$$

Solution 4.3

(a) The 0.95-quantile of $\chi^2(5)$ is 11.07.

(b) The value required is the 0.975-quantile of $\chi^2(8)$, namely 17.53.

(c) Looking along the row of the table corresponding to $\nu = 3$, the value 10.25 lies between the 0.975-quantile (9.35) and the 0.99-quantile (11.34). Thus

$$0.01 < P(W > 10.25) < 0.025.$$

Hence, from the table, the best lower bound for the probability is 0.01 and the best upper bound is 0.025. (The value of the probability is, in fact, approximately 0.01656.)

Solution 4.4

(a) To calculate the p value, the null distribution of the test statistic is required. Since Table 4.5 is a 3×2 table, the null distribution is approximately chi-squared with degrees of freedom $\nu = (3-1) \times (2-1) = 2$. Since the expected frequencies are all greater than 5, the approximation is adequate.

The 0.80-quantile of $\chi^2(2)$ is 3.22, which is greater than 3.11, the observed value of χ^2. Thus the p value is greater than 0.2.

(b) Since $p > 0.2$, there is little evidence of association between childhood asthma and gestational age.

Solution 4.5

The expected frequencies under the null hypothesis of no association are in Table S.4.

Table S.4 Ectopic pregnancy and history of genital infections — expected frequencies

History	Cases	Controls	Total
PID	103.02	220.98	324
Non-PID	179.34	384.66	564
None	497.64	1067.36	1565
Total	780	1673	2453

The value of the chi-squared test statistic is

$$\chi^2 = \frac{(212-103.02)^2}{103.02} + \frac{(157-179.34)^2}{179.34} + \frac{(411-497.64)^2}{497.64} + \frac{(112-220.98)^2}{220.98} + \frac{(407-384.66)^2}{384.66} + \frac{(1154-1067.36)^2}{1067.36} \simeq 195.23.$$

The null distribution of the test statistic is approximately chi-squared with degrees of freedom $\nu = (3-1)\times(2-1) = 2$. Since all expected frequencies are at least 5, the approximation is adequate.

The 0.999-quantile of the $\chi^2(2)$ distribution is 13.82. The observed value of the test statistic is greater than this, so the p value is less than 0.001. There is strong evidence of an association between prior genital infection and ectopic pregnancy.

Solution 4.6

(a) The expected frequencies are shown in brackets in Table S.5.

Table S.5 Observed and expected frequencies for lung cancer data

Exposure category	Cases	Controls	Total
Smoked	647 (634.50)	622 (634.50)	1269
Never smoked	2 (14.50)	27 (14.50)	29
Total	649	649	1298

The observed value of the chi-squared test statistic is

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i} = \frac{(647-634.50)^2}{634.50} + \frac{(622-634.50)^2}{634.50} + \frac{(2-14.50)^2}{14.50} + \frac{(27-14.50)^2}{14.50} \simeq 0.2463 + 0.2463 + 10.7759 + 10.7759 \simeq 22.04.$$

The null distribution of the test statistic is approximately chi-squared with degrees of freedom $\nu = (2-1)\times(2-1) = 1$. Since the expected frequencies are all at least 5, the approximation is adequate. (Note that one *observed* value is less than 5, but this does not matter.)

Using tables, the 0.999-quantile of $\chi^2(1)$ is 10.83. Since the value of the test statistic is greater than this, the p value is less than 0.001. There is strong evidence of an association between smoking and lung cancer.

(b) The estimated odds ratio is

$$\widehat{OR} = \frac{647\times 27}{622\times 2} \simeq 14.0426 \simeq 14.04.$$

To calculate an approximate 95% confidence interval, the estimated standard error $\widehat{\sigma}$ is required:

$$\widehat{\sigma} = \sqrt{\frac{1}{647} + \frac{1}{622} + \frac{1}{2} + \frac{1}{27}} \simeq 0.7350.$$

The 95% confidence limits are

$$OR^- \simeq 14.0426 \times \exp(-1.96\times 0.7350) \simeq 3.33,$$
$$OR^+ \simeq 14.0426 \times \exp(1.96\times 0.7350) \simeq 59.31.$$

So the 95% confidence interval for the odds ratio is $(3.33, 59.31)$.

(c) There is strong evidence of an association between smoking and lung cancer ($p < 0.001$). The estimated odds ratio is 14.04, with 95% confidence interval $(3.33, 59.31)$. This indicates a strong positive association. In conclusion, this study provides *strong evidence* of a *strong positive association* between smoking and lung cancer.

Solutions to Exercises

Solution 1.1

(a) The estimates are as follows:

$$\widehat{RR} = \frac{a/n_1}{c/n_2} = \frac{327/542}{76/277} \simeq 2.20,$$

$$\widehat{OR} = \frac{a \times d}{b \times c} = \frac{327 \times 201}{215 \times 76} \simeq 4.02.$$

(b) The relative risk and the odds ratio are both greater than 1. This suggests a positive association between pre-eclampsia or eclampsia during the first pregnancy and hypertension later in life.

Solution 2.1

(a) In Example 2.1, the odds ratio was estimated to be $\widehat{OR} \simeq 2.41$. As the odds ratio is to be used in intermediate calculations, greater accuracy is required. To four decimal places,

$$\widehat{OR} = \frac{893 \times 2783}{5724 \times 180} \simeq 2.4121.$$

To calculate a confidence interval, the estimated standard error $\widehat{\sigma}$ is required:

$$\begin{aligned}\widehat{\sigma} &= \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}} \\ &= \sqrt{\frac{1}{893} + \frac{1}{5724} + \frac{1}{180} + \frac{1}{2783}} \\ &\simeq 0.08491.\end{aligned}$$

For a 99% confidence interval, the 0.995-quantile of $N(0,1)$ is required, namely $z = 2.576$. The confidence limits are

$$\begin{aligned}OR^- &= \widehat{OR} \times \exp(-z \times \widehat{\sigma}) \\ &\simeq 2.4121 \times \exp(-2.576 \times 0.08491) \\ &\simeq 1.94, \\ OR^+ &= \widehat{OR} \times \exp(z \times \widehat{\sigma}) \\ &\simeq 2.4121 \times \exp(2.576 \times 0.08491) \\ &\simeq 3.00.\end{aligned}$$

Thus a 99% confidence interval for the odds ratio is $(1.94, 3.00)$.

(b) If deployment to the Gulf was not associated with PTSD then OR would be 1. This is implausible since the 99% confidence interval for OR is $(1.94, 3.00)$ and hence is located well above 1.

Solution 3.1

(a) In a cohort study, the groups would include a group of vaccinated children (the exposed group) and a group of unvaccinated children. Both groups would be followed through time and the numbers of cases of Hib meningitis in the two groups would be counted and compared.

(b) In a case-control study, the two groups would be a group of Hib meningitis cases and a group of children without Hib meningitis. The numbers of children previously vaccinated against Hib in the two groups would then be identified.

(c) Since Hib meningitis is rare, a cohort study would have to be very large. An advantage of a case-control study is that it would not need to be so large.

(d) The odds ratio provides a good approximation to the relative risk when the disease is rare, as is the case here.

Solution 4.1

The expected frequencies under the null hypothesis of no association are in Table S.6.

Table S.6 Wearing a seat belt and sustaining at least moderately severe injury — expected frequencies

	Sustained at least moderately severe injury		
Exposure category	Yes	No	Total
Not wearing a seat belt	10.48	22.52	33
Wearing a seat belt	16.52	35.48	52
Total	27	58	85

The number of children not wearing a seat belt who sustained at least moderately severe injury (14) is greater than expected (10.48).

The value of the test statistic is

$$\begin{aligned}\chi^2 &= \frac{(14 - 10.48)^2}{10.48} + \frac{(19 - 22.52)^2}{22.52} \\ &\quad + \frac{(13 - 16.52)^2}{16.52} + \frac{(39 - 35.48)^2}{35.48} \\ &\simeq 2.83.\end{aligned}$$

The null distribution of the test statistic is approximately chi-squared with degrees of freedom $\nu = (2-1)(2-1) = 1$. Since all the expected frequencies are at least 5, the approximation is adequate.

The 0.9-quantile of $\chi^2(1)$ is 2.71 and the 0.95-quantile is 3.84. The observed value of the test statistic lies between these values, so $0.05 < p < 0.1$. Thus, for children aged 4–14, there is weak evidence of an association between not wearing a seat belt and sustaining at least moderately severe injury in the event of a car accident.

Index