2

Basic Concepts from Traditional Statistical Analysis

The Bayesian approach, to a considerable extent, supplements rather than replaces the kind of analyses traditionally carried out in assessing health-care interventions, and in this chapter we shall briefly review some of the basic ideas that will subsequently be found useful. In particular, probability theory is fundamental to Bayesian analysis, and we therefore revise the basic concepts with a natural emphasis on Bayes theorem. We also consider random variables and probability distributions with particular emphasis on the normal distribution, which plays a vital role in summarising what the observed data can tell us about unknown quantities of interest. A particularly important practical aspect is the transformation of output from standard statistical packages into a form amenable to Bayesian interpretation.

Bayesian analysis makes a much wider use of probability distributions than traditional statistical methods, in that not only are sampling distributions required for summaries of data, but also a wide range of distributions are used to represent prior opinion about proportions, event rates, and other unknown quantities. The *shapes* of distributions therefore become particularly important, as they are intended to represent the plausibility of different values, and so we shall provide (in starred sections) extensive graphical displays as well the usual formulae.

Most of the issues addressed in this chapter are covered in a concise and readable manner in standard textbooks such as Altman (2001) and Berry *et al.* (2001b). In addition, Clayton and Hills (1993) consider a likelihood-based approach to many of the models that are frequently encountered in epidemiology and health-care evaluation.

Bayesian Approaches to Clinical Trials and Health-Care Evaluation D. J. Spiegelhalter, K. R. Abrams and J. P. Myles
 ISBN: 0-471-49975-7

2.1 PROBABILITY

2.1.1 What is probability?

Suppose a is some event which may or may not take place, such as the next toss of a coin coming up heads. Although we may casually speak of the 'probability' of a occurring, and give it a mathematical notation $p(a)$, it is perhaps remarkable that there is no universally agreed definition of what this term means. Perhaps the currently most accepted interpretation is the following: $p(a)$ is the proportion of times a will occur in an infinitely long series of repeated identical situations. This is known as the 'frequentist' perspective, as it rests on the frequency with which specific events occur. However, a number of other interpretations of probability have been made throughout history, and we shall consider a different, 'subjective', definition in Section 3.1.

There is little dispute, however, about the mathematical properties of probability. Let a and b be events, and H represent the context in which a and b might arise, and let $p(a|H)$ denote the probability of a given the context H: the vertical line represents 'conditioning'. Then $p(a|H)$ is a number that satisfies the following three basic rules:

1. *Bounds*.

$$0 \leq p(a|H) \leq 1,$$

   where $p(a|H) = 0$ if a is impossible and $p(a|H) = 1$ if a is certain in the context H.

2. *Addition rule*. If a and b are mutually exclusive (*i.e.* one at most can occur),

$$p(a \text{ or } b|H) = p(a|H) + p(b|H).$$

   (We note that, for technical reasons, it is helpful if Rule 2 is taken as holding for an infinite set of mutually exclusive events.)

3. *Multiplication rule*. For any events a and b,

$$p(a \text{ and } b|H) = p(a|b,H)p(b|H).$$

We say that a and b are independent if $p(a \text{ and } b|H) = p(a|H)p(b|H)$ or equivalently $p(a|b,H) = p(a|H)$: thus the fact that b has occurred does not alter the probability of a. The multiplication rule can equivalently be expressed as the definition of conditional probability,

$$p(a|b,H) = \frac{p(a \text{ and } b|H)}{p(b|H)},$$

provided $p(b|H) \neq 0$.

The explicit introduction of the context H is unusual in standard texts and we shall subsequently drop it to avoid accusations of pedantry: however, it is always useful to keep in mind that *all probabilities are conditional* and so, if the situation changes, then probabilities may change. We shall see in Section 3.1 that this notion forms the basis of *subjective probability*, in which H, the context, represents the information on which an individual bases their *own* subjective assessment of the *degree of belief, i.e.* probability, of an event occurring.

Example 2.1 illustrates that these rules can be given an immediate intuitive justification by comparison with a standard experiment.

Example 2.1 *Dice: Illustration of rules of probability*

Suppose H denotes the roll of two perfectly balanced six-sided dice, and let '$\equiv$' denote 'is equivalent to'.

Rule 1. For a single die: if $a \equiv$ 'throw 7', then $p(a) = 0$; if $a \equiv$ 'throw ≤ 6', then $p(a) = 1$. If c is the sum of the two dice: then if $c \equiv$ '13', then $p(c) = 0$; if $c \equiv$ '≤ 12', then $p(c) = 1$.

Rule 2. For a single die: if $a \equiv$ 'throw 3', $b \equiv$ 'throw 4', then

$$\begin{aligned} p(a \text{ or } b) &= p(a) + p(b) \text{ since } a \text{ and } b \text{ are mutually exclusive} \\ &= 1/6 + 1/6 = 1/3. \end{aligned}$$

Rule 3. If we throw two dice: if $a \equiv$ 'first die throw 2', $b \equiv$ 'second die throw 5', then

$$\begin{aligned} p(a \text{ and } b) &= p(a)p(b) \text{ since } a \text{ and } b \text{ are independent} \\ &= 1/6 \times 1/6 = 1/36. \end{aligned}$$

If $a \equiv$ 'total score of the two throws is greater than or equal to 6', $b \equiv$ 'first die throw 1', then

$$\begin{aligned} p(a \text{ and } b) &= p(a|b)p(b) \\ &= 1/3 \times 1/6 = 1/18. \end{aligned}$$

Suppose we also consider the events 'a and b' and 'a and $\overline{b}$', where $\overline{b}$ represents the event 'not b'. Then 'a and b' and 'a and $\overline{b}$' are mutually exclusive and together form the event a, and hence, using Rule 2, we have the identity

$$p(a) = p(a \text{ and } b) + p(a \text{ and } \overline{b}) \tag{2.1}$$

which is known as 'marginalisation'. Further, by using Rule 3, we obtain

$$p(a) = p(a|b)p(b) + p(a|\overline{b})p(\overline{b}), \tag{2.2}$$

which is known by the curious title of 'extending the conversation' (or 'extending the argument'). Example 2.2 shows these expressions follow naturally from considering the full 'joint' distribution over all possible combinations of events.

Example 2.2 *Prognosis: Marginalisation and extending the conversation*

Suppose we wish to determine the probability of survival (up to a specified point in time) following a particular cancer diagnosis, given that it depends on the stage of disease at diagnosis amongst other factors. Whilst directly specifying the probability of surviving, denoted b, may be difficult, by extending the conversation to include whether the cancer was at an early stage, denoted a, or not, denoted $\overline{a}$, we obtain from (2.1),

$$p(b) = p(b|a)p(a) + p(b|\overline{a})p(\overline{a}).$$

For example, suppose patients with early stage disease have a good prognosis, say $p(b|a) = 0.80$, but for late stage it is poor, say $p(b|\overline{a}) = 0.20$, and that of new diagnoses the majority, 90%, are early stage, *i.e.* $p(a) = 0.90$ and $p(\overline{a}) = 0.10$. Then the marginal probability of surviving is $p(b) = 0.80 \times 0.90 + 0.20 \times 0.10 = 0.74$.

Table 2.1 shows all possible combinations of events and their probabilities, as well as the marginal probabilities that, appropriately, appear in the margin of the table. The joint probabilities of events have been obtained by Rule 2 so that, for example, $p(b \text{ and } a) = p(b|a)p(a) = 0.80 \times 0.90 = 0.72$.

Table 2.1 Probabilities of all combinations of survival and stage, including marginal probabilities.

	Early stage a	Late stage $\overline{a}$	
Survive b	0.72	0.02	0.74
Not survive $\overline{b}$	0.18	0.08	0.26
	0.90	0.10	1.00

2.1.2 Odds and log-odds

Any probability p can also be expressed in terms of 'odds' O, where

$$O = \frac{p}{1-p}$$

and

$$p = \frac{O}{1+O},$$

so that, for example, a probability of 0.20 (20% chance) corresponds to odds of $O = 0.20/0.80 = 0.25$ or, in betting parlance, '4 to 1 against'. Conversely, betting odds of '7 to 4 against' correspond to $O = 4/7$, or a probability of $p = 4/11 = 0.36$.

The natural logarithm (denoted log) of the odds is termed the 'logit', so that

$$\text{logit}(p) = \log\left[\frac{p}{1-p}\right].$$

2.1.3 Bayes theorem for simple events

A number of properties can immediately be derived from Rules 1 to 3 of Section 2.1.1. Since $p(b \text{ and } a) = p(a \text{ and } b)$, Rule 3 implies that $p(b|a)p(a) = p(a|b)p(b)$, or equivalently

$$p(b|a) = \frac{p(a|b)}{p(a)} \times p(b). \tag{2.3}$$

We have proved Bayes theorem! In words, this vital result tells us how an initial probability $p(b)$ is changed into a conditional probability $p(b|a)$ when taking into account the event a occurring: it should be clear by this description that we are interpreting Bayes theorem as providing a formal mechanism for learning from experience.

Equation (2.3) also holds for $\overline{b}$, so that

$$p(\overline{b}|a) = \frac{p(a|\overline{b})}{p(a)} \times p(\overline{b}), \tag{2.4}$$

and dividing (2.3) by (2.4) we obtain the *odds form* for Bayes theorem:

$$\frac{p(b|a)}{p(\overline{b}|a)} = \frac{p(a|b)}{p(a|\overline{b})} \times \frac{p(b)}{p(\overline{b})}. \tag{2.5}$$

Thus $p(b)/p(\overline{b}) = p(b)/(1 - p(b))$, the odds on b before taking into account the event a, which is changed into the new odds $p(b|a)/p(\overline{b}|a)$ after conditioning on a. Equation (2.5) shows how Bayes theorem accomplishes this transformation without even explicitly calculating $p(a)$, and this insight is exploited in Section 3.2.

Example 2.3 *Prognosis (continued): Bayes theorem for single events*

Suppose we were given Table 2.1, and wanted to use Bayes theorem to tell us how knowing the stage of the disease at diagnosis revises our probability for survival a. Initially, before we know the stage, $p(b) = 0.74$ from the

marginal probability in Table 2.1. Suppose we find out that the disease is at an early stage, *i.e.* a, where we know from Table 2.1 that $p(a|b) = 0.72/0.74 = 0.97$ and $p(a) = 0.9$. Hence from (2.3) we obtain a revised probability of survival

$$p(b|a) = \frac{0.97}{0.9} \times 0.74 = 0.80,$$

matching what, in fact, we knew already.

To use the odds form of Bayes theorem (2.5) we first require the initial odds for survival, *i.e.* $p(b)/p(\overline{b}) = 0.74/0.26 = 2.85$, and the ratio $p(a|b)/p(a|\overline{b}) = 0.97/0.69 = 1.405$. Then from (2.5) we obtain the final odds on survival as $2.85 \times 1.41 = 4.01$, corresponding to a probability $p(b|a) = 0.80$ (up to rounding error).

The two forms of Bayes theorem both give the required results and can be thought of as a means of moving from a marginal probability in a table to a conditional probability having taken into account some evidence. As we shall see in Section 3.2, it is this use of Bayes theorem that is used in many diagnostic testing situations without any controversy.

2.2 RANDOM VARIABLES, PARAMETERS AND LIKELIHOOD

2.2.1 Random variables and their distributions

Random variables have a somewhat complex formal definition, but it is sufficient to think of them as unknown quantities that may take on one of a set of values: traditionally a random variable is denoted by a capital Latin letter, say Y, before being observed and by a lower-case letter y as a specific observed value. This convention tends to be broken in Bayesian analysis, in which all unknown quantities are considered as random variables, but we shall try to keep to it where it clarifies the exposition.

Loosely speaking, $p(y)$ denotes the probability of a random variable Y taking on each of its possible values y. $p(y)$ is formally known as the *probability density function*, and the probability that Y does not exceed y, $P(Y \leqslant y)$, is termed the *probability distribution function*. We shall tend to use 'probability distribution' as a generic term, hopefully without causing confusion.

Probability distributions may be:

Binary. When Y can take on one of two values, we shall generally use the notation $Y = 1$ for when an event of interest occurs, and $Y = 0$ when it does not: this is

known as a *Bernoulli trial*, after Jakob Bernoulli (1654–1705). The corresponding probability distribution obeys the rules $p(Y = 1) = 1 - p(Y = 0)$, and is said to have a Bernoulli distribution (Section 2.6.1); see Example 2.4.

Discrete. $p(y)$ forms a discrete distribution when Y can take on one of a list of values, say 0, 1, 2, 3, ... The binomial (Section 2.6.1) and Poisson (Section 2.6.2) distributions are used in this book.

Continuous. Suppose Y can, in theory, take on values measured to an arbitrary degree of precision (of course, in practice, rounding of measurements prevents this). This means that calculus is needed, and the probability of Y lying in any specified interval I is obtained by the integral $\int_I p(y)\, dy$. The continuous distributions met most often in this book are the normal (Section 2.3) and the uniform (Section 2.6.4), although a wide range of others are discussed in Section 2.6: many of these are useful as prior distributions for unknown quantities.

Following Rule 1 in Section 2.1.1, all probability distributions should assign total probability 1 to the set of all possible events – these are known as 'proper' probability distributions. For continuous distributions this would mean that they integrated to 1, *i.e.* $\int p(y)\, dy = 1$. In some theoretical exercises it can be useful to imagine 'improper' distributions that do not obey this rule, for example uniform distributions over the entire range $-\infty$ to ∞. In practice, however, all distributions used in our examples will be proper (this can in any case always be achieved by truncating such a distribution at very low and high values).

The expressions derived in Section 2.1 for simple events have their counterparts for continuous random variables x, y. To express how the probability of y is changed when taking into account an observation x, we write Bayes theorem as

$$p(y|x) = \frac{p(x|y)}{p(x)} \times p(y). \tag{2.6}$$

To obtain the (marginal) distribution $p(x)$ from the joint distribution $p(x,y)$, we require the continuous counterpart to (2.1),

$$p(x) = \int p(x,y)\, dy; \tag{2.7}$$

shows how this is particularly important in Bayesian analysis as there may be many unknown quantities but we may only be interested in one at a time. Finally, the notion of extending the conversation (see (2.2)), given by

$$p(x) = \int p(x|y)\, p(y)\, dy, \tag{2.8}$$

expresses how a conditional distribution $p(x|y)$ is 'averaged over' by a distribution $p(y)$ in order to produce a distribution on x.

Bayesian methods make repeated use of such integrations, and indeed the technical problems of carrying them out has, in the past, hampered the development of the approach. Fortunately, in subsequent chapters their use will be implicit and intuitive, with the necessary integrations made reasonably straightforward either by simplifying assumptions of normal distributions, or by using modern simulation methodology.

2.2.2 Expectation, variance, covariance and correlation

If we have a distribution, $p(y)$, for an unknown quantity, Y, and we require the expectation (mean) of Y then this is given by

$$\mathrm{E}(Y) = \sum_{i=1}^{k} y_i \; p(y_i) \tag{2.9}$$

if the distribution is discrete, and by

$$\mathrm{E}(Y) = \int y \; p(y) \; dy \tag{2.10}$$

if the distribution is continuous.

The variance of Y is defined as

$$\begin{aligned} \mathrm{V}(Y) &= \mathrm{E}(Y - \mathrm{E}(Y))^2 \\ &= \mathrm{E}(Y^2) - \mathrm{E}(Y)^2, \end{aligned}$$

which may be calculated, for example, using $\mathrm{E}(Y^2) = \int y^2 p(y) \; dy$. The standard deviation is then defined as $\mathrm{SD}(Y) = \sqrt{\mathrm{V}(Y)}$.

The 'covariance' of X and Y is defined as

$$\mathrm{Cov}(X,Y) = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y) \tag{2.11}$$

and measures the association between X and Y. However the covariance is not generally easy to interpret, and a better summary measure is the correlation, which is the covariance scaled by the standard deviations of the variables:

$$\mathrm{Corr}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\mathrm{SD}(X)\mathrm{SD}(Y)}. \tag{2.12}$$

$\mathrm{Corr}(X,Y)$ is a number between -1 and 1 which, loosely speaking, expresses how close X and Y are to lying on a straight line: $\mathrm{Corr}(X,Y)$ is near 1 for a positive relationship, near 0 when X and Y are unrelated, and near -1 for a negative relationship.

Conditional expectation and variance*

We return to the relationship between joint and marginal distributions introduced in (2.7). X has both a *conditional* mean and variance defined for each value y, *i.e.* $\mathrm{E}(X|y)$ and $\mathrm{V}(X|y)$, and a *marginal* mean and variance defined for the marginal distribution of X alone, *i.e.* $\mathrm{E}(X)$ and $\mathrm{V}(X)$. Their relationship can be shown to be as follows:

$$\mathrm{E}(X) = \mathrm{E}_Y[\mathrm{E}_X(X|Y)], \tag{2.13}$$

$$\mathrm{V}(X) = \mathrm{V}_Y[\mathrm{E}_X(X|Y)] + \mathrm{E}_Y[\mathrm{V}_X(X|Y)], \tag{2.14}$$

where the subscripts indicate the relevant variable for the expectation or variance. Some interpretation of these expressions might be obtained by assuming that Y will be the interim results of a study, and X will be the final results. Then (2.13) shows that our overall expectation of the final results can be calculated by first conditioning on the interim data as if they were known, and then taking our expectations (with respect to the interim data) of those conditional expectations. Equation (2.14) is more complex and says that our overall uncertainty about the final outcomes can be broken down into two components: our uncertainty about its conditional expectation given the interim data, and our expectation of its conditional variance.

We shall use these expressions in the context of prediction: first for normal variables in Section 3.13, and then in Section 9.8.3 within the context of microsimulation in complex cost-effectiveness models.

2.2.3 Parametric distributions and conditional independence

A central aspect of statistical inference is learning about the assumed underlying distribution of quantities we observe, and this is generally carried out by assuming that the probability distributions follow a particular *parametric* form $p(y|\theta)$, *i.e.* the distribution of Y depends on some currently unknown parameter θ. Parameters are usually given Greek letters: in Bayesian inference they are considered as random variables but the usual convention of capital and lowercase letters is ignored, to no apparent detriment.

For example, for a Bernoulli variable Y such that $p(Y = 0) = 1 - \theta$, $p(Y = 1) = \theta$, we may write this likelihood in the form

$$p(y|\theta) = \theta^y(1-\theta)^{1-y}; \qquad y = 0,\ 1. \tag{2.15}$$

A standard assumption in traditional statistics is that a set of random variables $Y_1, \ldots, Y_n$ are independent and identically distributed (i.i.d.). If we are willing to adopt a parametric distribution, this corresponds to assuming that each is drawn independently from a probability distribution $p(y|\theta)$ where θ is some unknown parameter or parameters, and hence by Rule 3 of Section 2.1.1 their joint distribution is

$$p(y_1, \ldots, y_n|\theta) = \prod_{i=1}^{n} p(y_i|\theta). \tag{2.16}$$

This is an example of what is known as *conditional independence*, since each Y_i is independent of the others, *conditional* on θ. We shall discuss in Section 3.4 how this expression can be derived rather than directly assumed.

2.2.4 Likelihoods

Much of traditional statistical inference is based on noting that, once data y have been observed, $p(y|\theta)$ can be considered as being a function of θ, and can tell us the extent to which different values of θ are *supported* by the data. When $p(y|\theta)$ is considered in this way it is known as the *likelihood*, and plays a very important role in Bayesian analysis, as it summarises all the information that the data y can provide about the parameter θ. It is important to note that any function of θ that is proportional to $p(y|\theta)$ can be considered as the likelihood, since multiplying $p(y|\theta)$ by any value that does not depend on θ does not affect the range of values of θ being supported.

The *likelihood function* expresses the relative plausibility of different values of θ, with the value of θ for which the likelihood is a maximum is referred to as the *maximum likelihood estimate*. We can use a range of values which are *best* supported by the data as an interval estimate for θ, and it can be argued (Clayton and Hills, 1993) that a reasonable range is defined by values of the likelihood above $\exp(-1.96^2/2) = 14.7\%$ of the maximum value – the reason for this choice will become apparent in Section 2.4.1. In practice, constructing intervals in such a manner is laborious, and in general we try to approximate likelihood functions by the normal distribution, as discussed in Section 2.4. Consider, for example, n individuals in a study; we measure whether the ith individual responds to treatment, $Y_i = 1$, or not, $Y_i = 0$. If we assume a set of independent Bernoulli trials such that the probability of response is θ, then, using (2.15) and (2.16), we can obtain the joint distribution for all n individuals as

$$\begin{aligned} p(y_1, \ldots, y_n|\theta) &= \prod_{i=1}^{n} p(y_i|\theta) \\ &= \prod_{i=1}^{n} \theta^{y_i}(1-\theta)^{1-y_i} \qquad (2.17) \\ &= \theta^{y_1+\ldots+y_n}(1-\theta)^{(1-y_1)+\ldots+(1-y_n)} \\ &= \theta^{y_1+\ldots+y_n}(1-\theta)^{n-(y_1+\ldots+y_n)} \\ &= \theta^{r}(1-\theta)^{n-r}, \qquad (2.18) \end{aligned}$$

where $r = \sum_i y_i$ is the number of responders. This likelihood is maximised at $\hat{\theta} = r/n$; hence the maximum likelihood estimate is the proportion of responders. The independence of the individual responses means that the probability (2.18) is the same regardless of the actual sequence, and hence if we were told that there were 3 successes out of 10 trials, our likelihood would be precisely the same.

Example 2.4 *Response: Combining Bernoulli likelihoods*

Suppose we observed the responses of 10 individuals to a drug, and the particular sequence observed is 0,1,0,0,0,1,0,1,0,0. Let θ be the probability of a random patient responding to the drug. There are 3 successes and 7 failures, and the probability of the data, *i.e.* the likelihood, is given by

$$p(y_1, \ldots, y_{10}|\theta) = \theta^3(1-\theta)^{10-3} = \theta^3(1-\theta)^7. \qquad (2.19)$$

Figure 2.1 shows this likelihood plotted for different values of θ and scaled to have maximum value 1. We return to this example in Section 2.4.1.

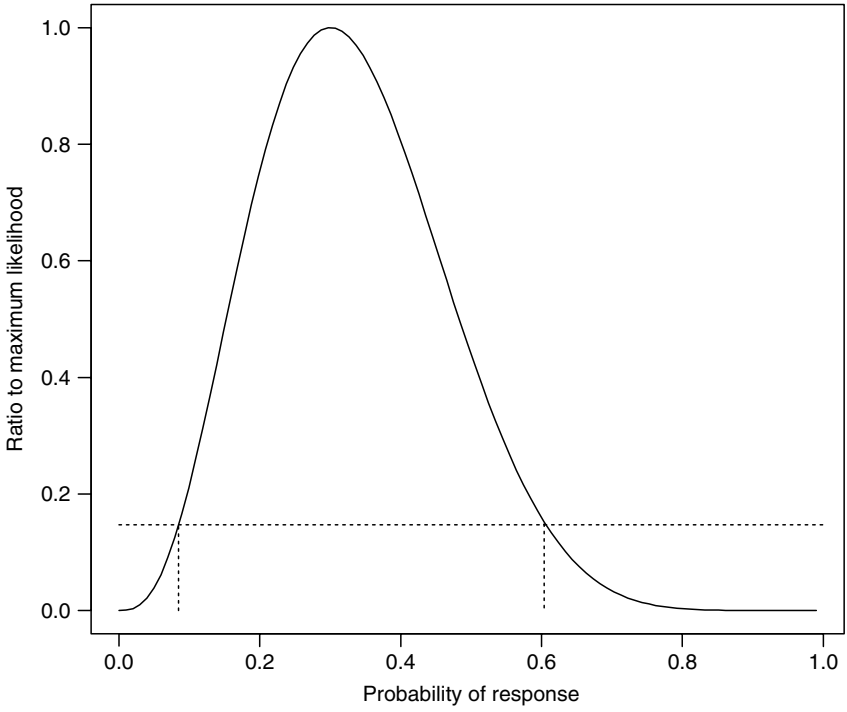

Figure 2.1 Likelihood function for the probability θ of response, after observing 10 individuals of whom 3 responded. The likelihood is scaled relative to its maximum value obtained at the maximum likelihood estimate $\hat{\theta} = 0.3$, and the interval (0.09, 0.61) is based on values with relative likelihood above $\exp(-1.96^2/2) = 0.147$.

2.3 THE NORMAL DISTRIBUTION

The normal (Gaussian) probability distribution is fundamental to much of statistical analysis and features in the majority of the examples covered in this book. We shall make frequent reference to properties of the normal distribution, and therefore it is worth some revision.

We shall use the expression

$$Y \sim \mathrm{N}[\theta, \sigma^2]$$

to represent the assumption that the random quantity Y comes from a normal distribution with mean θ and variance σ^2 (standard deviation σ), which means that

$$p(y) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\frac{(y-\theta)^2}{\sigma^2}\right); \qquad -\infty < y < \infty. \tag{2.20}$$

We also occasionally make use of the notation $p(y) = \mathrm{N}[y|\theta, \sigma^2]$. We note that the inverse of the variance, $1/\sigma^2$, is known as the *precision* of the distribution.

We shall often want to make use of areas under a normal distribution, for example the probability that Y is greater than 0 (a 'tail area'), or the range that comprises, say, 95% of the distribution (a '95% interval'). Let $Z \sim \mathrm{N}[0, 1]$ denote a standard normal variable with mean $\theta = 0$ and standard deviation $\sigma = 1$: the shape of its probability distribution is given in Figure 2.2. Tables or computer programs generally provide the standard normal 'distribution function' $\Phi(z) = P(Z \leqslant z)$, the probability that Z is less than or equal to z, and Table 2.2 displays some useful values for $\Phi(z)$.

We note the useful property

$$\Phi(z) = 1 - \Phi(-z). \tag{2.21}$$

For any tail area ϵ, we denote the corresponding normal deviate by z_ϵ, so that

$$P(Z \leqslant z_\epsilon) = \epsilon \tag{2.22}$$

$$z_\epsilon = \Phi^{-1}(\epsilon), \tag{2.23}$$

where Φ^{-1} represents the inverse of Φ. Hence (2.21) leads to the identity

$$z_\epsilon = -z_{1-\epsilon}.$$

Perhaps the most familiar value is $\Phi^{-1}(0.025) = z_{0.025} = -1.96 = -z_{0.975}$.

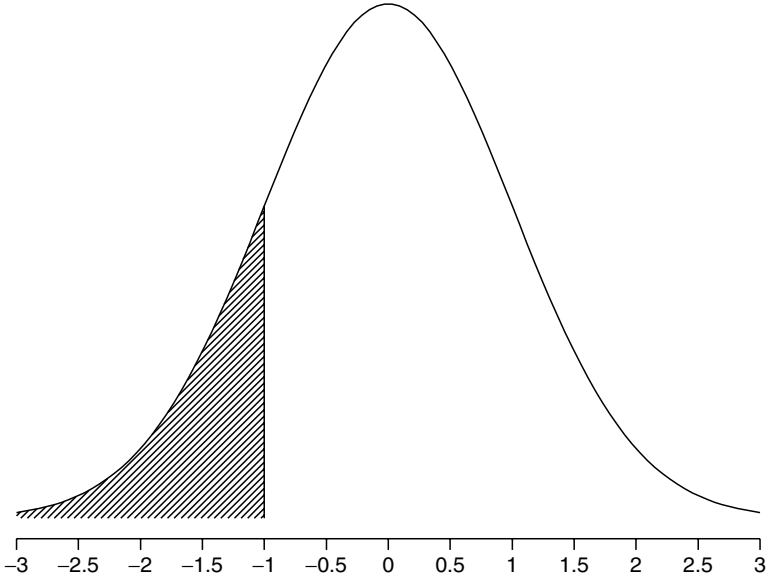

Figure 2.2 Probability distribution of a standard normal variable $Z \sim N[0,1]$. The shaded area represents $\Phi(-1) = P(Z \leqslant -1) = 0.159$.

For a general normal quantity we can easily derive tail areas and intervals from $\Phi(z)$, using the fact that if $Y \sim N[\theta, \sigma^2]$, then $(Y - \theta)/\sigma$ is a standard normal variable $Z \sim N[0, 1]$. Hence

$$P(Y \leqslant y) = P\left(\frac{Y-\theta}{\sigma} \leqslant \frac{y-\theta}{\sigma}\right) = P\left(Z \leqslant \frac{y-\theta}{\sigma}\right) = \Phi\left(\frac{y-\theta}{\sigma}\right). \qquad (2.24)$$

Thus, if we want to know $P(Y \leqslant y)$ we calculate the standardised statistic $z = (y - \theta)/\sigma$ and consult a table such as Table 2.2 to obtain $\Phi(z)$.

Alternatively, if we want, say, a 99% interval for Y, we use a table to find that the 99% interval for Z is $(-2.576, 2.576)$, and then transform this to an interval for Y of $(\theta - 2.576\sigma, \theta + 2.576\sigma)$.

An important property of normally distributed quantities is that they retain normality under addition or subtraction. For example, if Y_1 and Y_2 are independent quantities such that $Y_1 \sim N[\mu_1, \sigma_1^2]$, and $Y_2 \sim N[\mu_2, \sigma_2^2]$, then their sum has distribution

$$Y_1 + Y_2 \sim N[\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2], \qquad (2.25)$$

i.e. their sum is normally distributed with mean equal to the sum of the means, and variance equal to the sum of the variances. We shall find this property very helpful when making predictions (Section 3.13). In many health-care

applications we also frequently consider the difference between two independent quantities; when they are both normally distributed we have

$$Y_1 - Y_2 \sim \mathrm{N}[\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2], \tag{2.26}$$

i.e. their difference is normally distributed with mean equal to the difference of the means, and variance equal to the *sum* of the variances.

2.4 NORMAL LIKELIHOODS

In many contexts it will be reasonable to assume that the data relevant to a parameter θ will be, after m 'observations', summarised by a statistic Y_m with a normal distribution

$$Y_m \sim \mathrm{N}\left[\theta, \frac{\sigma^2}{m}\right], \tag{2.27}$$

where θ is the parameter of interest, generally a treatment effect defined on a suitable scale, and σ^2 is assumed known: note that 'observations' is in quotes as we will find it convenient to use this form even when m is an 'effective' number of observations. After having observed a particular y_m, in traditional statistical terms y_m can be considered as an estimate of the true treatment effect θ, with standard error $\sigma/\sqrt{m}$.

Table 2.2 Some normal tail areas, expressed as percentages, where $100\epsilon = 100\Phi(z_\epsilon) = 100P(\mathrm{Z} \leqslant z_\epsilon)$. From this table we can read, for example, that a symmetric 90% interval for Z would be $(-1.645, 1.645)$, while a one-sided 90% interval could be $(-\infty,\ 1.282)$ or $(-1.282,\ \infty)$.

z_ϵ	$100 \times \Phi(z_\epsilon)$	z_ϵ	$100 \times \Phi(z_\epsilon)$
		0.00	50.0
−0.50	30.8	0.50	69.2
−0.842	20.0	0.842	80.0
−1.00	15.9	1.00	84.1
−1.282	10.0	1.282	90.0
−1.50	6.7	1.50	93.3
−1.645	5.0	1.645	95.0
−1.960	2.5	1.960	97.5
−2.00	2.3	2.00	97.7
−2.326	1.0	2.326	99.0
−2.50	0.6	2.50	99.4
−2.576	0.5	2.576	99.5
−3.00	0.1	3.00	99.9
−3.090	0.1	3.090	99.9

Much of our approximate analysis is based on assuming a normal likelihood (2.27) in quite general contexts. These can be characterised as situations in which it is considered reasonable to quote the results of fitting a statistical model in terms of estimates and standard errors, for example after using standard statistical packages. This can, unfortunately, involve some effort transforming forwards and backwards between the quantities of interest and the somewhat unintuitive scales on which a normal likelihood is more appropriate. However, the examples in this book should demonstrate the value of becoming familiar with this process. It is worth emphasising that, since the likelihood is a function of θ and not a distribution for θ, it is not appropriate to speak, for example, of the mean, variance or tail-area of a likelihood.

We now consider a range of types of data on which the results of different interventions may be compared, detailing the parameters for which it may be appropriate to assume a normal likelihood, and describing how the results of standard regression analyses can be exploited. Obviously there are many areas, particularly with small samples, which cannot be adequately modelled assuming normality. This generally indicates a computational shift away from closed-form analysis and into simulation methodology, which will be discussed in Section 3.19.2.

2.4.1 Normal approximations for binary data

Suppose our data comprise a series of observations in which an event has occurred or not, and we wish to compare the probability of such events under two different interventions. For two events with probabilities p_1 and p_2, the odds ratio (OR) is

$$\text{OR} = \frac{p_1}{1-p_1} \Big/ \frac{p_2}{1-p_2}, \tag{2.28}$$

which is a standard way of reporting changes in the chances of events due to an intervention, on a scale between 0 and ∞. In many circumstances the event is 'negative' (*e.g.* death or disease recurrence) and the 'new' intervention is in the numerator of (2.28), making odds ratios less than 1 favour the new. However, this will not always be the case and care must be taken. We note that for rare events, $(1-p_1)$ and $(1-p_2)$ are near 1, and hence the odds ratio is approximately the relative risk or risk ratio (RR) $= p_1/p_2$, and an odds ratio of, say, 0.7 can also be referred to as a 30% risk reduction. However, we shall try to avoid the term 'relative risk' due to potential confusion.

In order to make the assumption of a normal likelihood more plausible, it is convenient to work with the natural logarithm of the odds ratio so that it takes values on the whole range between $-\infty$ and $+\infty$. Thus

$$\log(\text{OR}) = \theta = \log\left[\frac{p_1}{1-p_1}\right] - \log\left[\frac{p_2}{1-p_2}\right], \tag{2.29}$$

and so the interventions are compared through their difference on the logit scale (Section 2.1.2). This is the standard scale underlying logistic regression analysis. In our analyses we will tend to perform calculations on the log(OR) scale, but report results as odds ratios, which are more intuitive. To assist slightly in the interpretation of log(odds ratios), we note that for small values of $\theta = \log(\text{OR})$, we have the approximation

$$\theta \approx \log(1 + \theta)$$

so that, for example, $\log(\text{OR}) = -0.1$ corresponds roughly to $\text{OR} = 0.9$, or a 10% risk reduction (the exact figure is $\text{OR} = 0.905$). So for small treatment effects, $100 \times \log(\text{OR})$ is approximately the percentage change in risk.

Use of the logit scale has the effect of improving the normal approximation of the likelihood. For example, Figure 2.3 shows the likelihood from Example 2.4 plotted on both the original probability scale and on the log(odds) scale, and the improvement is clear. We now argue why it might be appropriate for likelihood-based intervals to comprise all parameter values with support greater than 14.7% of the maximum, as already quoted in Section 2.2.4 – the following paragraph may be skipped without loss of continuity.

First, note that if the likelihood really *were* $\text{N}[\theta, \sigma^2/m]$, then from (2.20) it has a maximum of $\sqrt{m}/(\sqrt{2\pi}\sigma)$. Hence, relative to its maximum, the likelihood has ordinate $\exp[-(y-\theta)^2/2\sigma^2]$. Second, a 95% interval would comprise values $\theta \pm 1.96\sigma/\sqrt{m}$. Plugging these values into the formula for the normal distribution (2.20) therefore reveals that the boundaries for the 95% interval would have ordinate relative to the maximum of $e^{-1.96^2/2} = 0.147$. Transforming the x-scale of the likelihood does not change the relative ordinates in any way, and hence exactly the same interval is obtained by using this value of 14.7% on the original likelihood on the untransformed scale. Therefore, as long as there is some transformation that can give a reasonable normal approximation, the value of 14.7% of the maximum is justified.

Suppose N observations have been cross-classified by two binary factors, say intervention and response, leading to the following 2×2 table:

		Intervention		
		New	Control	
Event	Death	a	b	$a+b$
	No death	c	d	$c+d$
		$a+c$	$b+d$	N

The maximum likelihood estimate of the odds of death under the new intervention is a/c (the number of deaths divided by the number of survivors), under the control is b/d, and of the odds ratio OR is $(a/c)/(b/d)$. $\theta = \log(\text{OR})$ could be estimated by $\log[(a/c)/(b/d)]$, but in fact the estimator of choice is

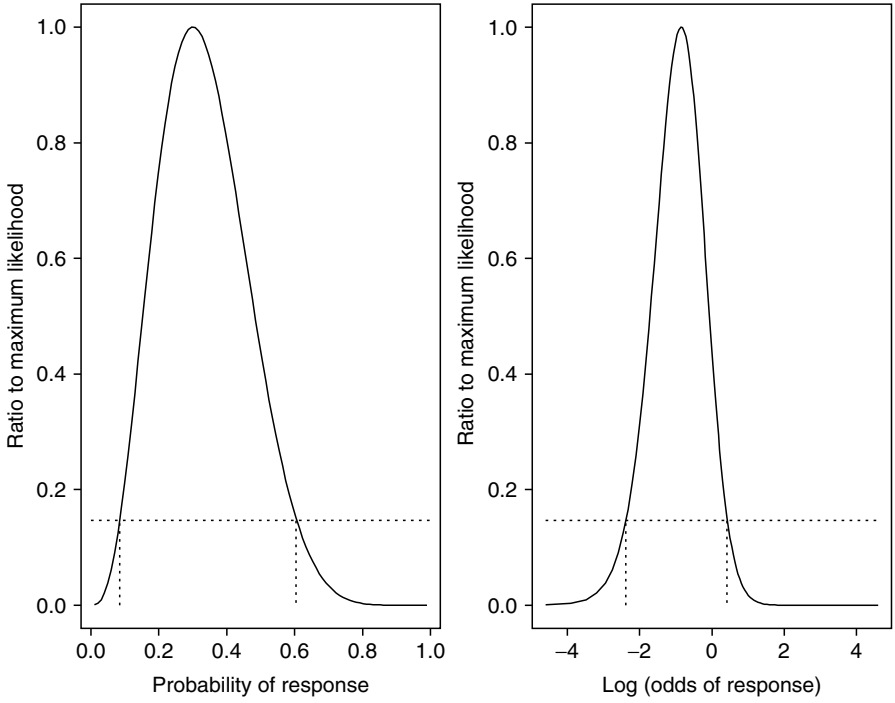

Figure 2.3 Likelihood function for the probability of disease, after treating 10 individuals of whom 3 were successes, plotted on both probability and log(odds) scale. The improvement to the normal approximation is clear.

$$\hat{\theta} = \log\left[\frac{(a+\frac{1}{2})(d+\frac{1}{2})}{(b+\frac{1}{2})(c+\frac{1}{2})}\right], \tag{2.30}$$

where $\hat{\theta}$ represents an estimate of θ. Lower mortality with the new intervention is represented by OR < 1, or negative values of θ. The estimator has approximate variance

$$\mathrm{V}(\hat{\theta}) = \frac{1}{a+\frac{1}{2}} + \frac{1}{b+\frac{1}{2}} + \frac{1}{c+\frac{1}{2}} + \frac{1}{d+\frac{1}{2}}. \tag{2.31}$$

The $\frac{1}{2}$s have the effect of lessening the bias of the estimator and preventing problems with small numbers of events, and will generally have a negligible effect with reasonable sample sizes. Adjustment for confounding factors, using either a Mantel–Haenszel analysis or logistic regression, will also provide an estimate $\hat{\theta}$ with estimated standard error s, and provided N is not too small it will be reasonable to assume a normal likelihood with $\mathrm{V}(\hat{\theta}) = s^2$.

In the notation of (2.27), we need to set $y_m = \hat{\theta}$ and $\sigma^2/m = \mathrm{V}(\hat{\theta})$. Strictly speaking, it is unnecessary to select appropriate values of σ^2 and m since we

could just use $V(\hat{\theta})$ in any analysis, but we shall find that this formulation is useful both for calculation and interpretation. There are two options:

1. We might fix m as the sample size N and so obtain $\sigma^2 = N\ V(\hat{\theta})$.
2. We might fix σ at some specific value, and choose m such that $m = \sigma^2/V(\hat{\theta})$. It turns out that in many contexts $\sigma = 2$ is a suitable choice. For example, consider a balanced randomised trial with a rare event occurring approximately equally often in each arm, so that $a \approx b$ and c and d are very large compared to a and b. Then, from (2.31),

$$V(\hat{\theta}) \approx \frac{2}{a} \approx \frac{4}{m},$$

where $m = a + b$ is the number of events. Thus if we take $\sigma = 2$ and $m = \sigma^2/V(\hat{\theta})$, we should find that m has an approximate interpretation as the number of events underlying the estimate of θ. This is likely to be easier to interpret than a variance on a log(OR) scale, which is fairly incomprehensible. We shall find in Section 2.4.2 that $\sigma = 2$ is also an appropriate choice in survival analysis, in that it also leads to m representing the effective number of events underlying the estimate.

If we are parameterising in terms of differences in proportions rather than the log(odds ratio), it may still be possible to assume a normal likelihood with large sample sizes, where y_m is the difference in sample response rates. Strictly speaking, σ^2 then depends upon the unknown response rates, but an estimate of σ^2 may be used.

Example 2.5 *GREAT: Normal likelihood from a 2 × 2 table*

The GREAT trial of early treatment for myocardial infarction, to be described in greater detail in Example 3.6, gave rise to the following data:

		Treatment		
		New	Control	
Event	Death	13	23	36
	No death	150	125	275
		163	148	311

Using (2.30) gives an estimated log(OR) of $y_m = -0.736$, with estimated variance (2.31) of $0.131 = 0.362^2$. Taking $\sigma = 2$, we obtain $m = 4/0.131 = 30.5$, which is reasonably near the observed number of events (36) and gives an intuitive idea of the amount of evidence underlying the estimate.

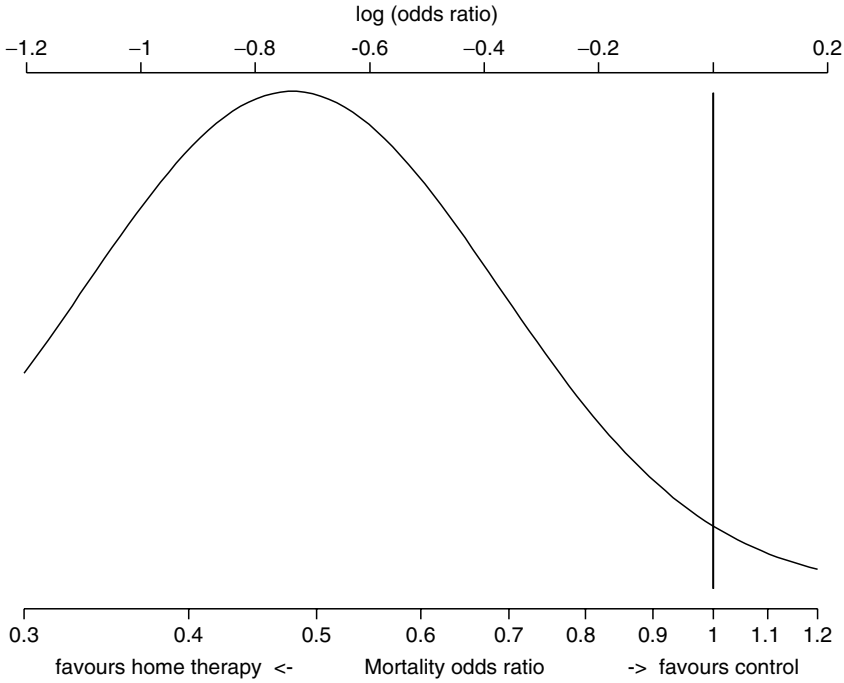

Figure 2.4 Normal likelihood for $\theta = \log(\text{OR})$ in the GREAT trial, with the upper axis labelled on the log(OR) scale. The lower scale is marked in terms of $\text{OR} = e^{\theta}$ for ease of interpretation.

Assuming a normal sampling distribution $y_m \sim \text{N}[\theta, \sigma^2/m]$ leads to the likelihood shown in Figure 2.4, which is plotted on the log(OR) scale but with axes labelled on both OR and log(OR) scales.

2.4.2 Normal likelihoods for survival data

Suppose we have a set of measurements of time to some event, say death or disease recurrence, often referred to as *survival data*. This event is assumed to occur with hazard rate $h(t)$, which is the chance of an event in a short interval of time following t. Survival under two different interventions with hazard rates $h_1(t)$ and $h_2(t)$ may be compared by their hazard ratio, $\text{HR} = h_1(t)/h_2(t)$: the common 'proportional hazards' assumption assumes HR is constant with time. The hazard ratio varies between 0 and ∞, and once again it is convenient to work with its natural logarithm,

$$\log(\text{HR}) = \theta = \log\left[\frac{h_1(t)}{h_2(t)}\right]. \tag{2.32}$$

In our analyses we will tend to perform calculations on the log(HR) scale, but report results as hazard ratios: generally events will be 'negative', such as death or disease recurrence, and so HR < 1 or $\theta < 0$ will favour the treatment in the numerator, which is usually the new intervention.

We note an important connection between hazard ratios and survival probabilities (although this derivation can be skipped). Let T be a random survival time with probability density $p(t)$, and let $S(t) = P(T > t)$ be the chance of surviving beyond t. The hazard rate $h(t)$ is the instantaneous chance of dying, given survival until t, and hence $h(t) = p(t)/S(t)$. Thus the cumulative hazard $H(t)$ obeys

$$H(t) = \int h(t)dt = \int p(t)/S(t)\ dt = -\log S(t).$$

Thus if we assume a proportional hazard model with HR $= h_1(t)/h_2(t)$, then we have

$$\text{HR} = \frac{h_1(t)}{h_2(t)} = \frac{H_1(t)}{H_2(t)} = \frac{\log S_1(t)}{\log S_2(t)}.$$

From this it follows that if p_1 and p_2 are the chances of surviving until some fixed time under the two interventions being compared, then under the proportional hazards assumption

$$\text{HR} = \frac{\log p_1}{\log p_2}, \tag{2.33}$$

$$\log(\text{HR}) = \theta = \log\left[\frac{\log p_1}{\log p_2}\right]. \tag{2.34}$$

This means that if we know the two survival proportions and are willing to assume proportional hazards, then we can transform onto a log(HR) scale. This relationship is shown in Figure 2.5, from which can be read approximate values of log(HR) corresponding to changes in survival probabilities. For example, if a new treatment is thought to change 5-year survival from $p_2 = 20\%$ to $p_1 = 40\%$, then Figure 2.5 suggests this corresponds to a log(hazard ratio) of around -0.5, or HR $= 0.61$. The precise value is given by $\theta = \log[\log(p_1)/\log(p_2)] = -0.56$, corresponding to HR $= 0.57$.

Suppose that the first intervention corresponds to an active treatment T, and the second to a control C. Often the results of a survival analysis may be given in terms of an observed log-rank test statistic L_m, which is defined as the excess of events under T, compared to that expected were there no treatment effect, where m is the total number of events observed. L_m is often denoted as $O - E$ (observed minus expected). Assuming proportional hazards, we have the

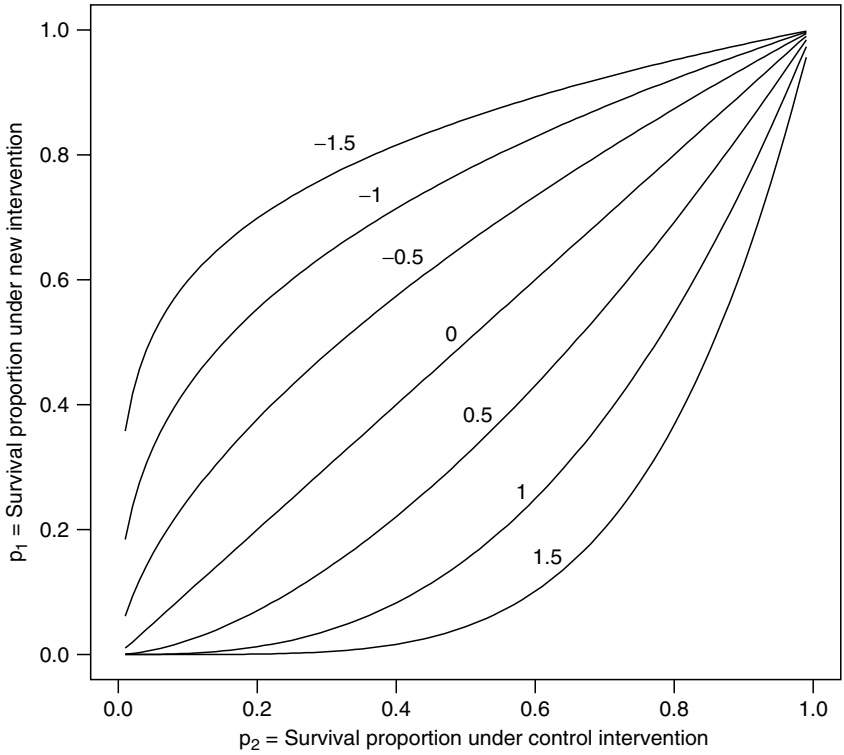

Figure 2.5 Log(hazard ratios) corresponding to changes from survival probability p_2 under a control treatment, to p_1 under a new treatment, where $\log(\text{HR}) = \theta = \log[\log(p_1)/\log(p_2)]$.

following approximation in the particular case of equal allocation and follow-up. If there have been O_T events on treatment, and O_C events on control, then the expected number of events in the treatment group under the null hypothesis is approximately $m/2$, and hence the log-rank statistic is $L_m = O_T - m/2 = (O_T - O_C)/2$. It can be shown (Tsiatis, 1981) that, for large trials, $y_m = 4L_m/m = 2(O_T - O_C)/m$ is an approximate estimate of the log(hazard ratio) θ, and

$$y_m \sim \text{N}[\theta, 4/m].$$

Hence we can set $\sigma = 2$ and adopt a normal likelihood.

If the estimated variance of the log-rank statistic, denoted $\text{V}[O - E]$, is provided in the report of the study, this will take into account different censoring, follow-up and so on. Now

$$\text{V}[O - E] = \text{V}[L_m] = \text{V}[my_m/4] = m^2\text{V}[y_m]/16 \approx m/4,$$

and hence $V[O-E]$ can be equated to $m/4$ in order to obtain the effective number of events m. In more general circumstances we might adjust for covariates using a Cox regression analysis, and hence obtain an estimate $\hat{\theta}$ and its standard error s: if we then set $\sigma = 2$ we may obtain an 'implicit' event count $m = \sigma^2/s^2$, in the same manner as in Section 2.4.1.

2.4.3 Normal likelihoods for count responses

Suppose events occur at a rate λ per unit of population or time. Then our responses will be a count y of the number of events in, say, T units of population or time, which will usually be assumed to have a Poisson distribution with mean λT (Section 2.6.2). For two series of events with rates λ_1 and λ_2, the rate ratio (RaR) λ_1/λ_2 is a standard way of reporting changes in the rates of events due to an intervention. The rate ratio varies between 0 and ∞.

It is again convenient to work with the natural logarithm of a rate ratio, $\theta = \log(\lambda_1/\lambda_2)$, which may be estimated either directly from observed rates or from a Poisson regression.

Suppose we have observed the following data:

	Treatment	
	New	Control
Events	r_1	r_2
Patient-years of follow-up	n_1	n_2

Here n_1 and n_2 are assumed to be large. The maximum likelihood estimate of the rate ratio is $(r_1/n_1)/(r_2/n_2)$, and $\theta = \log(\text{RaR})$ can be estimated by

$$\hat{\theta} = \log \frac{(r_1 + \frac{1}{2})/n_1}{(r_2 + \frac{1}{2})/n_2}. \tag{2.35}$$

RaR < 1, or negative values of θ, indicate a lower event rate with the new treatment. The estimator has approximate variance

$$V(\hat{\theta}) = \frac{1}{r_1 + \frac{1}{2}} + \frac{1}{r_2 + \frac{1}{2}}. \tag{2.36}$$

As with binary and survival data, a normal likelihood can be assumed provided the number of events is not too small, and once again we shall generally set $\sigma = 2$.

2.4.4 Normal likelihoods for continuous responses

Suppose that difference in mean response is the outcome measure of interest, m individuals are allocated to each treatment in a trial, and their individual responses are assumed normal with variance $\sigma^2/2$. Let θ be the true difference in mean response, and y_m be the difference in group sample means. Then $y_m \sim N[\theta, \sigma^2/m]$. (If σ^2 is unknown, then a full Bayesian analysis with a prior on σ^2 is possible: with a specific choice of prior one obtains the standard Student's t distribution for y_m (Section 5.5.1).)

2.5 CLASSICAL INFERENCE

In this section we give the briefest of summaries of standard statistical analysis when normal likelihoods can be assumed: for a comparative discussion of the basis for these and Bayesian techniques, we refer to Chapter 4.

The normal likelihood

$$y_m \sim \mathrm{N}\left[\theta, \frac{\sigma^2}{m}\right]$$

leads to θ being estimated by $\hat{\theta} = y_m$ with an accompanying two-sided 95% confidence interval of $y_m \pm 1.96 \times \sigma/\sqrt{m}$; this may be given the standard sampling-theory interpretation that 95% of the intervals produced using this procedure will contain the true parameter. If we wish to test a null hypothesis, say H_0: $\theta = 0$, we may examine whether the two-sided 95% interval excludes H_0, or equivalently use $z_m = y_m\sqrt{m}/\sigma$ as a standardised test statistic to refer to normal tables and, for example, declare the result 'statistically significant at the two-sided 5% level' if $|z_m| > 1.96$. We may also calculate the 'P-value' P_m associated with z_m, which is the probability of observing data as extreme as z_m under the null hypothesis. This can be taken as

$$P_m = \min\left(P(\mathrm{Z} \geqslant z_m),\ P(\mathrm{Z} \leqslant z_m)\right) = \min\left(\Phi(-z_m),\ \Phi(z_m)\right),$$

although generally the 'two-sided' P-value is considered a more appropriate summary of 'extremeness' for H_0: $\theta = 0$, being

$$2P_m = P(\mathrm{Z} > |z_m|) = \Phi(-|z_m|).$$

Suppose we are designing a clinical trial with proposed size n to detect an alternative hypothesis H_1: $\theta = \theta_A > 0$, and we decide that the result will be declared statistically significant and in favour of H_1 if a two-sided $100(1-2\epsilon)\%$ interval based on a future estimate Y_n lies wholly above 0, corresponding to the future standardised statistic $Z_n > -z_\epsilon$: typically $\epsilon = 0.025$ and so $-z_\epsilon = -z_{0.025} = 1.96$.

In this context this event is equivalent to $P_n \leqslant 2\epsilon$, and 2ϵ is therefore the probability of obtaining a statistically significant conclusion in either direction if the null hypothesis is in fact true. 2ϵ may be termed the 'significance level', the 'size', or the Type I error of the study, and is often denoted α. The null hypothesis will be rejected in favour of H_1 provided $Y_n > -z_\epsilon \sigma/\sqrt{n}$, which from (2.21) and (2.24) will occur with probability

$$1 - \Phi\left(\frac{-z_\epsilon \sigma/\sqrt{n} - \theta}{\sigma/\sqrt{n}}\right) = 1 - \Phi\left(-z_\epsilon - \frac{\theta\sqrt{n}}{\sigma}\right) = \Phi\left(\frac{\theta\sqrt{n}}{\sigma} + z_\epsilon\right).$$

The probability that a trial of n observations will lead to a statistically significant conclusion at the 2ϵ level, given that the alternative hypothesis is true, is known as the *power* of the study, conventionally denoted $1 - \beta$, and hence

$$1 - \beta = \Phi\left(\frac{\theta_A\sqrt{n}}{\sigma} + z_\epsilon\right). \tag{2.37}$$

From (2.37) we can easily see that the sample size necessary to obtain a specified power, say $100(1 - \beta)\%$, will obey

$$\frac{\theta_A\sqrt{n}}{\sigma} + z_\epsilon = \Phi^{-1}(1 - \beta) = z_{1-\beta},$$

and therefore

$$n = (z_{1-\beta} - z_\epsilon)^2 \frac{\sigma^2}{\theta_A^2}. \tag{2.38}$$

Typical values might be $\epsilon = 0.025$, $1 - \beta = 0.80$ and so, from Table 2.2, $(z_{1-\beta} - z_\epsilon)^2 = (0.842 + 1.96)^2 = 7.85$.

Note that some care is required in specifying σ and n. Our formulation is based on assuming that the estimate of the treatment effect has distribution $y_n \sim \mathrm{N}[\theta,\ \sigma^2/n]$. Suppose, however, that we are performing a two-arm study with n patients per group, in which $y_n = \overline{y_2} - \overline{y_1}$, the difference in group means. Then σ^2 must be the variance of the *difference* between the responses from a random pair of patients, one from each arm. This will be the sum of the sampling variances in the two arms.

Example 2.6 *Power: Choosing the sample size for a trial*

Suppose we are designing a trial for a new cancer treatment which it is hoped will raise 5-year survival from 20% to 40%. From the analysis in Section 2.4.2, this is equivalent to a hazard ratio of $\log(0.40)/\log(0.20) = 0.57$ when assuming proportional hazards, or a log(hazard ratio) of $\theta_A = -0.56$. We note the above discussion of power has assumed an

alternative hypothesis $\theta_A > 0$, whereas our θ_A is negative. However, we may simply reverse the role of null and alternative hypotheses and take $\theta_A = 0.56$: this is equivalent to redefining the hazard ratio as control hazard divided by new intervention hazard instead of its inverse. Taking $\sigma = 2$, the power of a study in which n events occur is given by (2.37): assuming $\epsilon = 0.025$ generates the power curve shown in Figure 2.6. From (2.38), 80% power is achieved at $n = 7.85 \times 2^2/(0.56)^2 = 100$: power rises slowly above this size of trial. Under the alternative hypothesis we expect about a 30% overall 5-year mortality in the trial, and so to observe 100 deaths we might recruit about 330 patients, 165 in each arm, and follow them for approximately 5 years.

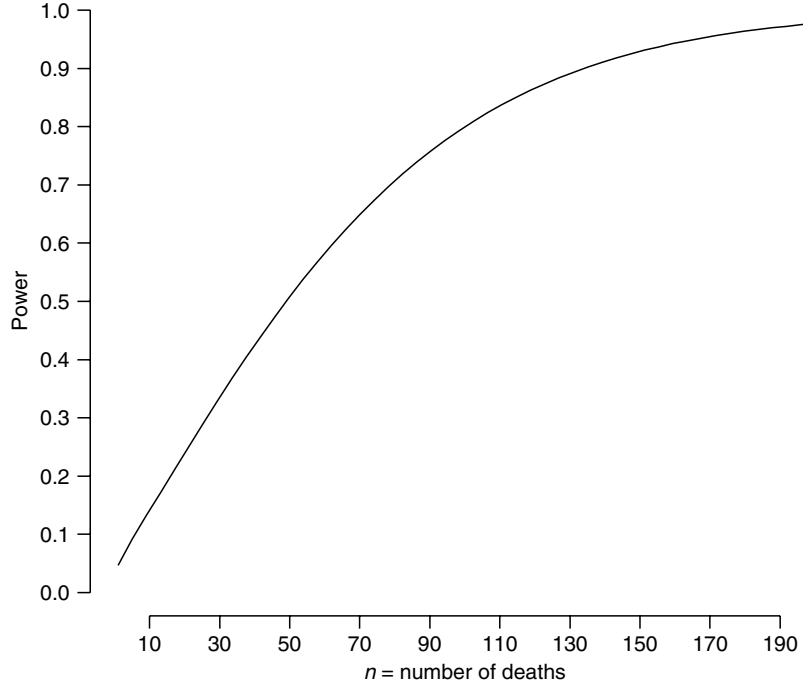

Figure 2.6 Power of a clinical trial in which n events are to be observed, and the alternative hypothesis is a rise from 20% survival to 40% survival, equivalent to a hazard ratio (control/new) of 1/0.57 (log(hazard ratio) $= \theta_A = 0.56$): Power $= \Phi(\theta_A\sqrt{n}/\sigma + z_\epsilon)$. 80% power is achieved at $n = 100$.

In Example 2.6 we took the alternative hypothesis as $\theta > 0$, leading to a power curve that rises for increasing values of θ. However, we shall be using many examples where low values of θ correspond to benefit of the new intervention, and hence care must be taken in using the equations. This rather technical point

is considered in detail in Section 6.5, where we also show how to take into account uncertainty about parameters when conducting power calculations.

2.6 A CATALOGUE OF USEFUL DISTRIBUTIONS*

Bayesian analysis makes use of a wide range of standard, and not so standard, parametric probability distributions in two contexts:

- *Sampling distributions for individual data points or summary statistics* form the basis for likelihoods, just as in classical statistical inference. We shall make use of standard distributional families such as the normal, binomial, and Poisson, but also more unusual choices such as the log-normal for cost data.
- *Prior distributions for parameters* form the very core of Bayesian inference, and the shape of the chosen distribution becomes vital as it represents the relative plausibility for different parameter values. It is therefore important to have a supply of flexible parametric families that can express properties such as skewness and having heavy tails, and so although many of the prior opinions used in this book can be approximated by a normal distribution, we shall also require less standard forms such as the beta, root-inverse-gamma, and half-normal.

These two contexts come together in the use of 'conjugate' distributions, which are families of prior distributions that 'fit together' with particular sampling distributions. These are discussed in Section 3.6.2 and are useful for illustrating Bayesian analysis in simple examples, but modern computational techniques have reduced their importance.

A familiarity with the uses, shapes and properties of different families of distributions can be very valuable, and Bayesian texts contain extensive catalogues of distributions and their mathematical properties: see, for example, Lee (1997), Bernardo and Smith (1994), Gelman *et al.* (1995) and Carlin and Louis (2000). Here we focus on the distributions that will be used in the examples in this book. We shall first discuss their derivation and give formal expressions for their distributional form, expectation and variance, but our primary focus will be on displaying their shapes and discussing their possible use in practical circumstances. We omit explicit restrictions on ranges of parameters when they are clear from the context.

This section might best be used as a reference throughout the book.

2.6.1 Binomial and Bernoulli

A discrete binomial variable Y arises as the sampling distribution of the total number of 'successes' in n independent Bernoulli trials, each with probability θ of success. The likelihood $\theta^y(1-\theta)^{n-y}$ gives the probability for a specific sequence of

$n - y$ 'failures' and y 'successes' (Section 2.2.3), and there are $\binom{n}{y}$ such sequences. Thus $Y \sim \text{Bin}[n,\ \theta]$ represents a binomial distribution with properties:

$$p(y|n,\ \theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}; \qquad y = 0, 1, \ldots, n, \tag{2.39}$$

$$\text{E}(Y|n,\ \theta) = n\theta, \tag{2.40}$$

$$\text{V}(Y|n,\ \theta) = n\theta(1-\theta). \tag{2.41}$$

The binomial with $n = 1$ is simply a Bernoulli distribution, denoted $Y \sim \text{Bern}[\theta]$.

Shape. The examples in Figure 2.7 illustrate the decreasing relative variability and the tendency to a normal distribution that occurs when sample size increases.

Use. The binomial is used as a sampling distribution for empirical counts that occur as proportions. Uses in this book include preference studies (Section 4.4.4), meta-analysis (Section 8.2.2, Example 8.2), and evidence synthesis (Example 8.6).

2.6.2 Poisson

Suppose there are a large number of opportunities for an event to occur, but the chance of any particular event occurring is very low. Then the total number of events occurring may often be represented by a discrete variable Y, where $Y \sim \text{Poisson}[\theta]$ represents a Poisson distribution with properties:

$$p(y|\theta) = \frac{\theta^y e^{-\theta}}{y!}; \quad y = 0,\ 1,\ 2,\ 3,\ \ldots, \tag{2.42}$$

$$\text{E}(Y|\theta) = \theta, \tag{2.43}$$

$$\text{V}(Y|\theta) = \theta. \tag{2.44}$$

In many applications it will arise as a total number of events occurring in a period of time T, where the events occur at an unknown rate λ per unit of time, in which case the expected value of Y is $\theta = \lambda T$.

Shape. The examples in Figure 2.8 show that if events happen with a constant rate, observing for longer periods of time leads to smaller relative variability and a tendency towards a normal shape. Comparison of Figure 2.8 with Figure 2.7 shows that, when sample size increases, a binomial might be approximated by a Poisson with the same mean.

Use. The Poisson distribution is used for count data, as in Example 8.3.

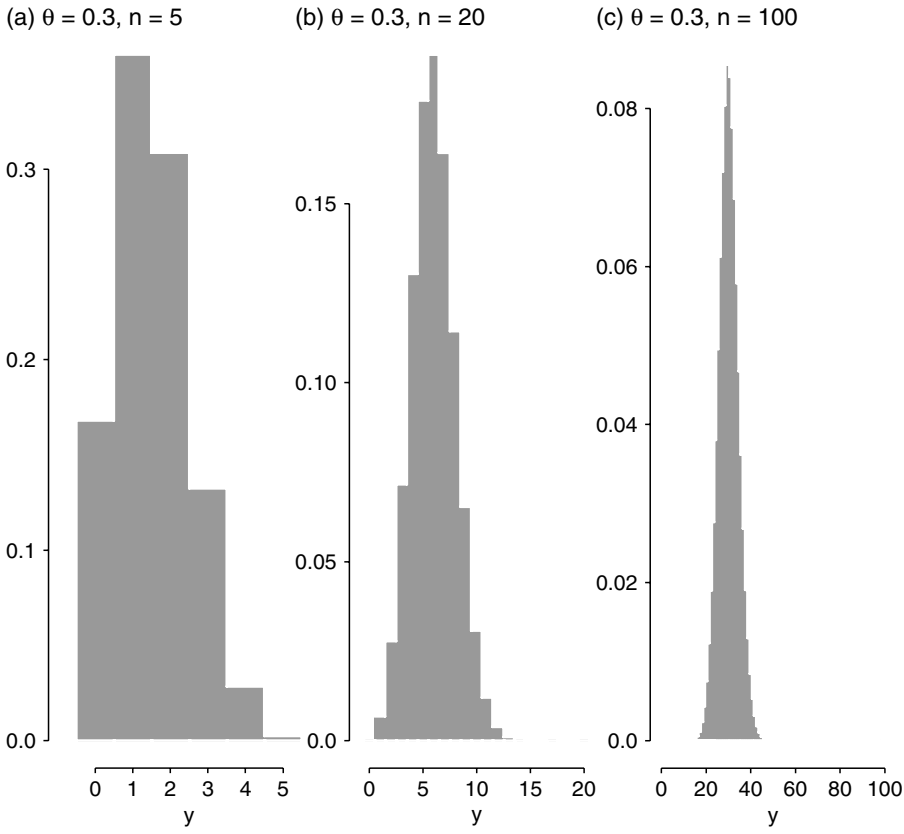

Figure 2.7 Binomial distributions for the number of successes in $n = 5,\ 20,\ 100$ Bernoulli trials, each with probability $\theta = 0.3$ of success.

2.6.3 Beta

Beta distributions form a flexible and mathematically convenient class for quantities constrained to lie between 0 and 1, and so can be used as a prior distribution for unknown proportions. $Y \sim \text{Beta}[a,\ b]$ represents a distribution with properties:

$$p(y|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} y^{a-1}(1-y)^{a-1}; \quad y \in (0,\ 1), \tag{2.45}$$

$$\text{E}(Y|a,b) = \frac{a}{a+b}, \tag{2.46}$$

$$\text{V}(Y|a,b) = \frac{ab}{(a+b)^2(a+b+1)}. \tag{2.47}$$

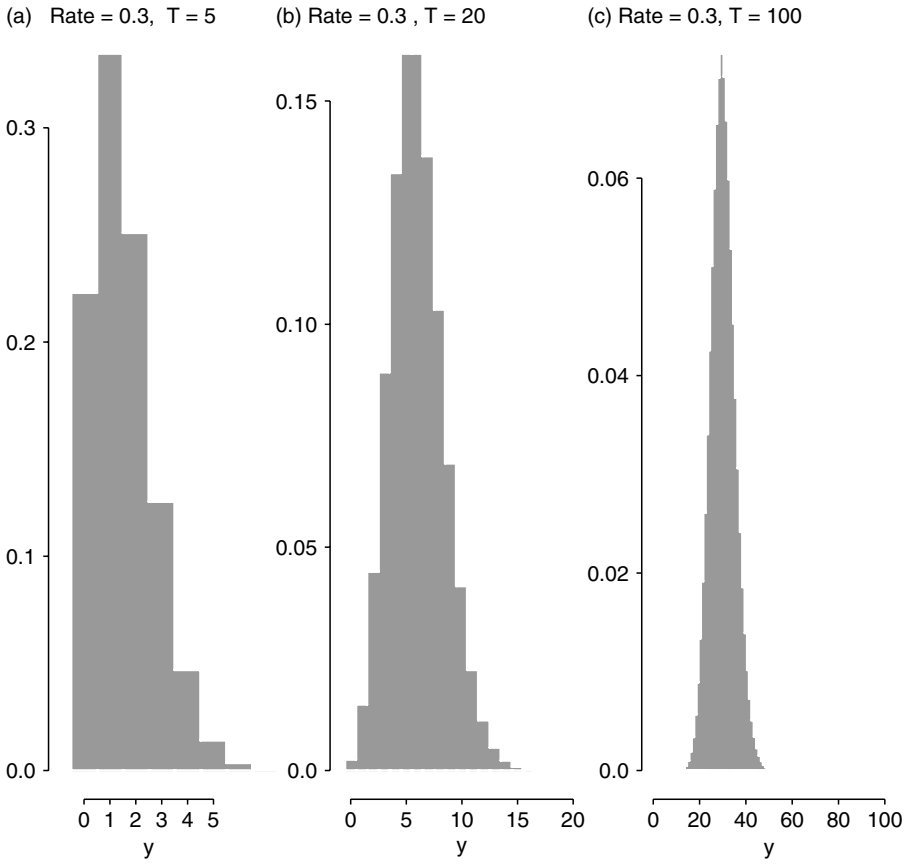

Figure 2.8 Poisson distributions representing the number of events occurring in time $T = 5,\ 20,\ 100$, when the rate at which an event occurs in a unit of time is $r = 0.3$: the Poisson distributions therefore correspond to $\theta = 1.5,\ 6$ and 30.

$\Gamma(a)$ represents the gamma function, a generalisation of the factorial for non-integers, in that $\Gamma(a) = (a-1)!$ if a is an integer. A Beta[1,1] distribution is uniform between 0 and 1 (see Figure 2.9(b) and Section 2.6.4).

Shape. The examples in Figure 2.9 show the flexibility of the family, with a tendency to normal as both parameters become larger.

Use. The sole use of beta distributions is for uncertain proportions where they are 'conjugate' to the binomial family of sampling distributions (Section 3.6) and hence make the necessary computations straightforward. However, we saw in Section 2.4.1 that in most applications with binary data it is much more flexible and convenient to transform the quantity of interest from a proportion (defined on a (0,1) scale) to log(odds) (defined on the full range of $-\infty$ to ∞).

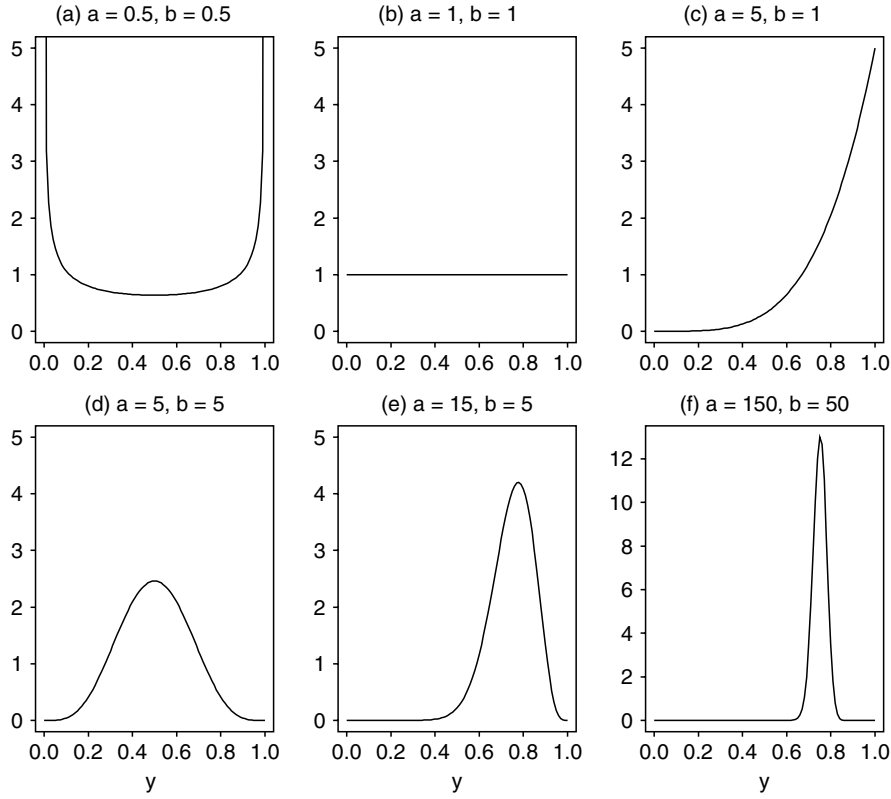

Figure 2.9 Beta distributions for different parameter values showing the flexibility of the family: note change in y-axis for (f).

Therefore, we shall find limited use for the beta except in tutorial examples (see Examples 3.3 and 8.6).

2.6.4 Uniform

Like the beta distribution, a uniform distribution on a range $(a,\ b)$ is generally adopted for an unknown parameter. $Y \sim \text{Unif}[a,\ b]$ means that:

$$p(y|a,b) = \frac{1}{b-a}; \quad y \in (a,b), \tag{2.48}$$

$$\text{E}(Y|a,b) = \frac{a+b}{2}, \tag{2.49}$$

$$\text{V}(Y|a,b) = \frac{(b-a)^2}{12}. \tag{2.50}$$

Shape. The shape of this distribution hardly needs plotting, but an example is given in Figure 2.9(b). Uniform distributions can also be given over a discrete set of values (see Example 3.2).

Use. The only use in this book is as a means of expressing indifference concerning the prior plausibility of a range of values – a so-called 'non-informative' or reference prior (Section 5.5.1). We shall frequently use it in this manner and merely refer to a 'uniform prior', which means uniform over a range that is large enough to encompass all plausible values of θ.

2.6.5 Gamma

Gamma distributions form a flexible and mathematically convenient class for quantities constrained to be positive. $Y \sim \text{Gamma}[a,\ b]$ represents a gamma distribution with properties:

$$p(y|a, b) = \frac{b^a}{\Gamma(a)} y^{a-1} e^{-by}; \qquad y \in (0,\ \infty), \tag{2.51}$$

$$\text{E}(Y|a,b) = \frac{a}{b}, \tag{2.52}$$

$$\text{V}(Y|a,b) = \frac{a}{b^2}. \tag{2.53}$$

Particular cases include the Gamma$[1, b]$ distribution, which is exponential with mean $1/b$, and the Gamma$[\frac{1}{2}v, \frac{1}{2}]$, which is the same as the chi-squared distribution χ^2_v on v degrees of freedom. A useful piece of distribution theory is that if $Y_1, \ldots, Y_n$ are a set of i.i.d. N$[\theta, \sigma^2]$ variables with mean $\overline{Y}$ and sample variance $S^2 = \sum_i (Y_i - \overline{Y})^2/n$, then $\sum_i (Y_i - \theta)^2/\sigma^2 \sim \chi^2_n$, and $nS^2/\sigma^2 \sim \chi^2_{n-1}$. We shall use this in Example 8.4.

Shape. The examples in Figure 2.10 show the family to be reasonably flexible.

Use. One justification is that the gamma distribution 'conjugate' to the Poisson family (Section 3.6.2). However, as with binary data, we shall see in Section 2.4.3 that in most applications it is much more flexible and convenient to transform the quantity of interest from a rate (defined on a $(0, \infty)$ scale) to a log-rate (defined on the full range of $-\infty$ to ∞), and then use normal approximations.

An alternative popular use has been as a prior distribution for the precision parameter (1/variance) of a normal distribution, for which it is also conjugate (Section 3.6.2). This is equivalent to using a root-inverse-gamma distribution for the standard deviation (see Section 2.6.6).

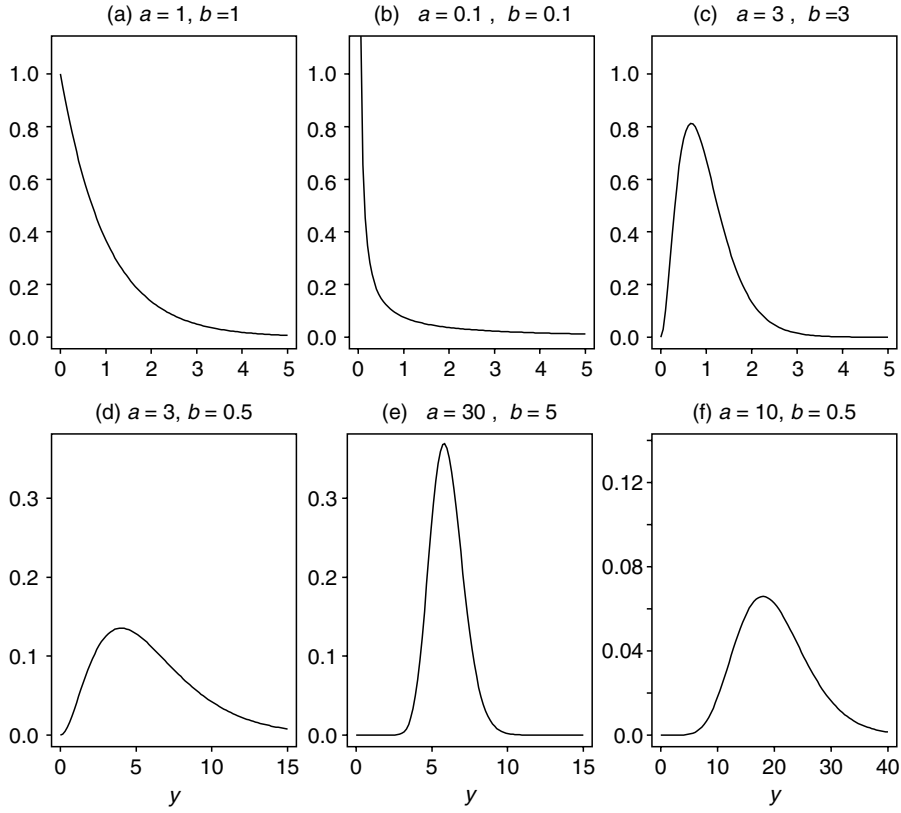

Figure 2.10 Gamma distributions. (a) is exponential with mean 1, (a), (b) and (c) all have the same mean but different shapes, (d) is a χ^2_6 distribution with mean 6, while (e) has the same mean as (a) but a different shape and becomes increasingly close to normal as the parameters both increase. (f) is a χ^2_{20} distribution.

2.6.6 Root-inverse-gamma

If $X \sim \text{Gamma}[a,b]$, then $1/\sqrt{X} \sim \text{RIG}[a,b]$. $Y \sim \text{RIG}[a,b]$ represents a root-inverse-gamma distribution with properties (Bernardo and Smith, 1994, p. 431):

$$p(y|a,b) = \frac{2b^a}{\Gamma(a)} \frac{1}{y^{2a+1}} e^{-b/y^2}; \qquad y \in (0, \infty), \tag{2.54}$$

$$\text{E}(Y|a,b) = \frac{\sqrt{b}\ \Gamma(a - \frac{1}{2})}{\Gamma(a)}, \tag{2.55}$$

$$\text{V}(Y|a,b) = \frac{b}{a-1} - \text{E}^2(Y|a,b). \tag{2.56}$$

We note that the variance is only defined for $a > 1$.

Shape. The examples in Figure 2.11 show that the family can have the somewhat curious property of forcing the quantity away from 0.

Use. The RIG is the implied prior distribution for a standard deviation when a gamma distribution is used for a precision, and so is frequently implicitly adopted in Bayesian analysis. However, it is almost never plotted, and the shape is perhaps not what was intended in many applications, given its property of rejecting low values. We shall therefore adopt it with some caution in Section 5.7.3 and in Example 8.1.

2.6.7 Half-normal

The half-normal arises by folding a normal distribution around 0: formally, if $X \sim \mathrm{N}[0, \sigma^2]$, then $|X| \sim \mathrm{HN}[\sigma^2]$. Thus $Y \sim \mathrm{HN}[\sigma^2]$ represents a half-normal distribution with properties:

$$p(y|\sigma^2) = \sqrt{\frac{2}{\pi\sigma^2}}\, e^{\frac{-y^2}{2\sigma^2}}; \quad y \in (0, \infty), \tag{2.57}$$

$$\mathrm{E}(Y|\sigma^2) = \sqrt{\frac{2}{\pi}}\,\sigma, \tag{2.58}$$

$$\mathrm{V}(Y|\sigma^2) = \sigma^2\left(1 - \frac{2}{\pi}\right), \tag{2.59}$$

and a median of $\Phi^{-1}(0.75)\ \sigma = z_{0.75}\ \sigma = 0.773\,\sigma$, using the notation of Section 2.3.

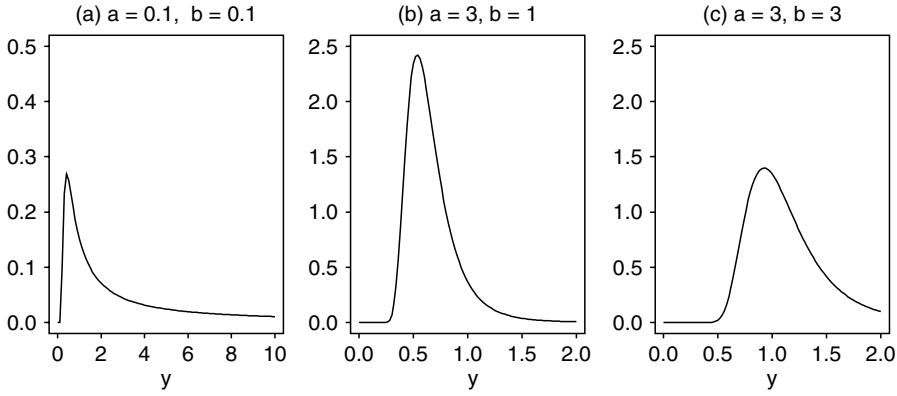

Figure 2.11 Root-inverse-gamma distributions. Note the different scale for (a), which has a very long right-hand tail. Comparing (c) with (b) shows that increasing b retains the shape but multiplies the mean and standard deviation by b.

Shape. The examples in Figure 2.12 show the family to express maximum support for 0, with the rate of decline governed by σ.

Use. The half-normal is useful to express support for values near 0, with σ controlling the upper range of support. This is applied to standard deviations in Section 5.7.3, and illustrated in Examples 8.1 and 8.5.

2.6.8 Log-normal

The log-normal is a distribution on positive values, like the gamma, root-inverse-gamma, and half-normal. It is defined as the *exponential* of a normal variable (this can cause confusion). Thus if $Y \sim \text{LN}[\mu, \sigma^2]$, then $\log(Y) \sim \text{N}[\mu, \sigma^2]$. $Y \sim \text{LN}[\mu, \sigma^2]$ represents a log-normal distribution with properties:

$$p(y|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma y} e^{-(\log y-\mu)^2/2\sigma^2}; \qquad y \in (0,\infty), \tag{2.60}$$

$$\text{E}(Y|\mu,\sigma^2) = e^{\mu+\sigma^2/2}, \tag{2.61}$$

$$\text{V}(Y|\mu,\sigma^2) = e^{2\mu+\sigma^2}(e^{\sigma^2} - 1). \tag{2.62}$$

Shape. The examples in Figure 2.13 show that a range of skewed distributions can be represented, although the right-hand tail is remarkably long. For example, Figure 2.13(b) has a broadly similar shape to the Gamma[0.1, 0.1] shown in Figure 2.11(a): however, while the latter has mean 1 and standard deviation $\sqrt{10} = 3.2$, the LN[0, 3] has mean $e^{4.5} = 90$, and standard deviation

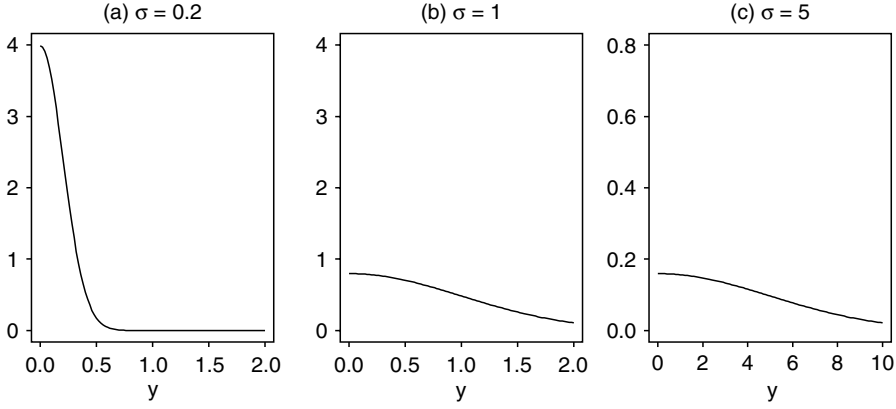

Figure 2.12 Half-normal distributions, with maximum at 0 and declining support for increasing y.

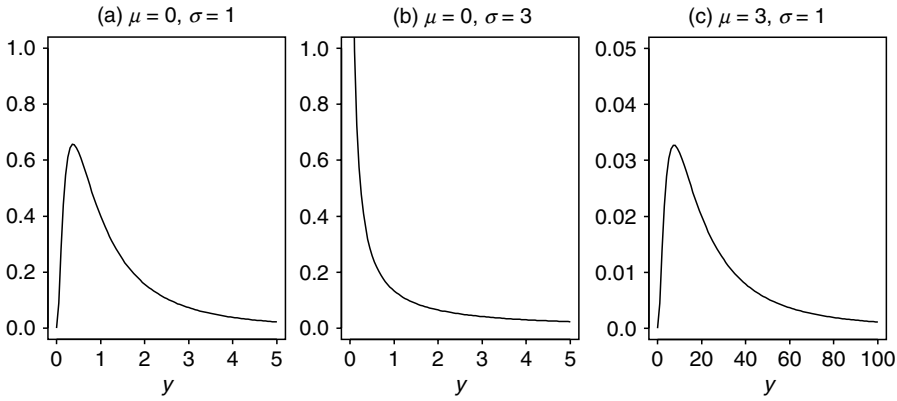

Figure 2.13 Log-normal distributions. Comparing (c) with (b) shows that μ acts as a scale parameter and does not change the shape of the distribution.

$\sqrt{e^9(e^9 - 1)} = 8100$. Thus although the gamma and log-normal are sometimes considered as alternative options for skewed distributions, the much heavier tail of the log-normal should be kept in mind.

Use. The log-normal can be used as a sampling distribution for positive observations such as costs (Example 9.2), or as a prior distribution for positive parameters such as variances (Examples 6.10 and 9.2). We have seen in Section 2.4 that in many situations we carry out inferences on logarithms of quantities, and then transform results back to a more interpretable scale. Thus in our examples that use normal theory, our posterior distributions of odds ratios, hazard ratios and rate ratios are in fact log-normal distributions.

2.6.9 Student's *t*

A standardised Student's t distribution arises as the ratio of a standard normal variable to the square root of an independent χ^2 variable divided by its degrees of freedom, and has a prominent role in classical statistics as the sampling distribution of a sample mean divided by its estimated standard error. It also occurs as a posterior distribution for the mean of a normal distribution given a specific choice of prior for the unknown variance (DeGroot, 1970). $Y \sim t[\mu, \sigma^2, \nu]$ represents a Student's t distribution with ν degrees of freedom, which has properties:

$$p(y|\mu,\sigma^2,\nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}\sigma} \frac{1}{\left(1 + \frac{(y-\mu)^2}{\nu\sigma^2}\right)^{\frac{\nu+1}{2}}}; \qquad y \in (\infty, \infty), \tag{2.63}$$

$$\mathrm{E}(Y|\mu,\sigma^2,\nu) = \mu, \tag{2.64}$$

$$\mathrm{V}(Y|\mu,\sigma^2,\nu) = \sigma^2 \frac{\nu}{\nu-2}; \tag{2.65}$$

the mean only exists if $\nu > 1$, and the variance only exists if $\nu > 2$.

Shape. Figure 2.14 shows the heavy-tailed nature of the t distribution, with high degrees of freedom looking increasingly normal.

Use. Apart from arising as a posterior distribution, it can also be used as a sampling distribution when some outliers are expected.

2.6.10 Bivariate normal

X and Y are said to have a bivariate normal distribution, denoted $X,Y \sim \mathrm{BN}[\theta_X,\theta_Y,\sigma_X,\sigma_Y,\rho]$, if

$$p(x,y|\theta_X,\theta_Y,\sigma_X,\sigma_Y,\rho) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left(-\frac{Q}{2(1-\rho^2)}\right); \quad x, y \in (\infty,\infty), \tag{2.66}$$

where Q is the quadratic expression

$$Q = \frac{(x-\theta_X)^2}{\sigma_X^2} - \frac{2\rho(x-\theta_X)(y-\theta_Y)}{\sigma_X\sigma_Y} + \frac{(y-\theta_Y)^2}{\sigma_Y^2}.$$

The distribution has properties

$$\mathrm{E}(X) = \theta_X, \quad \mathrm{E}(Y) = \theta_Y, \quad \mathrm{V}(X) = \sigma_X^2, \quad \mathrm{V}(Y) = \sigma_Y^2,$$

and covariance and correlation

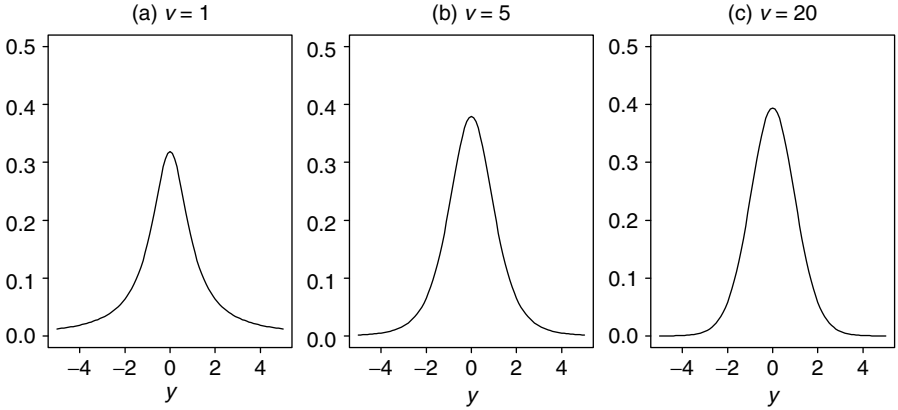

Figure 2.14 Student's t distributions with $\mu = 0$, $\sigma = 1$: other values of μ and σ will change the location and scale but not the shape.

$$\mathrm{Cov}(X,Y) = \rho\sigma_X\sigma_Y, \quad \mathrm{Corr}(X,Y) = \rho.$$

In addition, the conditional distribution of $Y|x$ is normal with mean and variance

$$\begin{aligned} \mathrm{E}(Y|x) &= \theta_Y + \frac{\rho\,\sigma_Y}{\sigma_X}(x - \theta_X), \\ \mathrm{V}(Y|x) &= \sigma_Y^2(1 - \rho^2). \end{aligned} \tag{2.67}$$

The conditional variance $\sigma_Y^2(1 - \rho^2)$ is never more than the unconditional variance σ_Y^2, showing that knowing the value of X never increases our uncertainty about Y. In addition, the conditional mean is a linear function of x – this is known as the 'regression' of Y on X. The bivariate normal generalises naturally to higher dimensions but we shall not require this extension for this book.

Shape. Figure 2.15 shows a 'contour plot' of a bivariate normal distribution, where contours are ellipses obtained as solutions of $Q =$ constant.

Use. The bivariate normal can be used as a sampling distribution of two correlated quantities, such as in Example 9.1 where it is used to describe the joint

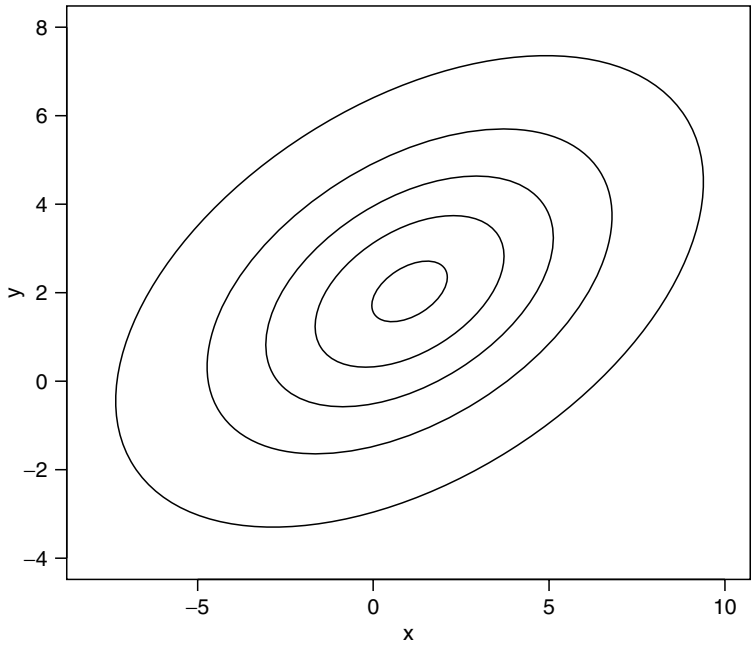

Figure 2.15 A bivariate normal distribution with parameters $\theta_X = 1$, $\theta_Y = 2$, $\sigma_X = 3$, $\sigma_Y = 2$, $\rho = 0.5$, with expanding ellipses enclosing 5%, 25%, 50%, 75% and 95% of the probability distribution.

distribution of costs and benefits. It also arises naturally as a prior distribution for two possibly correlated unknown parameters, such as the baseline rate and treatment effect in a clinical trial or epidemiological study (Section 8.2.3): see Example 8.3 for an example in a meta-analysis of observational studies.

2.7 KEY POINTS

1. Bayesian analysis rests wholly on probability theory, and all inferences can be derived from three basic rules.
2. The sampling distributions for data are used to derive likelihoods for unknown parameters, and so familiarity with classical methods helps in Bayesian analysis.
3. Normal approximations for likelihoods play a very important role.
4. Bayesian analysis makes use of a wide range of parametric probability distributions, both as a basis for likelihoods and as prior distributions.

EXERCISES

2.1. A coin is tossed and lands 'heads'.
   (a) What is *your* assessment of the probability that a second toss of the coin will also yield a 'head'?

   Before the coin was tossed for the first time it was randomly selected from two possible coins, one a 'fair' coin, *i.e.* with with both 'head' and 'tail', and the other a 'double-headed' coin.
   (b) What is your assessment of the probability that the second toss of the coin will now yield a 'head'?

2.2. Consider a case of disputed paternity, and the blood groups of the mother, the child and the alleged father. The mother has blood type O and the alleged father has blood type AB: let F denote the event that he is the true father. If the child has blood group O then the alleged father can be excluded from the paternity case. After testing, the child has blood type B, and Mendelian genetics implies $P(B|F) = 0.5$. The blood bank gives $P(B|\bar{F}) = 0.09$ for Caucasians. What is $P(F|B)$, *i.e.* the probability that the alleged father really is the father given that the child has blood type B, (a) as a general function of $P(F)$, and (b) when $P(F) = 0.5$?

2.3. Lee (1997) considers the case of twins and whether they are monozygotic (M) or dizygotic (D). Monozygotic twins develop from the same egg, look very similar (often being referred to as identical twins) and are *always* of the same sex, whilst dizygotic twins can look very similar too, but can be of different sexes. Therefore, $P(GG|M) = P(BB|M) = 0.5$, $P(GG|D) = P(BB|D) = 0.25$, and $P(GB|M) = 0$, $P(GB|D) = 0.5$.

(a) By extending the argument, express $P(GG)$ in terms of $p(M)$, the prior probability that a set of twins is monozygotic.
(b) Again in terms of $p(M)$, find the probability that if twins are both girls they are dizygotic, *i.e.* $P(D|GG)$.
(c) Find $P(D|GG)$ when $p(M) = 0.5$.

2.4. In a study of a drug, 20 out of 50 patients respond. (a) Find the maximum likelihood estimate for the response rate, and use a normal approximation for the likelihood for the log(odds) to find a 95% interval of values for the response rate which are supported by the data. A second study is performed, but due to time constraints only 20 patients are observed, of whom 8 respond. (b) For the second study, what is the most likely value for the response rate and an approximate 95% interval?

2.5. Gardner *et al.* (2000) report the results of a trial to investigate whether a progesterone emitting intra-uterine device (IUD) can reverse endometrial changes in women being treated for breast cancer with tamoxifen. At the end of the trial 5 out 56 women in the IUD group were discovered to have a submucous fibroid, whilst the corresponding number in the control group was 13 out of 53. Obtain a normal approximation to the likelihood for the log(odds ratio), and hence give a 95% interval for the odds ratio.

2.6. In the breast cancer trial of Exercise 2.5, women recruited had received tamoxifen for varying lengths of time, and the investigators felt that it was important to adjust for this and other possible confounders (including parity, menopausal status, body-mass index and age) in any analysis. They therefore used logistic regression to obtain an adjusted odds ratio of 0.23 with associated 95% confidence interval (CI) from 0.07 to 0.76. Obtain a normal approximation to the likelihood for the adjusted log(odds ratio).

2.7. Allen-Mersh *et al.* (1994) reported the results of a trial in which patients undergoing chemotherapy for liver metastases were randomised to receive it either systematically, as was standard, or via hepatic arterial infusion (HAI). Of 51 randomised to HAI 44 died, and of 49 randomised to systemic therapy 46 died.
(a) Obtain a rough normal approximation to the likelihood for the log(hazard ratio).
(b) The reported hazard ratio was 0.60 (95% CI from 0.40 to 0.95). Why might the approximation be so poor?

2.8. Shepherd *et al.* (2002) report the results of the PROSPER placebo-controlled RCT to evaluate the use of pravastin in elderly patients on a combined primary endpoint of death from coronary heart disease, non-fatal myocardial infarction, or stroke (fatal or non-fatal). Of 2891 patients randomised to pravastatin, 408 experienced the primary endpoint, whilst in the placebo group of 2913 patients 473 experienced it. (a) Obtain a rough estimate of the log(hazard ratio), assuming equal follow-up. The authors reported the results of a Cox proportional hazards regression

model adjusting for a large number of baseline characteristics, which resulted in a 15% proportionate reduction in the hazard of the primary endpoint with 95% CI from 3% to 26%. (b) Obtain a normal approximation to the likelihood for the adjusted log(hazard ratio).

2.9. The PROSPER RCT in Exercise 2.8 also considered whether cancer incidence was higher in those patients receiving statin therapy. In the statin arm 245 cancers occurred out of 2891 patients, and in the placebo arm 199 cancers occurred in 2913 patients.
   (a) Obtain a normal approximation to the likelihood for the log(odds ratio).
   (b) Calculate a classical two-sided *P*-value.
   (c) Assess whether the data support a change in cancer incidence with statin use.

2.10. Suppose that 10% of patients taking anti-retroviral therapy currently experience a particular adverse event. Preliminary evidence suggests a new therapy might reduce this rate to 5%.
   (a) What is the hypothesised log(odds ratio)?
   (b) Estimate the number of events that would be required in an RCT in order to detect such a change, assuming a two-sided 5% level of statistical significance is to be used with a required power of 80%.
   (c) How many patients would be required in each arm of an RCT in order to observe this many events?