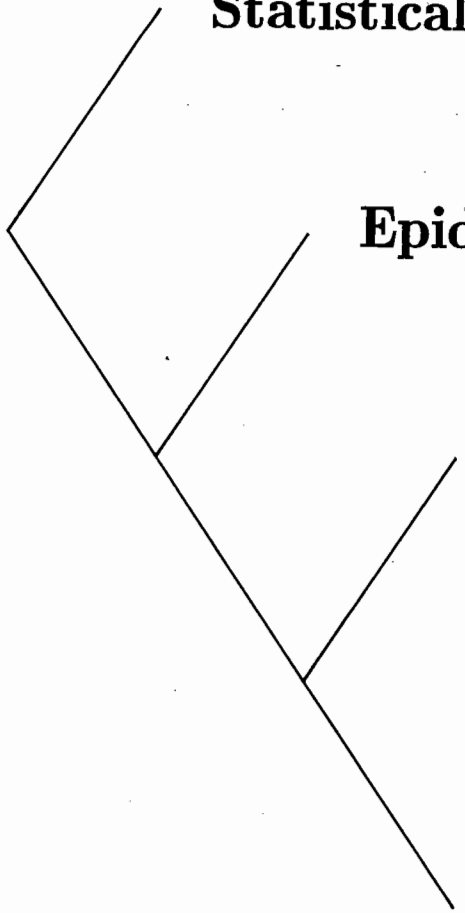
Statistical Models in Epidemiology



David Clayton
Medical Research Council, Cambridge

and

Michael Hills
London School of Hygiene and Tropical Medicine

OXFORD • NEW YORK • TOKYO
OXFORD UNIVERSITY PRESS

Oxford University Press, Walton Street, Oxford OX2 6DP
Oxford New York Toronto
Delhi Bombay Calcutta Madras Karachi
Kuala Lumpur Singapore Hong Kong Tokyo
Nairobi Dar es Salaam Cape Town
Melbourne Auckland Madrid
and associated companies in
Berlin Ibadan

Oxford is a trade mark of Oxford University Press

Published in the United States
by Oxford University Press Inc., New York

First published 1993
Reprinted 1994 (twice)

A catalogue record for this book is available from the British Library

Library of Congress Cataloging in Publication Data
Clayton, David, statistician.
Statistical models in epidemiology/David Clayton and Michael Hills.
Includes bibliographical references and index.
1. Epidemiology—Statistical methods. I. Hills, Michael. II. Title.
[DNLM: 1. Epidemiology. 2. Models, Statistical. WA 105 C622s 1993]
RA652.2M8C53 1993 614.4'072—dc20 93-19448

ISBN 0 19 852221 5 (Hbk)

Printed in Great Britain by
Biddles Ltd, Guildford & King's Lynn

Preface

The aim of this book is to give a self-contained account of the statistical basis of epidemiology. The book is intended primarily for students enrolled for a masters degree in epidemiology, clinical epidemiology, or biostatistics, and should be suitable both as the basis for a taught course and for private study.

Although we anticipate that most readers will have taken a first course in statistics, no previous knowledge is assumed, and the mathematical level of the book has been chosen to suit readers whose basic training is in biology. Some of the material in the book could be omitted at first reading, either because it is rather more demanding of mathematical skills or because it deals with rather specialized points. We have been careful to gather such material either into complete chapters or complete sections and to indicate these with a marginal symbol, as here. ⋆

Epidemiologists today have ready access to computer programs of great generality, but to use these sensibly and productively it is necessary to understand the ideas which lie behind them. The most important of these is the idea of a *probability model*. All statistical analysis of data is based on probability models, even though the models may not be explicit. Only by fully understanding the model can one fully understand the analysis.

Models depend on parameters, and values must be chosen for these parameters in order to match the model to the data. In showing how this is done we have chosen to emphasize the role of likelihood because this offers an approach to statistics which is both simple and intuitively satisfying. An additional advantage of this approach is that it requires the model and its parameters to be made explicit, even in the simplest situations. More complex problems can then be tackled by natural extensions of simple methods and do not require a whole new way of looking at things.

Most of the material in this book was developed during successive residential summer courses in epidemiology and statistics, held in Florence under the auspices of the European Educational Programme in Epidemiology. We are grateful to the International Agency for Cancer Research, the Regional Office for Europe of the World Health Organization, the Commission of the European Communities, and the Tuscany Regional Government, for sponsoring the program, and to Walter Davies, Organizing Secretary, and Rodolfo Saracci, Course Director, whose respective skills ensured that the course took place each year. We also acknowledge with thanks helpful

comments on earlier drafts from Damien Jolley, Bendix Carstensen, Dave Leon, and Nick Hills.

Cambridge
London
February 1993

David Clayton
Michael Hills

Dedication

To the students of the Florence course, 1988 – 92, without whose help and encouragement this book would never have appeared.

Contents

★ Denotes a chapter which could be omitted from a first course.

Part I

Probability models and likelihood

1
Probability models

1.1 Observation, experiments and models

Science proceeds by endless repetition of a three-stage process,

1. observation;
2. building a model to describe (or 'explain') the observations; and
3. using the model to predict future observations. If future observations are not in accord with the predictions, the model must be replaced or refined.

In quantitative science, the models used are mathematical models. They fall into two main groups, *deterministic* models and probability (or *stochastic*) models. It is the latter which are appropriate in epidemiology, but the former are more familiar to most scientists and serve to introduce some important ideas.

DETERMINISTIC MODELS

The most familiar examples of deterministic models are the laws of classical physics. We choose as a familiar example *Ohm's law*, which applies to the relationship between electrical potential (or voltage), V, applied across a conductor and the current flowing, I. The law holds that there is a strict proportionality between the two — if the potential is doubled then the current will double. This relationship is represented graphically in Fig. 1.1.

Ohm's law holds for a wide range of conductors, and simply states that the line in Fig. 1.1 is straight; it says nothing about the gradient of the line. This will differ from one conductor to another and depends on the resistance of the conductor. Without knowing the resistance it will not be possible to predict the current which will flow in any *particular* conductor. Physicists normally denote the resistance by R and write the relationship as

$$I = \frac{V}{R}.$$

However, R is a different sort of quantity from V or I. It is a *parameter* — a number which we must fix in order to apply the general law to a specific case. Statisticians are careful to differentiate between observable variables

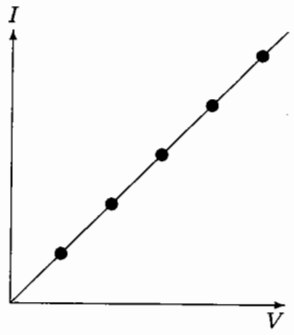


Fig. 1.1. A deterministic model: Ohm's law.

(such as V and I) and parameters (such as R) and use Greek letters for the latter. Thus, if Ohm were a modern statistician he would write his law as

$$I = \frac{V}{\rho}$$

In this form it is now clear that ρ, the resistance, is a parameter of a simple mathematical model which relates current to potential. Alternatively, he could write the law as

$$I = \gamma V$$

where γ is the conductance (the inverse of the resistance). This is a simple example of a process called *reparametrization* — writing the model differently so that the parameters take on different meanings.

STOCHASTIC MODELS

Unfortunately the phenomena studied by scientists are rarely as predictable as is implied by Fig. 1.1. In the presence of measurement errors and uncontrolled variability of experimental conditions it might be that real data look more like Fig. 1.2. In these circumstances we would not be in a position to predict a future observation with certainty, nor would we be able to give a definitive estimate of the resistance parameter. It is necessary to extend the deterministic model so that we can predict a range of more probable future observations, and indicate the uncertainty in the estimate of the resistance.

Problems such as this prompted the mathematician Gauss to develop his *theory of errors*, based on the Gaussian distribution (often also called the *Normal* distribution), which is the most important probability model for these problems. A very large part of statistical theory is concerned with this model and most elementary statistical texts reflect this. Epidemiology,

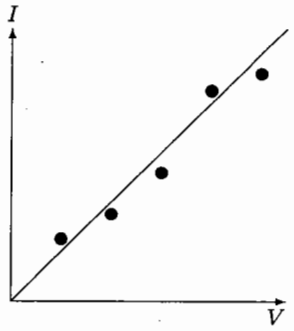


Fig. 1.2. Experimental/observational errors.

however, is more concerned with the occurrence (or not) of certain events in the natural history of disease. Since these occurrences cannot be described purely deterministically, probability models are also necessary here, but it is the models of Bernoulli and Poisson which are more relevant. The remainder of this chapter discusses a particularly important type of data generated by epidemiological studies, and the nature of the models we use in its analysis.

1.2 Binary data

Many epidemiological studies generate data in which the response measurement for each subject may take one of only two possible values. Such a response is called a *binary* response. Two rather different types of study generate such data.

COHORT STUDIES WITH FIXED FOLLOW-UP TIME

In a *cohort* study a group of people are followed through some period of time in order to study the occurrence (or not) of a certain event of interest. The simplest case is a study of *mortality* (from any cause). Clearly, there are only two possible outcomes for a subject followed, say, for five years — death or survival.

More usually, it is only death from a specified cause or causes which is of interest. Although there are now three possible outcomes for any subject — death from the cause of interest, death from another cause, or survival — such data are usually dealt with as binary data. The response is taken as death from cause of interest as against survival, death from other causes being treated as premature termination of follow-up. Premature termination of follow-up is a common feature of epidemiological and clinical follow-up studies and may occur for many reasons. It is called *censoring*, a word which reflects the fact that it is the underlying binary response which

we would have liked to observe, were it not for the removal of the subject from observation.

In *incidence studies* the event of interest is new occurrence of a specified disease. Again our interest is in the binary response (whether the disease occurred or not) although other events may intervene to censor our observation of it.

For greater generality, we shall use the word *failure* as a generic term for the event of interest, whether incidence, mortality, or some other (undesirable) outcome. We shall refer to non-failure as *survival*. In the simplest case, we study N subjects, each one being followed for a fixed time interval, such as five years. Over this time we observe D failures, so that $N - D$ survive. We shall develop methods for dealing with censoring in later chapters.

CROSS-SECTIONAL PREVALENCE DATA

Prevalence studies have considerable importance in assessing needs for health services, and may also provide indirect evidence for differences in incidence. They have the considerable merit of being relatively cheap to carry out since there is no follow-up of the study group over time. Subjects are simply categorized as affected or not affected, according to agreed clinical criteria, at some fixed point in time. In a simple study, we might observe N subjects and classify D of them as affected. An important example is serological studies in infectious-disease epidemiology, in which subjects are classified as being seropositive or seronegative for a specified infection.

1.3 The binary probability model

The obvious analysis of our simple binary data consisting of D failures out of N subjects observed is to compute the proportion failing, D/N. However, knowing the proportion of a cohort which develops a disease, or dies from a given cause, is of little use unless it can be assumed to have a wider applicability beyond the cohort. It is in making this passage from the particular to the general that statistical models come in. One way of looking at the problem is as an attempt to predict the outcome for a new subject, similar to the subjects in the cohort, but whose outcome is unknown. Since the outcome for this new subject cannot be predicted with certainty the prediction must take the form of *probabilities* attached to the two possible outcomes. This is the *binary probability model*. It is the simplest of all probability models and, for the present, we need to know nothing of the properties of probability save that probabilities are numbers lying in the range 0 to 1, with 0 representing an impossible outcome and 1 representing a certain outcome, and that the probability of occurrence of either one of two distinct outcomes is the sum of their individual probabilities (the *additive* rule of probability).

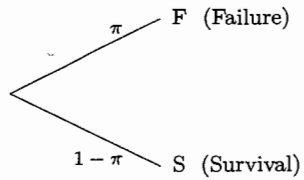


Fig. 1.3. The binary probability model.

THE RISK PARAMETER

The binary probability model is illustrated in Figure 1.3. The two outcomes are labelled F (failure) and S (survival). The model has one *parameter*, π, the probability of failure. Because the subject must either fail or survive, the sum of the probabilities of these two outcomes must be 1, so the probability of survival is $1-\pi$. In the context where π represents the probability of occurrence of an event in a specified time period, it is usually called the *risk*.

THE ODDS PARAMETER

An important alternative way of parametrizing the binary probability model is in terms of the *odds* of failure versus survival. These are

$$\pi : (1 - \pi),$$

which may also be written as

$$\frac{\pi}{1-\pi} : 1.$$

It is convenient to omit the : 1 in the above expression and to measure the odds by the fraction

$$\frac{\pi}{1-\pi}.$$

This explains why, although the word odds is plural, there is often only one number which measures the odds.

Exercise 1.1. Calculate the odds of F to S when the probability of failure is (a) 0.75, (b) 0.50, (c) 0.25.

In general the relationship between a probability π and the corresponding odds Ω is

$$\Omega = \frac{\pi}{(1-\pi)}.$$

This can be inverted to give

$$\pi = \frac{\Omega}{1+\Omega}, \quad 1-\pi = \frac{1}{1+\Omega}.$$

Exercise 1.2. Calculate the probability of failure when Ω, the odds of F to S is (a) 0.3, (b) 3.0.

RARE EVENTS

In this book we shall be particularly concerned with *rare events*, that is, events with a small probability, π, of occurrence in the time period of interest. In this case $(1-\pi)$ is very close to 1 and the odds parameter and the risk parameter are nearly equal:

$$\Omega \approx \pi.$$

This approximation is often called the *rare disease assumption*, but this is a misleading term, since even the common cold has a small probability of occurrence within, say, a one-week time interval.

1.4 Parameter estimation

Without giving a value to the parameter π, this model is of no use for prediction. Our next problem is to use our observed data to estimate its value. It might seem obvious to the reader that we should estimate π by the proportion of failures, D/N. This corresponds to estimating the odds parameter Ω by $D/(N-D)$, the ratio of failures to survivors.

It might also seem obvious that we should place more reliance on our estimate (and upon any predictions based on it) if N is 1000 than if N is 10. The formal statistical theory which provides a quantitative justification for these intuitions will be discussed in later chapters.

1.5 Is the model true?

A model which states that every one of a group of patients has the same probability of surviving five years will seem implausible to most clinicians. Indeed, the use of such models by statisticians is a major reason why some practitioners, brought up to think of each patient as unique, part company with the subject!

The question of whether scientific models are *true* is not however, a sensible one. Instead, we should ask ourselves whether our model is *useful* in describing past observations and predicting future ones. Where there remains a choice of models, we must be guided by the criterion of *simplicity*. In epidemiology probability models are used to describe past observations of disease events in study cohorts and to make predictions for future individuals. If we have no further data which allows us to differentiate subjects

in the cohort from one another or from a future individual, we have no option save to assign the same probability of failure to each subject. Further data allows elaboration of the model. For example, if we can identify subjects as exposed or unexposed to some environmental influence, the model can be extended to assign different probabilities to exposed and unexposed subjects. If additionally we know the level of exposure we can extend the model by letting the probability of failure be some increasing function of exposure.

In this book we shall demonstrate the manner in which more complicated models may be developed to deal with more detailed data. The binary model has been our starting point since it is the basic building brick from which more elaborate models are constructed.

Solutions to the exercises

1.1 (a) Odds $= 0.75/0.25 = 3$.
(b) Odds $= 0.50/0.50 = 1$.
(c) Odds $= 0.25/0.75 = 0.3333$.

1.2 (a) Probability $= 0.3/1.3 = 0.2308$.
(b) Probability $= 3/4 = 0.75$.

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:

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

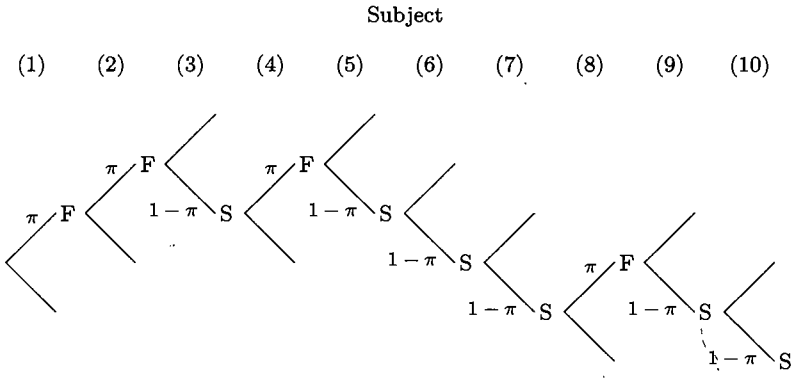


Fig. 3.1. Study outcomes for 10 subjects.

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,

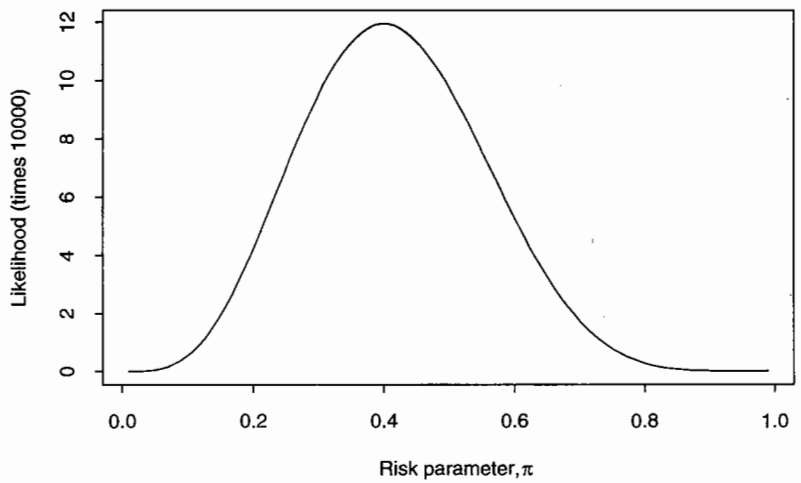


Fig. 3.2. The likelihood for π.

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

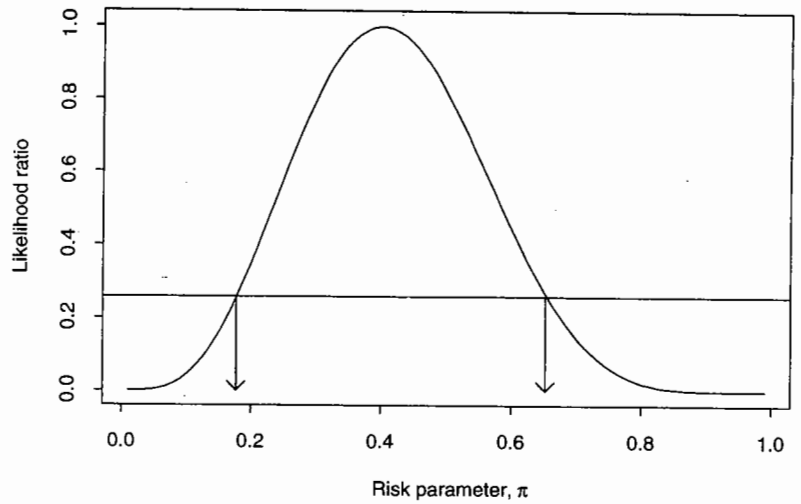


Fig. 3.3. The likelihood ratio for π.

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.

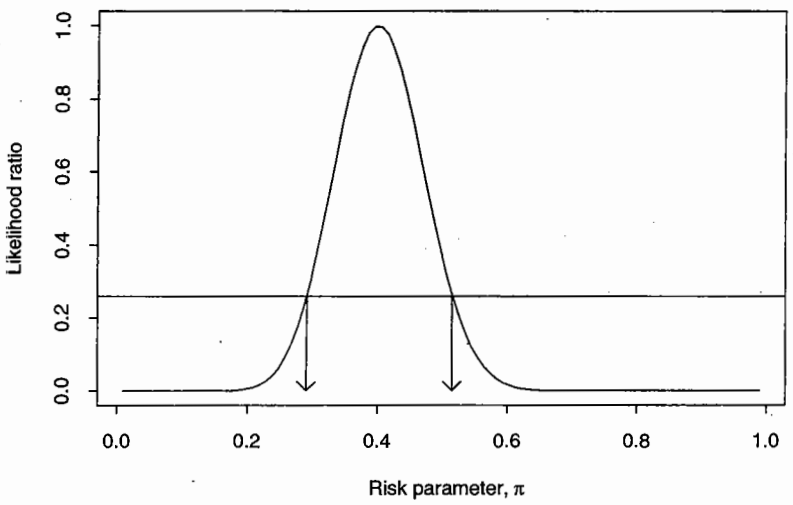


Fig. 3.4. The likelihood ratio based on 20 failures in 50 subjects.

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.

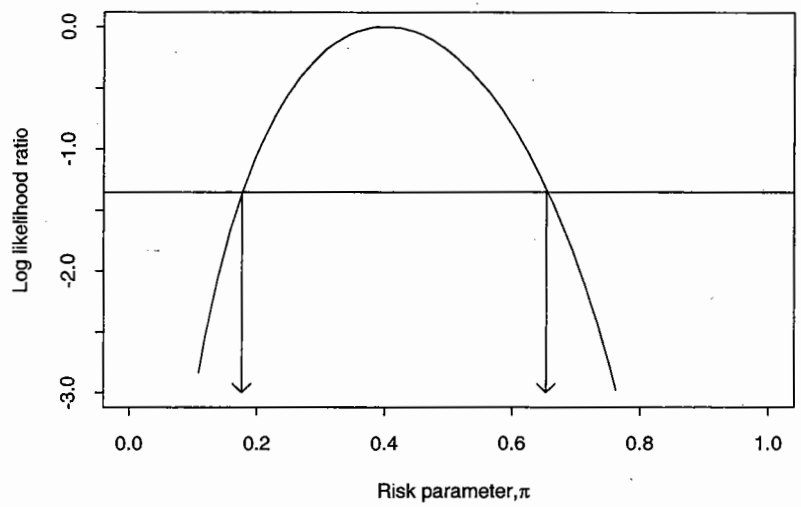


Fig. 3.5. The log likelihood ratio for π.

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.

4
Consecutive follow-up intervals

In the last chapter we touched on the difficulty of estimating the probability of failure during a fixed follow-up period when the observation times for some subjects are censored. A second problem with fixed follow-up periods is that it may be difficult to compare the results from different studies; a five-year probability of failure can only be compared with other five-year probabilities of failure, and so on. Finally, by ignoring *when* the failures took place, all information about possible changes in the probability of failure during follow-up is lost.

The way round these difficulties is to break down the total follow-up period into a number of shorter consecutive intervals of time. We shall refer to these intervals of time as *bands*. The experience of the cohort during each of these bands can then be used to build up the experience over any desired period of time. This is known as the *life table* or *actuarial* method. Instead of a single binary probability model there is now a sequence of binary models, one for each band. This sequence can be represented by a conditional probability tree.

4.1 A sequence of binary models

Consider an example in which a three-year follow-up interval has been divided into three one-year bands. The experience of a subject during the three years may now be described by a sequence of binary probability models, one for each year, as shown by the probability tree in Fig.4.1. The four possible outcomes for this subject, corresponding to the tips of the tree, are

1. failure during the first year;
2. failure during the second year;
3. failure during the third year;
4. survival for the full three-year period.

The parameter of the first binary model in the sequence is π^1, the probability of failure during the first year; the parameter of the second binary model is π^2, the probability of failure during the second year, given the subject has not failed before the start of this year, and so on. These are

analysis. If survival is analyzed by time in study there are no late entries, but in an analysis of the same study by age, or by time since entering an occupation, there will be late entries.

Solutions to the exercises

7.1 The estimated 5-year risk of myocardial infarction is 27/1000 while that for stroke is 8/1000. The risk of a cardiovascular event is 35/1000.

7.2 The outcomes and their probabilities are listed below.

Outcome	Probability
Band 1	
F1	0.1
F2	0.2
Band 2	
F1	$0.7 \times 0.1 = 0.07$
F2	$0.7 \times 0.2 = 0.14$
Band 3	
F1	$0.7 \times 0.7 \times 0.1 = 0.049$
F2	$0.7 \times 0.7 \times 0.2 = 0.098$
S	$0.7 \times 0.7 \times 0.7 = 0.343$

8
The Gaussian probability model

Until now we have been concerned only with the binary probability model. In this model there are two possible outcomes and the total probability of 1 is shared between them. It is an appropriate model when studying the occurrence of events, but not when studying a response for which there are many possible outcomes, such as blood pressure. For this the *Gaussian* or *normal* probability model is most commonly used.

In the Gaussian model the total probability of 1 is shared between many values. This is illustrated in the left panel of Fig. 8.1. When measurements are recorded to a fixed number of decimal places, there is a finite number of possible outcomes but, in principle, such measurements have infinitely many possible outcomes, so the probability attached to any one is effectively zero. For this reason it is the probability *density* per unit value which is specified by the model, not the probability of a given value. This is illustrated in the right panel of the figure. If π is the probability shared between values in a very narrow range, width h units, the probability density is π/h.

8.1 The standard Gaussian distribution

The standard Gaussian distribution has probability density centred at 0. The probability density at any value z (positive or negative) is given by

$$0.3989 \exp\left[-\frac{1}{2}(z)^2\right].$$

A graph of this probability density for different values of z is shown in Fig. 8.2. There is very little probability outside the range ± 3.

Tables of the standard Gaussian distribution are widely available, and these readily allow calculation of the probability associated with specified ranges of z. For our purposes it is necessary only to record that the probability corresponding to the range $(-1.645, +1.645)$ is 0.90 and that for the range $(-1.960, +1.960)$ is 0.95.

If the probability model for z is a standard Gaussian distribution then the probability model for $(z)^2$ is called the *chi-squared* distribution on one degree of freedom. Tables of chi-squared distributions can be used to find

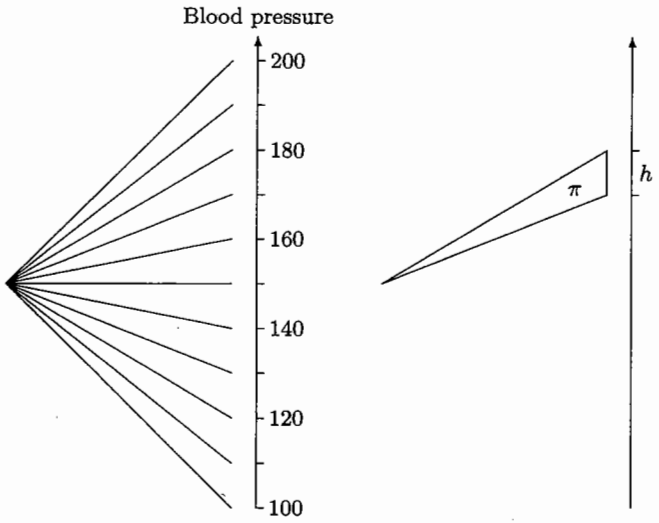


Fig. 8.1. Probability shared between many outcomes.

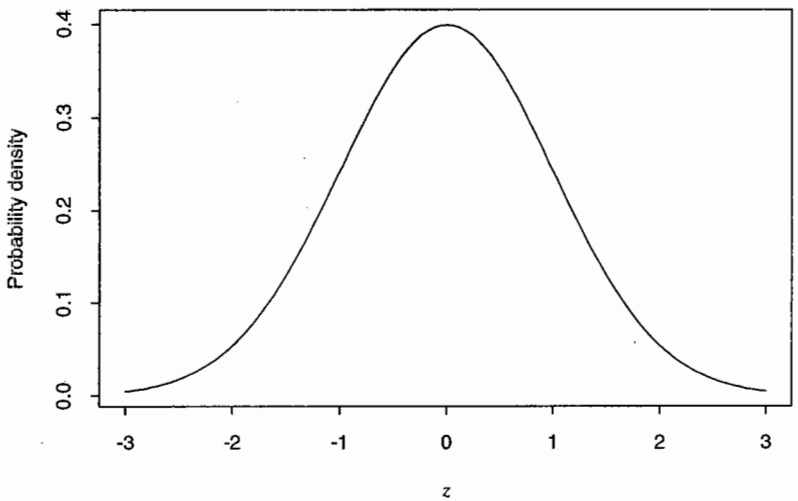


Fig. 8.2. The standard Gaussian distribution.

the probabilities of exceeding specified values of $(z)^2$ in the same way as tables of the standard Gaussian distribution are used to find probabilities of exceeding specified values of z.

Exercise 8.1. Use the tables in Appendix D to find the probability of exceeding the value 2.706 in a chi-squared distribution on one degree of freedom.

Note that, for $(z)^2$ to exceed 2.706, z must lie outside the range ± 1.645 of the standard normal distribution.

8.2 The general Gaussian model

It would be remarkable if the data we are analysing fell into the range -3 to $+3$, so for modelling the variability of real data, it is necessary to generalize the model to incorporate two parameters, one for the central value or *location*, and one for the spread or *scale* of the distribution. These are called the *mean* parameter and *standard deviation* parameter and are usually denoted by μ and σ respectively. A variable with such a distribution is derived by multiplying z by the scale factor and adding the location parameter. Thus

$$x = \mu + \sigma z.$$

has a distribution of the same general shape as the standard Gaussian distribution but centred around μ with most of its probability between $\mu - 3\sigma$ and $\mu + 3\sigma$.

Exercise 8.2. If the mean and standard deviation of a general Gaussian distribution are 100 and 20 respectively, what ranges of values correspond to probabilities of 0.90 and 0.95 respectively?

Similarly, when x has a Gaussian distribution with mean μ and standard deviation σ then

$$z = \left(\frac{x-\mu}{\sigma}\right)$$

will have a *standard* Gaussian distribution. This fact can be used get the probability for a range of values of x using tables of z.

The probability density per unit of x when x has a Gaussian distribution with mean μ and standard deviation σ is

$$\frac{0.3989}{\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right].$$

This expression is obtained by substituting $(x-\mu)/\sigma$ for z in the probability density of a standard Gaussian distribution to obtain the probability density per σ units of x, and then dividing by σ to obtain the probability density per unit of x. Sometimes the distribution is described in terms of the square of σ, which is called the *variance*.

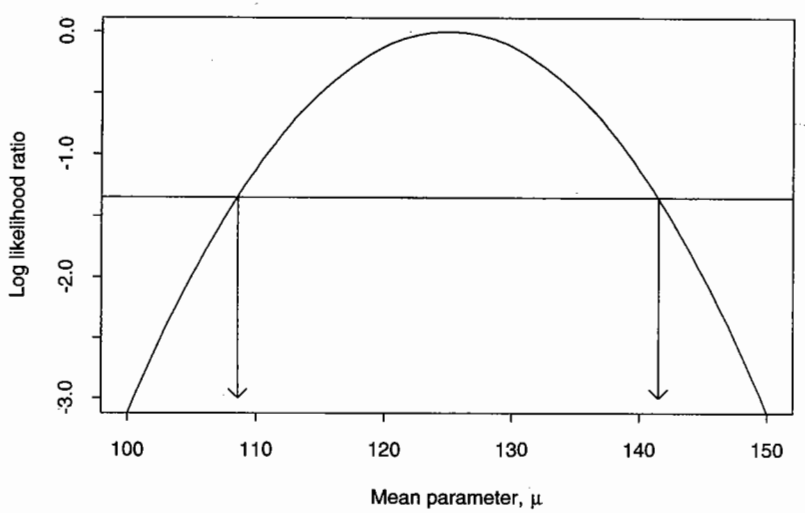


Fig. 8.3. The log likelihood ratio for the Gaussian mean, μ.

8.3 The Gaussian likelihood

Suppose a single value of x, say $x = 125$ is observed. Using the probability model that this is an observation from a Gaussian distribution with parameters μ and σ, the log likelihood for μ and σ is given by the log of the corresponding Gaussian probability density:

$$\log(0.3989) - \log(\sigma) - \frac{1}{2}\left(\frac{125-\mu}{\sigma}\right)^2.$$

This log likelihood depends on two unknown parameters, but to keep things simple we shall assume that one of them, σ, is known from past experience to have the value 10. Omitting constant terms, the log likelihood for μ is then

$$-\frac{1}{2}\left(\frac{125-\mu}{10}\right)^2.$$

The most likely value of μ is 125 and, since the above expression is zero at this point, this expression also gives the log likelihood *ratio* for μ. This is plotted in Fig. 8.3; curves with this shape are called *quadratic*.

We saw in Chapter 3 that we take the extremes of the supported range for a parameter to correspond to the value -1.353 for the log likelihood ratio. To find the limits of the supported range for μ we must therefore

solve the simple equation

$$-\frac{1}{2}\left(\frac{125-\mu}{10}\right)^2 = -1.353.$$

This takes only a few lines:

$$\begin{aligned}\left(\frac{125-\mu}{10}\right)^2 &= 2.706,\\ \left(\frac{125-\mu}{10}\right) &= \pm 1.645,\\ \mu &= 125 \pm 1.645 \times 10,\end{aligned}$$

so that supported values of μ are those between 108.6 and 141.5. In general, the log likelihood ratio for μ is

$$-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2,$$

the most likely value of μ is the observation x, and the supported range for μ is

$$x \pm 1.645\sigma,$$

where σ is the standard deviation (which we assume to be known).

We saw in Exercise 8.1 that the probability of exceeding 2.706 in a chi-squared distribution is 0.10, and the probability corresponding to the range ± 1.645 in the standard Gaussian distribution is 0.90. The fact that these numbers turn up in the above calculation is no accident and suggests that the log likelihood ratio criterion of -1.353 leads to supported ranges which have something to do with a probability of 0.90. This is indeed the case, but the relationship is not altogether straightforward and we shall defer this discussion to Chapter 10.

8.4 The likelihood with N observations

When there are N observations

$$x_1, x_2, \ldots, x_N,$$

the log likelihood for μ is obtained by adding the separate log likelihoods for each observation giving

$$\sum -\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2.$$

Let M refer to the mean of the observations,

$$M = \frac{x_1 + x_2 + \cdots + x_N}{N}.$$

It can be shown that the log likelihood can be rearranged as

$$-\frac{1}{2}\left(\frac{M-\mu}{S}\right)^2 + \sum -\frac{1}{2}\left(\frac{x_i - M}{\sigma}\right)^2$$

where $S = \sigma/\sqrt{N}$, sometimes called the *standard error of the mean.* This rearrangement involves only elementary algebra and the details are omitted. The second part of this new expression for the log likelihood does not depend on μ and cancels in the log likelihood ratio for μ which is

$$-\frac{1}{2}\left(\frac{M-\mu}{S}\right)^2,$$

The most likely value of μ is M, and setting the log likelihood ratio equal to -1.353 to obtain a supported range for μ gives

$$\mu = M \pm 1.645S.$$

As we would expect, with larger N, the value of S becomes smaller and the supported range narrower.

Exercise 8.3. The following measurements of systolic blood pressure were obtained from a sample of 20 men.

98	160	136	128	130	114	123	134	128	107
123	125	129	132	154	115	126	132	136	130

What is the most likely value for μ? Assuming that $\sigma = 14$, calculate the range of supported values for μ.

This exercise continues to make the unrealistic assumption, made throughout this chapter, that σ is *known.* In practice it must almost invariably be estimated from the data. We shall defer discussion of this until Chapter 34.

Solutions to the exercises

8.1 The probability of exceeding 2.706 in the chi-squared distribution with one degree of freedom is 0.10.

8.2 The range corresponding to a probability of 0.9 is

$$100 \pm 1.645 \times 20 = (67.1, 132.9)$$

and, for a probability of 0.95,

$$100 \pm 1.96 \times 20 = (60.8, 139.2).$$

8.3 The mean of the 20 measurements is 128.00 and this is the most likely value of μ. To calculate the supported range for μ, we first calculate

$$S = \frac{\sigma}{\sqrt{N}} = \frac{14}{\sqrt{20}} = 3.13$$

so that the range lies between

$$\mu = 128.00 \pm 1.645 \times 3.13$$

that is from 122.9 to 133.1 .

9
Approximate likelihoods

Because the Gaussian log likelihood for the mean parameter, μ, takes the simple form

$$-\frac{1}{2}\left(\frac{M-\mu}{S}\right)^2$$

the supported range for μ also takes a simple form, namely

$$M \pm 1.645S.$$

For log likelihoods such as the Bernouilli and Poisson there is no simple algebraic expression for the supported range, and the values of the parameters at which the log likelihood is exactly -1.353 must be found by systematic trial and error. However, the shapes of these log likelihoods are *approximately* quadratic, and this fact can be used to derive simple formulae for approximate supported ranges. Methods based on quadratic approximation of the log likelihood are particularly important because the quadratic approximation becomes closer to the true log likelihood as the amount of data increases.

9.1 Approximating the log likelihood

Consider a general likelihood for the parameter, θ, of a probability model and let M be the most likely value of θ. Since the quadratic expression

$$-\frac{1}{2}\left(\frac{M-\theta}{S}\right)^2$$

has a maximum value of zero when $\theta = M$ it can be used to to approximate the true log likelihood ratio, after an appropriate value of S has been chosen. Small values of S give quadratic curves with sharp peaks and large values of S give quadratic curves with broad peaks. We shall refer to S as the standard deviation of the estimate of θ. Alternatively, it is sometimes called the *standard error* of the estimate.

Once M has been found and S chosen, an approximate supported range for θ is found by solving the equation

$$-\frac{1}{2}\left(\frac{M-\theta}{S}\right)^2 = -1.353,$$

to give

$$\theta = M \pm 1.645S.$$

Full details of how S is chosen are given later in the chapter, but for the moment we shall give formulae for S, without justification, and concentrate on how to use these in practice.

THE RISK PARAMETER

The log likelihood for π, the probability of failure, based on D failures and $N - D$ survivors is

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

The most likely value of π is D/N. To link with tradition we shall also refer to the most likely value of π as P (for proportion). The value of S which gives the best approximation to the log likelihood ratio is

$$S = \sqrt{\frac{P(1-P)}{N}}.$$

For the example we worked through in Chapter 3, $D = 4$ and $N = 10$ so that the value of P is 0.4 and

$$S = \sqrt{\frac{0.4 \times 0.6}{10}} = 0.1549.$$

An approximate supported range for π is given by

$$0.4 \pm 1.645 \times 0.1549$$

which is from 0.15 to 0.65, while the supported range obtained from the true curve lies from 0.17 to 0.65. The true and approximate log likelihood curves are shown in Fig. 9.1. The curve shown as a solid line is the true log likelihood ratio curve, while the broken line indicates the Gaussian approximation.

THE RATE PARAMETER

The log likelihood for a rate λ based on D cases and Y person years is

$$D\log(\lambda) - \lambda Y.$$

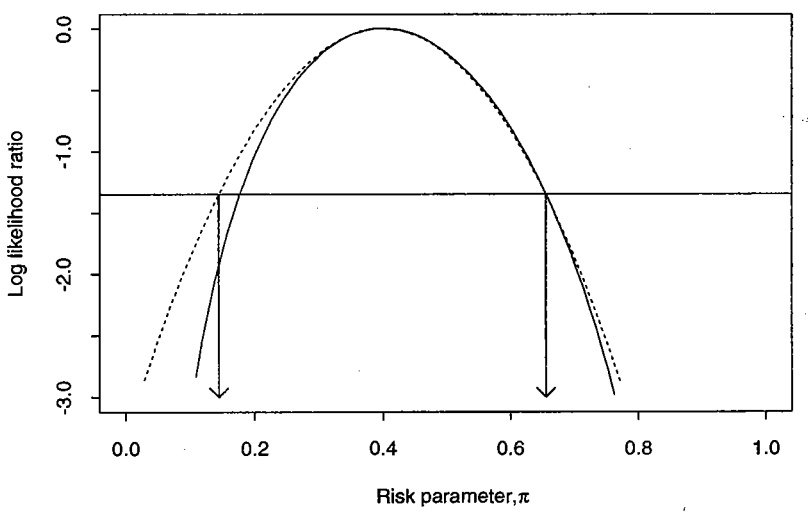


Fig. 9.1. True and approximate Bernouilli log likelihoods.

The most likely value of λ is D/Y and the value of S which gives the best approximation to the log likelihood ratio is

$$S = \frac{\sqrt{D}}{Y}.$$

For the example in Chapter 5, $D = 7$ and $Y = 500$. The most likely value of λ is 0.014 and

$$S = \sqrt{7}/500 = 0.00529.$$

An approximate supported range for λ is therefore

$$0.014 \pm 1.645 \times 0.00529$$

which is from 5.3/1000 to 22.7/1000. The true (solid line) and approximate (broken line) log likelihood ratio curves are shown in Fig. 9.2. The range of support obtained from the true curve spans from 7.0 to 24.6 per 1000.

Exercise 9.1. Find the approximate supported range for π, the probability of failure, based 7 failures and 93 survivors. Find also the approximate supported range for λ, the rate of failure, based on 30 failures over 1018 person-years.

9.2 Transforming the parameter

The Gaussian log likelihood curve for μ is symmetric about M and extends *indefinitely* to either side. However, the parameters of some probability

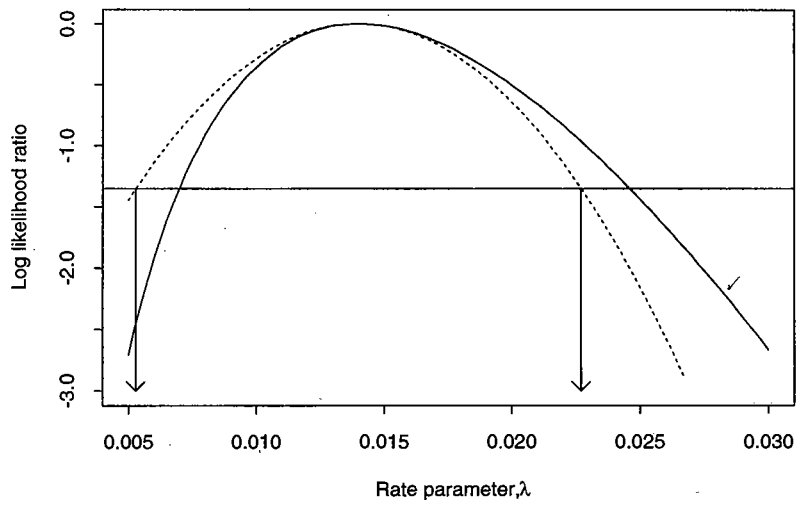


Fig. 9.2. True and approximate Poisson log likelihoods.

models are not free to vary in this manner. For example, the rate parameter λ can take only positive values, and the risk parameter must lie between 0 and 1. Approximate supported ranges for such parameters calculated from the Gaussian approximation can, therefore, include impossible values.

The solution to this problem is to find some function (or *transformation*) of the parameter which is unrestricted and to first find an approximate supported range for the transformed parameter.

THE LOG RATE PARAMETER

The rate parameter λ can take only positive values, but its logarithm is unrestricted. To calculate an approximate supported range for λ it is better, therefore, to first calculate a range for $\log(\lambda)$, and then to convert this back to a range for λ. Note that the range for $\log(\lambda)$ will always convert back to positive values for λ. To find the approximate range for $\log(\lambda)$ we need a new value of S — that which gives the best Gaussian approximation to the log likelihood ratio curve when plotted against $\log(\lambda)$. When a rate λ is estimated from D failures over Y person-years, this value of S is given by

$$S = \sqrt{1/D}.$$

Fig. 9.3 illustrates this new approximation for our example in which $D = 7$ and $Y = 500$ person-years. Here,

$$S = \sqrt{1/7} = 0.3780,$$

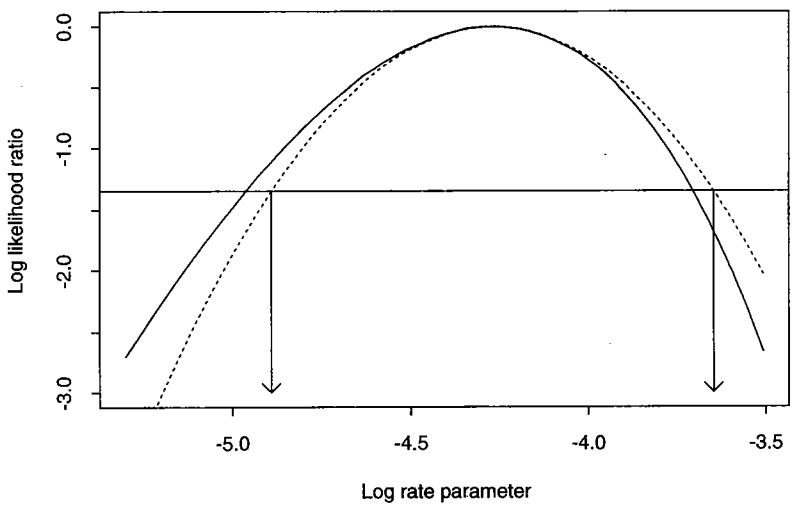


Fig. 9.3. Approximating the log likelihood for $\log(\lambda)$.

and an approximate supported range for $\log(\lambda)$ is

$$\log(7/500) \pm 1.645 \times \sqrt{1/7},$$

which is from -4.890 to -3.647. The range for λ is therefore from $\exp(-4.890)$ to $\exp(-3.647)$ which spans from 7.5/1000 to 26.1/1000.

A more convenient way of carrying out this calculation is suggested by noting that the limits of the range for λ are given by

$$\frac{7}{500} \overset{\times}{\div} \exp\left(1.645\sqrt{\frac{1}{7}}\right) = 0.014 \overset{\times}{\div} 1.862.$$

The range is then from $0.014/1.862 = 7.5/1000$ to $0.014 \times 1.862 = 26.1/1000$, as before. We shall refer to the quantity

$$\exp(1.645S)$$

as an *error factor*.

THE LOG ODDS PARAMETER

The same thing can be done when calculating a supported range for the risk parameter π based on D failures in N subjects. The value of π is restricted on both sides, by 0 on the left and by 1 on the right. The value of $\log(\pi)$ is

still restricted on the right by zero because $\log(1) = 0$, but $\log(\Omega)$, where Ω is the odds corresponding to π, is not restricted at all. Hence we first find a range for $\log(\Omega)$ and then convert this back to a range for π. The most likely value of $\log(\Omega)$ is

$$M = \log\left(\frac{D}{N-D}\right)$$

and the value of S for approximating the log likelihood for $\log(\Omega)$ is

$$S = \sqrt{\frac{1}{D} + \frac{1}{N-D}}.$$

For the example where $D = 4$ and $N - D = 6$,

$$S = \sqrt{\frac{1}{4} + \frac{1}{6}} = 0.6455,$$

and an approximate supported range for $\log(\Omega)$ is given by

$$\log\left(\frac{4}{6}\right) \pm 1.645 \times 0.6455,$$

that is, from -1.4673 to 0.6564. This is a range for $\log(\Omega)$ and it is equivalent to a range for Ω from $\exp(-1.4673) = 0.231$ to $\exp(0.6564) = 1.928$. This can be calculated more easily by first calculating the error factor

$$\exp(1.645 \times 0.6455) = 2.892.$$

The most likely value of Ω is $4/6 = 0.667$, so that the supported range for Ω is

$$0.667 \overset{\times}{\div} 2.892$$

that is, from 0.231 to 1.928 as before. Finally, remembering that $\pi = \Omega/(1+\Omega)$, the range for π is given by

$$\frac{0.231}{1.231} \text{ to } \frac{1.928}{2.928}$$

which is from 0.19 to 0.66.

Some of the more commonly used values of S obtained by approximating the log likelihood are gathered together in Table 9.1.

Exercise 9.2. Repeat Exercise 9.1 by first finding 90% intervals for $\log(\Omega)$ and $\log(\lambda)$ respectively, and then converting these to intervals for π and λ.

Exercise 9.3. Repeat the above exercise using error factors.

Table 9.1. Some important Gaussian approximations

Parameter	M	S
π	$D/N = P$	$\sqrt{P(1-P)/N}$
λ	D/Y	$\sqrt{D}/Y$
$\log(\Omega)$	$\log[D/(N-D)]$	$\sqrt{1/D + 1/(N-D)}$
$\log(\lambda)$	$\log(D/Y)$	$\sqrt{1/D}$

★ 9.3 Finding the best quadratic approximation

We now return to the problem of how to determine the values for M and S. To do this we need some elementary ideas of calculus summarized in Appendix B. In particular, we need to be able to find the *gradient* (or *slope*) of the log-likelihood curve together with its *curvature*, which is defined as the rate of change of the gradient. The mathematical terms for these quantities are the first and second *derivatives* of the log likelihood function.

The value of M can be found by a direct search for that value of of θ which maximizes the log likelihood, but it is often easier to find the value of θ for which the gradient of the log likelihood is zero; this occurs when $\theta = M$.

The value of S is chosen to make the curvature of the quadratic approximation equal to that of the true log likelihood curve at M, thus ensuring that the true and approximate log likelihoods are very close to each other near $\theta = M$. The quadratic approximation to the log likelihood ratio is

$$-\frac{1}{2}\left(\frac{M-\theta}{S}\right)^2,$$

and the rules summarized in Appendix B show that the curvature of this is constant and takes the value

$$-\frac{1}{(S)^2}.$$

We therefore choose the value of S to make $-1/(S)^2$ equal to the curvature of the true log likelihood curve at its peak.

THE RATE PARAMETER

The log likelihood for a rate λ is

$$D\log(\lambda) - \lambda Y.$$

Using the rules of calculus given in Appendix B the gradient of $\log(\lambda)$ is $1/\lambda$ and the gradient of λ is 1. Hence the gradient of the log likelihood is

$$\frac{D}{\lambda} - Y.$$

The maximum value of the log likelihood occurs when the gradient is zero, that is, when $\lambda = D/Y$, so the most likely value of λ is D/Y. The curvature of a graph at a point is defined as the rate of change of the gradient of the curve at that point. The rules of calculus show this to be

$$-\frac{D}{(\lambda)^2}.$$

The peak of the log likelihood occurs at $\lambda = D/Y$ so the curvature at the peak is found by replacing λ by D/Y in this expression to obtain

$$-\frac{(Y)^2}{D}.$$

Setting this equal to $-1/(S)^2$ gives

$$S = \sqrt{D}/Y,$$

which is the formula quoted earlier.

THE RISK PARAMETER

The log likelihood for the probability π based on D positive subjects out of a total of N is

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

The gradient of the log likelihood is

$$\frac{D}{\pi} - \frac{N-D}{1-\pi}$$

which is zero at $\pi = D/N$, also referred to as P. The gradient of the gradient is

$$-\frac{D}{(\pi)^2} - \frac{N-D}{(1-\pi)^2},$$

so the curvature at $\pi = P$ is

$$-\frac{D}{(P)^2} - \frac{N-D}{(1-P)^2}.$$

Replacing D by NP and $N-D$ by $N(1-P)$, this reduces to

$$-\frac{N}{P(1-P)}$$

so

$$S=\sqrt{\frac{P(1-P)}{N}}.$$

⋆ 9.4 Approximate likelihoods for transformed parameters

When the log likelihood for a parameter is plotted against the log of the parameter rather than the parameter itself, the curvature at the peak will be different. For example, the log likelihood for a rate parameter λ is

$$D\log(\lambda)-\lambda Y.$$

Plotting this against $\log(\lambda)$ is the same as expressing the log likelihood as a function of $\log(\lambda)$. To do this we introduce a new symbol β to stand for $\log(\lambda)$, so

$$\beta=\log(\lambda),\quad \lambda=\exp(\beta).$$

In terms of β the log likelihood is

$$D\beta-Y\exp(\beta).$$

The gradient of this with respect to β is

$$D-Y\exp(\beta)$$

and the curvature is

$$-Y\exp(\beta).$$

The most likely value of $\exp(\beta)$ (which equals λ) is D/Y, so the curvature at the peak is

$$-Y\times(D/Y)=-D.$$

It follows that

$$S=\sqrt{1/D}.$$

In general, derivations such as that above can be simplified considerably by using some further elementary calculus which provides a general rule for the relationship between the values of S on the two scales. In the case of the log transformation, this rule states that multiplying the value of S on the scale of λ by the gradient of $\log(\lambda)$ at $\lambda=M$ gives the value of S on the scale of $\log(\lambda)$. The rules of calculus tell us that, at $\lambda=M$, the gradient

of the graph of $\log(\lambda)$ against λ is $1/M$. Since, on the λ scale, $M=D/Y$ and $S=\sqrt{D}/Y$, the rule tells us that the value of S for $\log(\lambda)$ is

$$\frac{\sqrt{D}}{Y}\times\frac{Y}{D}=\sqrt{\frac{1}{D}}.$$

This agrees with the expression obtained by the longer method.

A similar calculation shows that the curvature of the Bernouilli log likelihood, when plotted against $\log(\Omega)$, the log odds, is given by

$$S=\sqrt{\frac{1}{D}+\frac{1}{N-D}}.$$

Solutions to the exercises

9.1 An approximate supported range for π is given by

$$0.07\pm 1.645S$$

where $S=\sqrt{0.07\times 0.93/100}$. This gives a range from 0.028 to 0.112.
An approximate supported range for λ is given by

$$30/1018\pm 1.645S$$

where $S=\sqrt{30}/1018$. This gives a range from 21/1000 to 38/1000.

9.2 The approximate supported range for $\log(\Omega)$ is given by

$$\log(7/93)\pm 1.645S$$

where

$$S=\sqrt{\frac{1}{7}+\frac{1}{93}}=0.3919.$$

This gives a range from -3.231 to -1.942. The range for Ω is from 0.040 to 0.143, and the range for π is from 0.038 to 0.125.
The approximate supported range for $\log(\lambda)$ is given by

$$\log(30/1018)\pm 1.645S$$

where

$$S=\sqrt{1/30}=0.1826.$$

This gives a range from -3.825 to -3.224. The range for λ is from 22/1000 to 40/1000.

9.3 The error factor for Ω is

$$\exp(1.645 \times 0.3919) = 1.905.$$

The most likely value for Ω is $7/93 = 0.075$ and the range for Ω is from $0.075/1.905 = 0.040$ to $0.075 \times 1.905 = 0.143$. The range for π is from 0.038 to 0.125.
The error factor for the rate is

$$\exp(1.645 \times 0.1826) = 1.350.$$

The most likely value of the rate is 29/1000 with range from $29/1.350 = 22$ per 1000 to $29 \times 1.350 = 40$ per 1000.

10
Likelihood, probability, and confidence

The supported range for a parameter has so far been defined in terms of the cut-point -1.353 for the log likelihood ratio. Some have argued that the scientific community should accept the use of the log likelihood ratio to measure support as *axiomatic*, and that supported ranges should be reported as 1.353 unit supported ranges, or 2 unit supported ranges, with the choice of how many units of support left to the investigator. This notion has not met with widespread acceptance because of the lack of any intuitive feeling for the log likelihood ratio scale — it seems hard to justify the suggestion that a log likelihood ratio of -1 indicates that a value is supported while a log likelihood ratio of -2 indicates lack of support. Instead it is more generally felt that the reported plausible range of parameter values should be associated in some way with a *probability*. In this chapter we shall attempt to do this, and in the process we shall finally show why -1.353 was chosen as the cut-point in terms of the log likelihood ratio.

There are two radically different approaches to associating a probability with a range of parameter values, reflecting a deep philosophical division amongst mathematicians and scientists about the nature of probability. We shall start with the more orthodox view within biomedical science.

10.1 Coverage probability and confidence intervals

Our first argument is based on the frequentist interpretation of probability in terms of relative frequency of different outcomes in a very large number of repeated "experiments". With this viewpoint the statement that there is a probability of 0.9 that the parameter lies in a stated range does not make sense; there can only be one correct value of the parameter and it will either lie within the stated range or not, as the case my be. To associate a probability with the supported range we must imagine a very large number of repetitions of the study, and assume that the scientist would calculate the supported range in exactly the same way each time. Some of these ranges will include the true parameter value and some will not. The relative frequency with which the ranges include the true value is called the *coverage probability* for the range, although strictly speaking

Part II

Regression models

22
Introduction to regression models

One of the main problems discussed in Part I was how to compare two rate parameters, λ_0 and λ_1, using their ratio λ_1/λ_0. To do this the log likelihood for the parameters λ_0 and λ_1 was re-expressed in terms of λ_0 and θ, where $\theta = \lambda_1/\lambda_0$. This technique was then extended to deal with comparisons stratified by a confounding variable by making the assumption that the parameter θ was constant over strata. In this second part of the book, the technique will be further extended to deal with the joint effects of several exposures and to take account of several confounding variables.

A common theme in all these situations is a change from the original parameters to new parameters which are more relevant to the comparisons of interest. This change can be described by the equations which express the old parameters in terms of the new parameters. These equations are referred to as *regression* equations, and the statistical model is called a *regression model*. To introduce regression models we shall first express some of the comparisons discussed in Part I in these terms. We use models for the rate parameter for illustration, but everything applies equally to models for the odds parameter.

22.1 The comparison of two or more exposure groups

When comparing two rate parameters, λ_0 and λ_1, the regression equations which relate the original parameters to the new ones are

$$\lambda_0 = \lambda_0, \qquad \lambda_1 = \lambda_0\theta,$$

where the first of these simply states that the parameter λ_0 is unchanged.

When there are three groups defined by an exposure variable with three levels, corresponding (for example) to no exposure, moderate exposure, and heavy exposure, the original parameters are λ_0, λ_1, and λ_2, and there are now more ways of choosing new parameters. The most common choice is to change to

$$\lambda_0, \qquad \theta_1 = \lambda_1/\lambda_0, \qquad \theta_2 = \lambda_2/\lambda_0.$$

With this choice of parameters the moderate and heavy exposure groups

Table 22.1. A regression model to compare rates by exposure levels

	Exposure	
Age	0	1
0	λ_0^0	$\lambda_0^0\theta$
1	λ_0^1	$\lambda_0^1\theta$
2	λ_0^2	$\lambda_0^2\theta$

are compared to the unexposed group. The regression equations are now

$$\lambda_0 = \lambda_0, \qquad \lambda_1 = \lambda_0\theta_1, \qquad \lambda_2 = \lambda_0\theta_2.$$

22.2 Stratified comparisons

When the comparison between exposure groups is stratified by a confounding variable such as age the change to new parameters is first made separately for each age band; for two exposure groups the regression equations for age band t are

$$\lambda_0^t = \lambda_0^t \qquad \lambda_1^t = \lambda_0^t\theta^t.$$

The parameter θ^t is age-specific and to impose the constraint that it is constant over age bands it is set equal to the constant value θ, in each age band. The regression equations are now

$$\lambda_0^t = \lambda_0^t \qquad \lambda_1^t = \lambda_0^t\theta.$$

This choice of parameters is the same as for the proportional hazards model, introduced in Chapter 15. The model is written out in full in Table 22.1 for the case of three age bands.

Although our main interest is whether the rate parameter varies with exposure, within age bands, we might also be interested in investigating whether it varies with age, within exposure groups. The parameter θ does not help with this second comparison because it has been chosen to compare the exposure groups. When making the comparison the other way round the age bands are the groups to be compared and the exposure groups are the strata. To combine the comparison across these strata requires the assumption that the rate ratios which compare levels 1 and 2 of age with level 0 are the same in both exposure groups. This way of choosing parameters is shown in Table 22.2, where the parameters ϕ^1 and ϕ^2 are the rate ratios for age, assumed constant within each exposure group. Note that there are two parameters for age because there are three age bands being compared.

Putting these two ways of choosing parameters together gives the regression model shown in Table 22.3. The parameter λ_0^0 has now been written as λ_C, for simplicity and to emphasize that it refers to the (top left-hand)

Table 22.2. A regression model to compare rates by age bands

	Exposure	
Age	0	1
0	λ_0^0	λ_1^0
1	$\lambda_0^0\phi^1$	$\lambda_1^0\phi^1$
2	$\lambda_0^0\phi^2$	$\lambda_1^0\phi^2$

Table 22.3. A regression model for exposure and age

	Exposure	
Age	0	1
0	λ_C	$\lambda_C\theta$
1	$\lambda_C\phi^1$	$\lambda_C\theta\phi^1$
2	$\lambda_C\phi^2$	$\lambda_C\theta\phi^2$

corner of the table. Both sorts of comparison can now be made in the same analysis. It is no longer necessary to regard one variable as the exposure, and the other as a confounder used to define strata; the model treats both types of variable symmetrically. To emphasize this symmetry the term *explanatory* variable is often used to describe both exposures and confounders in regression models. Although this is useful in complex situations where there are many variables, there are also dangers. Although it makes no difference to a computer program whether an explanatory variable is an exposure or confounder it makes a great deal of difference to the person trying to interpret the results. Perhaps the single most important reason for misinterpreting the results of regression analyses is that regression models can be used without the user thinking carefully about the status of different explanatory variables. This will be discussed at greater length in Chapter 27.

Exercise 22.1. Table 22.4 shows a set of values for the rate parameters (per 1000 person-years) which satisfy exactly the model shown in Table 22.3. What are the corresponding values of $\lambda_C, \theta, \phi^1, \phi^2$?

Exercise 22.2. When the model in Table 22.3 is fitted to data it imposes the constraint that the rate ratio for exposure is the same in all age bands, and equally, that each of the two rate ratios for age is constant over both levels of exposure. Is the constraint on the rate ratios for age a new constraint, or does it automatically follow whenever the rate ratio for exposure is the same in all age bands?

Table 22.4. Parameter values (per 1000) which obey the constraints

	Exposure	
Age	0	1
0	5.0	15.0
1	12.0	36.0
2	30.0	90.0

Table 22.5. A regression model using names for parameters

	Exposure	
Age	0	1
0	Corner	Corner × Exposure(1)
1	Corner × Age(1)	Corner × Age(1) × Exposure(1)
2	Corner × Age(2)	Corner × Age(2) × Exposure(1)

22.3 Naming conventions

Using Greek letters for parameters is convenient when developing the theory but less so when applying the methods in practice. With many explanatory variables there will be many parameters and it is easy to forget which letter refers to which parameter. For this reason we shall now move to using names for parameters instead of Greek letters.

The first of the parameters in Table 22.3, λ_C, is called the Corner. The θ parameter, which is the effect of exposure controlled for age, is referred to as Exposure(1); when the exposure variable has three levels there are two effects and these are referred to as Exposure(1) and Exposure(2), and so on. When the exposure variable is given a more specific name such as Alcohol then the effects are referred to as Alcohol(1) and Alcohol(2). The ϕ parameters, which are the effects of age controlled for exposure, are referred to as Age(1) and Age(2). The model in Table 22.3 is written using names in Table 22.5.

Because writing out models in full is rather cumbersome, particularly when using names for parameters, we shall use a simple abbreviated form instead. The entries in Tables 22.3 and 22.5 refer to the right-hand sides of the regression equations; the left-hand sides are the original rate parameters which are omitted. Such a set of regression equations is abbreviated to

$$\text{Rate} = \text{Corner} \times \text{Exposure} \times \text{Age}.$$

It is important to remember that this abbreviation is not itself an equation (even though it looks like one!); it represents a set of equations and is shorthand for tables like Table 22.5. The regression model is sometimes abbreviated even further and referred to simply as a *multiplicative model* for exposure and age.

Table 22.6. Energy intake and IHD incidence rates per 1000 person-years

	Unexposed (≥ 2750 kcals)			Exposed (< 2750 kcals)			Rate
Age	Cases	P-yrs	Rate	Cases	P-yrs	Rate	ratio
40–49	4	607.9	6.58	2	311.9	6.41	0.97
50–59	5	1272.1	3.93	12	878.1	13.67	3.48
60–69	8	888.9	9.00	14	667.5	20.97	2.33

Table 22.7. Estimated values of the parameters for the IHD data

Parameter	Estimate
Corner	0.00444
Exposure(1)	×2.39
Age(1)	×1.14
Age(2)	×2.00

22.4 Estimating the parameters in a regression model

Table 22.6 shows the data from the study of ischaemic heart disease and energy intake. There are two explanatory variables, age with three levels and exposure with two. The two levels of exposure refer to energy intakes above and below 2750 kcals per day.

Although the rate ratio for exposure is rather lower in the first age band than in the other two age bands, it is based on only 6 cases, and a summary based on the assumption of a common rate ratio seems reasonable. In the new terminology this means fitting the regression model

$$\text{Rate} = \text{Corner} \times \text{Exposure} \times \text{Age}.$$

The most likely values of the parameters in this model, obtained from a computer program, are shown in Table 22.7. Note that the most likely value of the Exposure(1) parameter is the same, to two decimal places, as the Mantel–Haenszel estimate of the common rate ratio, given in Chapter 15.

Exercise 22.3. Use the most likely values of the parameters in the regression model, shown in Table 22.7, to predict the rates for the six cells in Table 22.6.

Computer programs differ in the precise details of how the output is

Table 22.8. Estimated parameters and SDs on a log scale

Parameter	Estimate (M)	SD (S)
Corner	−5.4180	0.4420
Exposure(1)	0.8697	0.3080
Age(1)	0.1290	0.4753
Age(2)	0.6920	0.4614

labelled. In particular you may see the word *variable* where we have used *parameter*, and the word *coefficient* where we have used *estimate*. We have used the term *corner* for the parameter which measures the level of response in the first age band of the unexposed group but several other terms are in widespread use, for example *constant*, *intercept*, *grand mean*, and (most cryptically of all) the number 1. We have numbered strata and exposure categories starting from zero, but some programs start numbering from one.

22.5 Gaussian approximations on the log scale

Gaussian approximations to the likelihood are used to obtain approximate confidence intervals for the parameter values. For the simple multiplicative models discussed so far the approximation is always made on the log scale, and in many programs the output is also in terms of logarithms. Table 22.8 shows the output on a log scale for the ischaemic heart data; the second column shows the most likely values (M) of the logarithms of the parameters and exponentials of these give the values on the original scale. For example,

$$\exp(0.8697) = 2.39,$$

which is the rate ratio for exposure. The third column shows the standard deviations (S) of the estimates, obtained from Gaussian approximations to the profile log likelihoods for each parameter. The standard deviation of the effect of exposure, on the log scale, is 0.3080, so the error factor for a 90% confidence interval for this parameter is $\exp(1.645 \times 0.3080) = 1.66$, and the limits are from $2.39/1.66 = 1.44$ to $2.39 \times 1.66 = 3.96$.

Exercise 22.4. Use Table 22.8 to calculate the 90% confidence limits for the first effect of age.

When the regression model is fitted on a log scale it is written in the form

$$\log(\text{Rate}) = \text{Corner} + \text{Exposure} + \text{Age}.$$

Table 22.9. A more complete description of the age effects

Parameter	Estimate	SD
Age(1)	0.1290	0.4753
Age(2)	0.6920	0.4614
Age(2) − Age(1)	0.5630	0.3229

Table 22.10. An abbreviated table for the age effects

Parameter	Estimate	SD	
Age(1)	0.1290	0.4753	0.3229
Age(2)	0.6920	0.4614	

Strictly speaking, the parameters on the right-hand side of this expression should be written as log(Corner) etc., but in practice the log on the left-hand side is enough to signal the fact that the parameter estimates will be on a log scale.

For variables with more than two categories, comparisons other than those with the first category are sometimes of interest. Taking the variable age in the ischaemic heart disease data as an example, the effect of changing from level 1 to level 2 of age is the difference between the two age effects, namely $0.6920 - 0.1290 = 0.5630$. Because the two age effects are based on some common data the standard deviation of their difference cannot be obtained from the simple formula

$$\sqrt{0.4753^2 + 0.4614^2} = 0.6624,$$

which was used in Chapter 13. To obtain the correct standard deviation we usually need to resort to a trick, such as recoding age so that the corner parameter refers to the *second* age band rather than the first. Table 22.9 shows how a fuller analysis of age effects could be reported; an option to obtain output in this form would be a useful feature not currently available in most computer programs.

An abbreviated way of conveying the same information is shown in Table 22.10. This provides the standard deviations for all three comparisons but leaves the user to do the subtraction to find the effect of changing from level 1 to level 2. The method extends naturally for factors with more than three levels; for example, a four-level factor would need a triangular array of 6 standard deviations for the six possible pairwise comparisons.

22.6 Additive models

When comparing two groups, in the first section of this chapter, the two parameters λ_0 and λ_1 were replaced by λ_0 and $\theta = \lambda_1/\lambda_0$. This change of parameters made it possible to estimate the rate ratio θ along with its standard deviation. The parameters could equally well have been changed to λ_0 and $\theta = \lambda_1 - \lambda_0$, thus making it possible to estimate the rate difference instead of the rate ratio.

The choice between the rate ratio and the rate difference is usually an empirical one, depending on which of the two is more closely constant over strata. In the early years of epidemiology, when age was often the only explanatory variable apart from exposure, methods of analysis were all based (implicitly) on multiplicative models. This is because most rates vary so much with age that the rate ratio is almost always more closely constant over age bands than the rate difference. More recently, particularly when investigating the joint effects of several exposures, epidemiologists have shown a greater interest in rate differences.

To impose the constraint that the rate difference is constant over age strata, the regression model

$$\text{Rate} = \text{Corner} + \text{Exposure} + \text{Age}$$

is fitted. This is called an *additive model* for exposure and age. Note that it is the rate and not the log rate which now appears on the left-hand side. The same likelihood techniques are used as with the additive model as with the multiplicative model, but because the estimated values of the parameters in the additive model must be restricted so that they predict positive rates, it is much harder to write foolproof programs to fit these models. We shall return to additive models in Chapter 28.

22.7 Using computer programs

There is a certain amount of specialized terminology connected with computer programs which we shall introduce briefly in this section.

VARIABLES AND RECORDS

The information collected in a study is best viewed as a rectangular table in which the columns refer to the different kinds of information collected for each subject, and the rows to the different subjects. In computer language the columns are called *variables* and the rows are called *records*. Variables such as age and observation time are called *quantitative* because they measure some quantity. Variables such as exposure group are called *categorical* because they record the category into which a subject falls. The different categories are called the *levels* of the variable. Another name for a categorical variable is *factor*. Categorical variables with only two categories (or levels) are also known as *binary* variables.

DERIVED VARIABLES

The raw data which is collected in a study may not be in exactly the right form for analysis. For example, in a follow-up study the observation time will usually be recorded as date of entry to the study and date of exit. The computer can be instructed to derive the observation time from these two dates by subtraction. Another example is where the grouped values of a quantitative variable are required in an analysis; it is then convenient to derive a new categorical variable which records the group into which each subject falls.

VARIABLE NAMES

In order to give instructions to a computer program each of the variables needs a name. These can usually be at least eight characters long and it is a good idea to make full use of this and to choose names which will mean something to you (and someone else) in a year's time.

SUMMARY TABLES

It is always important when using computer programs to keep in close touch with the data you are analyzing. The simplest way of doing this is to start by looking at tables which show the estimated rate or odds parameters for different combinations of the values of the explanatory variables. When there are two explanatory variables the table is called two-way, and so on. Three-way tables are presented as a series of two-way tables. When an explanatory variable is quantitative it will usually be necessary to group the values of the variable before using it to define a table. Only after inspecting various summary tables to get some feel for the main results should you use regression models to explore the data more fully.

FREQUENCY OR INDIVIDUAL RECORDS

Computer programs are generally able to accept either *individual records* or *frequency records* based on groups of subjects. For example, in the ischaemic heart disease study, we could use the data records for each subject, or frequency records showing the number of subjects in each combination of age band and exposure group. Entering a frequency record for 25 subjects has exactly the same effect as entering 25 identical individual records.

When an explanatory variable is quantitative its values must be grouped before frequency records can be formed, while the actual values can be used with individual records. Frequency records can be stored more compactly than individual records, and log likelihood calculations are correspondingly faster, but using frequency records requires two computer programs — one

to compute the frequency records and one to carry out the regression analysis — and communication between these programs may be inconvenient. For case-control studies the number of subjects is usually relatively small and the data are usually entered as individual records. For cohort studies there may be tens of thousands of individual records, possibly further subdivided between time-bands, so the data are usually entered as frequency records.

MISSING VALUES

Most studies contain records which have some missing values, and it is essential to have some way of indicating this to the computer program. The most convenient code for a missing value is the character *, but when a program insists on a numeric code it is best to choose some large number like 9999. When there are many variables in a study the analyses are usually on some subset of the variables, and the program will automatically include those records with complete data on the subset being used.

Solutions to the exercises

22.1 $\lambda_C = 5.0$ per 1000, $\theta = 3.0$, $\phi^1 = 2.4$, $\phi^2 = 6.0$.

22.2 It is not a new constraint. Table 22.1 shows that when the rate ratio for exposure is constant over age bands then the rate ratios for age will automatically be constant over exposure groups.

22.3 The predicted rates for the six combinations of age and exposure are

Age	Unexposed	Exposed
40 – 49	4.44	10.61
50 – 59	5.06	12.10
60 – 69	8.88	21.22

22.4 The effect of age level 1 is $\exp(0.1290) = 1.14$. The 90% confidence interval for this effect is

$$1.14 \overset{\times}{\div} \exp(1.645 \times 0.4753)$$

which is from 0.52 to 2.49.

23
Poisson and logistic regression

In principle the way a computer program goes about fitting a regression model is simple. First the likelihood is specified in terms of the original set of parameters. Then it is expressed in terms of the new parameters using the regression equations, and finally most likely values of these new parameters are found. In studies of event data the two most important likelihoods are Poisson and Bernouilli, and the combinations of these with regression models are called *Poisson* and *logistic* regression respectively. Gaussian regression is the combination of the Gaussian likelihood with regression models and will be discussed in Chapter 34.

23.1 Poisson regression

When a time scale, such as age, is divided into bands and included in a regression model, the observation time for each subject must be split between the bands as described in Chapter 6. This is illustrated in Fig. 23.1, where a single observation time ending in failure (the top line) has been split into three parts, the last of which ends in failure. These parts can then be used to make up frequency records containing the number of failures and the observation time, as was done for the ischaemic heart disease data in Table 23.1, or they can be analysed as though they were individual records.

If they are to be analysed as though they were individual records then each of these new records must contain variables which describe which time band is being referred to, how much observation time is spent in the time band, and whether or not a failure occurs in the time band. Values of

Table 23.1. The IHD data as frequency records

Cases	Person-years	Age	Exposure
4	607.9	0	0
2	311.9	0	1
5	1272.1	1	0
12	878.1	1	1
8	888.9	2	0
14	667.5	2	1