3
Likelihood

The purpose of models is to allow us to use past observations (*data*) to make predictions. In order to do this, however, we need a way of choosing a value of the parameter (or parameters) of the model. This process is called parameter *estimation* and this chapter discusses the most important general approach to it. In simple statistical analyses, these stages of model building and estimation may seem to be absent, the analysis just being an intuitively sensible way of summarizing the data. However, the analysis is only scientifically useful if we can generalize the findings, and such generalization must imply a model. Although the formal machinery of modelling and estimation may seem heavy handed for simple analyses, an understanding of it is essential to the development of methods for more difficult problems.

In modern statistics the concept which is central to the process of parameter estimation is *likelihood*. Likelihood is a measure of the *support* provided by a body of data for a particular value of the parameter of a probability model. It is calculated by working out how probable our observations would be if the parameter were to have the assumed value. The main idea is simply that parameter values which make the data more probable are better supported than values which make the data less probable. In this chapter we develop this idea within the framework of the binary model.

3.1 Likelihood in the binary model

Fig. 3.1 illustrates the outcomes observed in a small study in which 10 subjects are followed up for a fixed time period. There are two possible outcomes for each subject: *failure*, such as the development of the disease of interest, or *survival*. We adopt a binary probability model for the outcome for each subject in which failure has probability π and survival has probability $1-\pi$. The complete tree would have many branches but only those corresponding to the observed study result is shown in full. To calculate the probability of occurrence of this result we simply multiply probabilities along the branches of the tree in the usual way:

$$\pi \times \pi \times (1-\pi) \times \cdots \times (1-\pi) = (\pi)^4(1-\pi)^6.$$

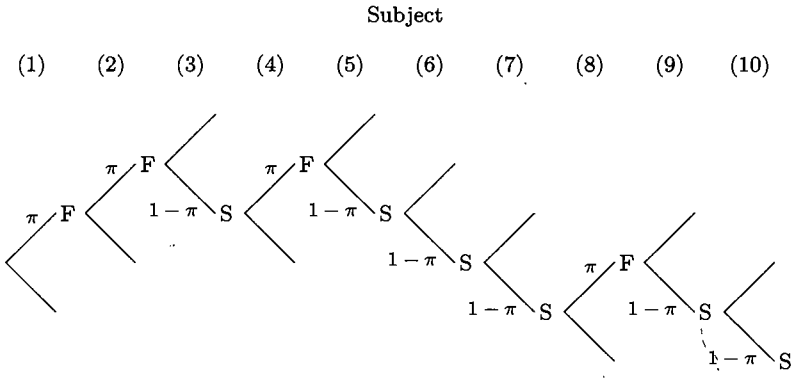

Fig. 3.1. Study outcomes for 10 subjects.

This expression can be used to calculate the probability of the observed study result for any specified value of π. For example, when $\pi = 0.1$ the probability is

$$(0.1)^4 \times (0.9)^6 = 5.31 \times 10^{-5}$$

and when $\pi = 0.5$ it is

$$(0.5)^4 \times (0.5)^6 = 9.77 \times 10^{-4}.$$

The results of these calculations show that the probability of the observed data is greater for $\pi = 0.5$ than for $\pi = 0.1$. In statistics this is often expressed by saying that $\pi = 0.5$ is more *likely* than $\pi = 0.1$, meaning that the former value is better supported by the data. In everyday use the words probable and likely mean the same thing, but in statistics the word likely is used in this more specialized sense.

Exercise 3.1. Is $\pi = 0.4$ more likely than $\pi = 0.5$?

The result of the expression

$$(\pi)^4(1-\pi)^6,$$

is a probability, but when we use it to assess the amount of support for different values of π it is called a *likelihood*. More generally, if we observed D failures in N subjects, the likelihood for π would be

$$(\pi)^D(1-\pi)^{N-D},$$

and we shall call this expression the *Bernoulli* likelihood, after the Swiss mathematician. Because there are so many possible outcomes to the study,

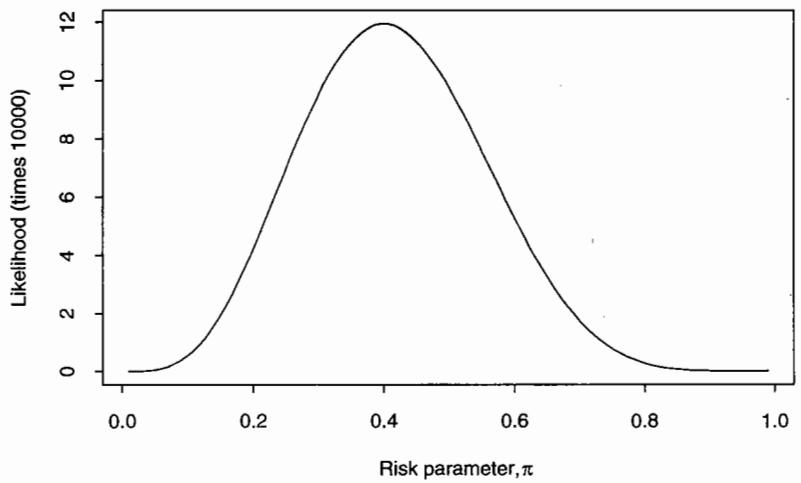

Fig. 3.2. The likelihood for π.

the likelihood (which is the probability of just one of these) is a small number. However, it is not the *absolute* value of the likelihood which should concern us, but its *relative* value for different choices of π.

Returning to our numerical example, Fig. 3.2 shows how the likelihood varies as a function of π. The value $\pi = 0.4$ gives a likelihood of 11.9×10^{-4}, which is the largest which can be achieved. This value of π is called the *most likely value* or, more formally, the *maximum likelihood estimate* of π. It coincides with the observed proportion of failures in the study, 4/10.

3.2 The supported range for π

The most likely value for π is 0.4, with likelihood 11.9×10^{-4}. The likelihood for any other value of π will be less than this. How much less is measured by the *likelihood ratio*, which takes the value 1 when $\pi = 0.4$ and values less than 1 for any other values of π. This provides a more convenient measure of the degree of support than the likelihood itself. It can be used to classify values of π as either supported or not according to some critical value of the likelihood ratio. Values of π with likelihood ratios above the critical value are reported as 'supported', and values with likelihood ratios below this critical value as 'not supported'. The *supported range* for π is the set of values of π with likelihood ratios above the critical value. The choice of the critical value is a matter of convention.

For our observation of 4 failures and 6 survivors, the likelihood ratio as a function of π is shown in Figure 3.3. We have used the number 0.258

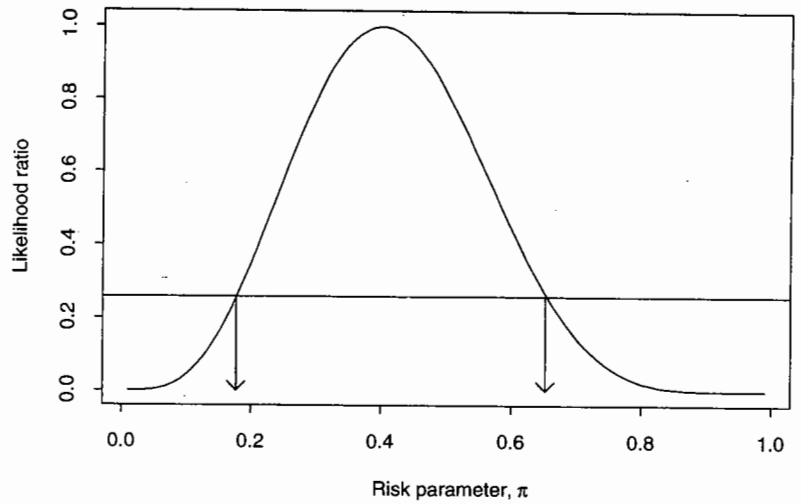

Fig. 3.3. The likelihood ratio for π.

for the critical value of the likelihood ratio and indicated the limits of the supported range with the two arrows. The range of supported values for π is rather wide in this case: from 0.17 to 0.65.* For any choice of critical value the width of the supported range reflects the uncertainty in our knowledge about π. The main thing which determines this is the quantity of data used in calculating the likelihood. For example, if we were to observe 20 failures in 50 subjects, the most likely value of π would still be 0.4, but the supported range would be narrower (see Figure 3.4).

Although the concept of a supported range based on likelihood ratios is intuitively simple, it requires some consensus about the choice of critical value. The achievement of this has not proved easy, since many scientists lack an intuitive feel for the amount of uncertainty corresponding to a stated numerical value for the likelihood ratio. As a result, statistical theorists have tried to find ways to measure the uncertainty about the value of a parameter in terms of *probability* which, it is argued, is more easily interpreted. The way of doing this which is most widely accepted in the scientific community is by imagining a large number of repetitions of the study. This approach is known as the *frequentist* theory of statistics and leads to a *confidence interval* for π rather than a supported range. Another approach, often favoured by mathematicians, is based on a probability measure for the subjective 'degree of belief' that the parameter value lies in a stated *credible*

*These values were obtained from the graph, as illustrated. We shall be describing more convenient approximate methods for their computation in Chapter 9.

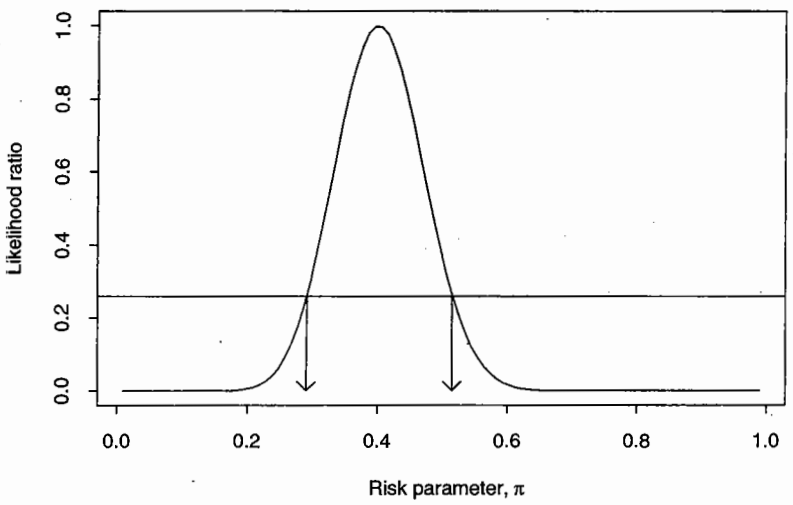

Fig. 3.4. The likelihood ratio based on 20 failures in 50 subjects.

interval. This is the *Bayesian* theory of statistics.

Luckily for applied scientists, these philosophical differences can be resolved, at least for the analysis of moderately large studies. In this case, we will show in Chapter 10 that the supported range based on a likelihood ratio criterion of 0.258 coincides approximately with a 90% confidence interval in the frequentist theory of statistics and a 90% credible interval in the Bayesian theory. We shall, therefore, set aside these difficulties for the present and continue to develop the idea of likelihood, which holds a central place in both theories of statistics and from which most of the statistical methods of modern epidemiology can be derived.

3.3 The log likelihood

The likelihood, when evaluated for a particular value of the parameter, can turn out to be a very small number, and it is generally more convenient to use the (natural) logarithm of the likelihood in place of the likelihood itself.† When combining log likelihoods from independent sets of data the separate log likelihoods are added to form the combined likelihood. This is because the likelihoods themselves, being the probabilities of independent sets of data, are combined by multiplication. The log likelihood for π, in this example, is

$$4\log(\pi) + 6\log(1-\pi).$$

†Readers not completely familiar with the logarithmic function, $\log(x)$ and its inverse, the exponential function, $\exp(x)$, are referred to Appendix A.

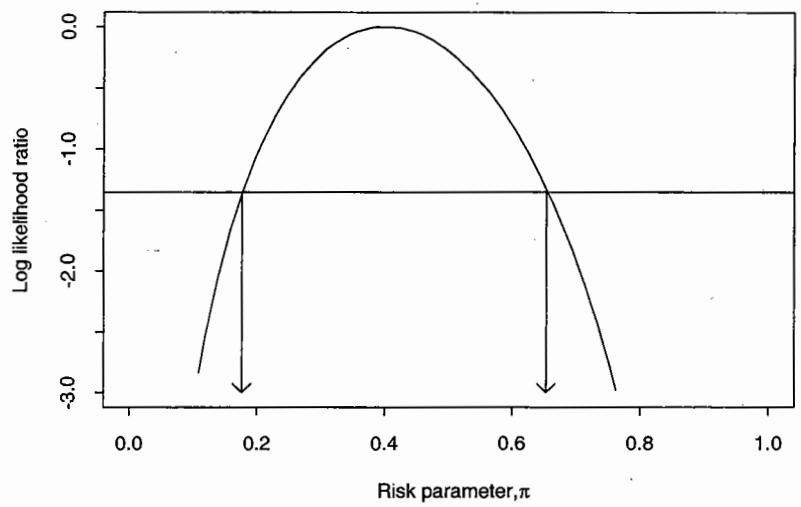

Fig. 3.5. The log likelihood ratio for π.

Exercise 3.2. Calculate the log likelihood when $\pi = 0.5$ and when $\pi = 0.1$.

The log likelihood takes its maximum at the same value of π as the likelihood, namely $\pi = 0.4$, so its maximum is

$$4\log(0.4) + 6\log(0.6) = -6.730.$$

To obtain the log likelihood *ratio*, this maximum must be *subtracted* from the log likelihood. A graph of the log likelihood ratio is shown in Fig. 3.5. The supported range for π can be found from this graph in the same way as from the likelihood ratio graph, by finding those values of π for which the log likelihood ratio is greater than

$$\log(0.258) = -1.353.$$

Exercise 3.3. Calculate the log likelihood ratios for $\pi = 0.1$ and $\pi = 0.5$. Are these values of π in the supported range?

In general, the log likelihood for π, when D subjects fail and $N - D$ survive, is

$$D\log(\pi) + (N-D)\log(1-\pi).$$

We shall show in Chapter 9 that this expression takes its maximum value when $\pi = D/N$, the observed proportion of subjects who failed.

If the binary model is parametrized in terms of the odds parameter, Ω, by substituting $\Omega/(1+\Omega)$ for π and $1/(1+\Omega)$ for $(1-\pi)$, we obtain the log likelihood

$$D\log(\Omega) - N\log(1+\Omega).$$

This takes its maximum value when $\Omega = D/(N-D)$, the ratio of the number or failures to the number of survivors. The maximum value of the log likelihood is the same whether the log likelihood is expressed in terms of π or Ω.

3.4 Censoring in follow-up studies

In our discussion of follow-up studies of the occurrence of disease events, or failures, we have assumed that all subjects are potentially observed for the same fixed period. In most practical studies there will be some subjects whose follow-up is incomplete. This will occur

- when they die from other causes before the end of the follow-up interval;
- when they migrate and are no longer covered by the record system which registers failures;
- when they join the cohort too late to complete the follow-up period.

In all three cases the observation time for the subject is said to be censored. In fact, the first type of loss to follow-up, failure due to a *competing cause*, is rather different from the remaining two, but they are usually grouped together and dealt with in the same way. In Chapter 7 we shall discuss the justification for this practice. For the moment, we assume it to be reasonable.

Censoring puts our argument in some difficulty. The model allows for only two outcomes, failure and survival, while our data contains three, failure, survival, and censoring. For the present we shall avoid this difficulty with a simple pretence. As an illustration, suppose we have followed 1000 men for five years, during which 28 suffered myocardial infarction and 972 did not, but observation of 15 men was censored before completion of five years follow-up. If all 15 men were withdrawn from study on the *first* day of the follow up period, the size of the cohort would be 985 rather than 1000. Conversely, if they were all withdrawn on the *last* day, censoring could be ignored and the cohort size treated as a full 1000. When censoring is evenly spread over the study interval, we would expect an answer which lies somewhere in between these two extreme assumptions. This suggests treating the effective cohort size as 992.5 — mid-way between 985 and 1000. This convention is equivalent to the assumption that 7.5 subjects are censored on the first day of follow up and 7.5 on the last day.

Table 3.1. Genotypes of 7 probands and their parents

Proband's genotype	Parents' genotypes		Number
	Mother	Father	
(a,c)	(a,b)	(c,d)	4
(b,d)	(a,b)	(c,d)	1
(a,c)	(a,b)	(c,c)	2

With only 15 subjects lost to follow up through censoring, this crude strategy for dealing with censoring is quite satisfactory, but if 150 were censored it could be seriously misleading. In Chapter 4 we shall see how this problem can be dealt with by extending the model.

3.5 Applications in genetics ★

The use of the log likelihood as a measure of support is of considerable importance in genetics. However, in that field it is conventional to use logarithms to the base 10 rather than natural logarithms. Since the two systems of logarithms differ only by a constant multiple (see Appendix A), this is only a trivial modification of the idea.

As an illustration of the use of log likelihood in genetics, we continue the example introduced in Exercises 2.4 and 2.5. Table 3.1 shows some hypothetical data which might have formed part of that collected in a study of an association between disease risk and presence of a certain HLA haplotype. If we were to observe a set of families over time, in order to relate the genotype to the eventual occurrence or non-occurrence of disease, then we could calculate a likelihood based on the probability of disease conditional upon genotype. However, such studies are logistically very difficult and are rarely done. Instead it is more usual to obtain, usually from clinicians, a collection of known cases of disease (*probands*) and their relatives, and to compare the genotypes of probands with the predictions from the model.

As in Exercise 2.5, we shall consider the model in which presence of a given haplotype, (a) say, leads to a risk of disease θ times as high as in its absence. Table 3.1 shows data concerning 7 probands and their parents. For each of the genetic configurations shown in the table, we derived the conditional probability of the genotype of a proband conditional on the genotypes of parents in Exercise 2.5 and we showed that these probabilities depend only on the risk ratio parameter θ.

Exercise 3.4. Write down the expression for the log likelihood as a function of the unknown risk ratio, θ, associated with presence of haplotype (a). What is the log likelihood ratio for the value $\theta = 1$ (corresponding to there being no increase in risk) as compared with $\theta = 6.0$ (which is the most likely value of θ in this case). Is the value $\theta = 1$ supported?

Solutions to the exercises

3.1 The probability of the observed data when $\pi = 0.4$ is

$$0.4^4 \times 0.6^6 = 1.19 \times 10^{-3}.$$

which is more than the probability when $\pi = 0.5$. It follows that $\pi = 0.4$ is more likely than $\pi = 0.5$.

3.2 The log likelihood when π=0.5 is

$$4\log(0.5) + 6\log(0.5) = -6.93.$$

The log likelihood when $\pi = 0.1$ is

$$4\log(0.1) + 6\log(0.9) = -9.84.$$

3.3 The maximum log likelihood, occurring at $\pi = 0.4$, is

$$4\log(0.4) + 6\log(0.6) = -6.73$$

so that the log likelihood ratio for $\pi = 0.5$ is $-6.93 - (-6.73) = -0.20$. For $\pi = 0.1$ it is $-9.84 - (-6.73) = -3.11$. Thus 0.5 lies within the supported range and 0.1 does not.

3.4 From the solution to Exercise 2.5, the conditional probabilities for each of the three genetic configurations are $\theta/(2\theta + 2)$, $1/(2\theta + 2)$, and $\theta/(\theta + 1)$. Thus, the log likelihood is

$$4\log\left(\frac{\theta}{2\theta+2}\right) + 1\log\left(\frac{1}{2\theta+2}\right) + 2\log\left(\frac{\theta}{\theta+1}\right).$$

At $\theta = 1.0$ this takes the value

$$4\log\left(\frac{1}{4}\right) + 1\log\left(\frac{1}{4}\right) + 2\log\left(\frac{1}{2}\right) = -8.318,$$

and at $\theta = 6.0$ (the most likely value) it is

$$4\log\left(\frac{6}{14}\right) + 1\log\left(\frac{1}{14}\right) + 2\log\left(\frac{6}{7}\right) = -6.337.$$

The log likelihood ratio for $\theta = 1$ is the difference between these, -1.981. Thus the parameter value $\theta = 1$ lies outside the limits of support we have suggested in this chapter.