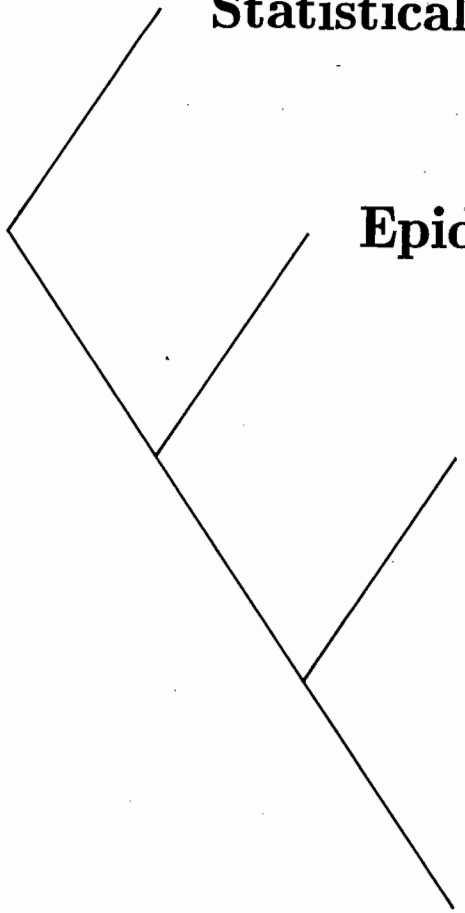
Statistical Models in Epidemiology



David Clayton
Medical Research Council, Cambridge

and

Michael Hills
London School of Hygiene and Tropical Medicine

OXFORD • NEW YORK • TOKYO
OXFORD UNIVERSITY PRESS

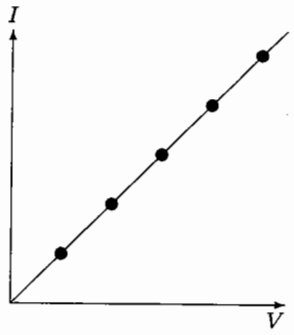


Fig. 1.1. A deterministic model: Ohm's law.

(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,

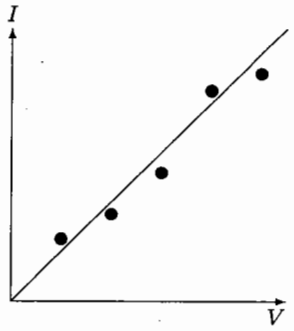


Fig. 1.2. Experimental/observational errors.

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).

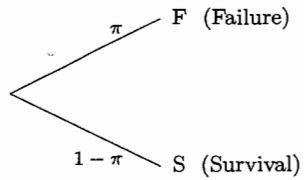


Fig. 1.3. The binary probability model.

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

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

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

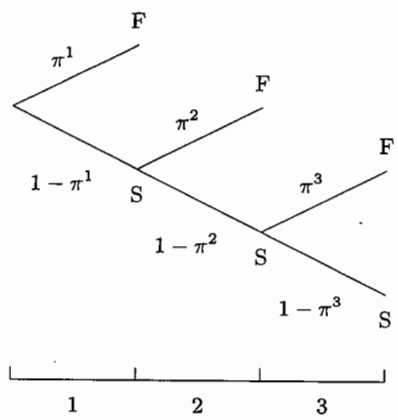


Fig. 4.1. A sequence of binary probability models.

all conditional probabilities — conditional on not having failed before the start of the year in question. The reason the probabilities are written with superscripts is that we have adopted the convention that a superscript is used to index *time*, and a subscript is used to index subjects or groups of subjects. It is important to distinguish these two situations, and using subscripts for both can be confusing.*

Suppose, for illustration, that the probability of failure is 0.3 in the first year; 0.2 in the second year, given the subject survives the first year without failure; and 0.1 in the third year, given the subject survives the first two years without failure. These illustrative values for the three conditional probabilities are shown on the conditional probability tree in Fig.4.2.

In this tree, the four final outcomes listed above correspond to the tips of the tree, and their probabilities can be calculated by multiplying conditional probabilities along the branches of the tree in the usual way. For example, the probability of the second outcome is made up from the probability that the subject survives the first year (0.7), multiplied by the probability that the subject fails during the second year (0.2). Using this rule, the four possible outcomes for any subject occur with probabilities:

$$0.3$$
$$0.7 \times 0.2$$
$$0.7 \times 0.8 \times 0.1$$
$$0.7 \times 0.8 \times 0.9$$

*Note that π^2 does not refer to $\pi \times \pi$. To avoid confusion we shall always use brackets when taking powers; for example, the square of π will be written $(\pi)^2$.

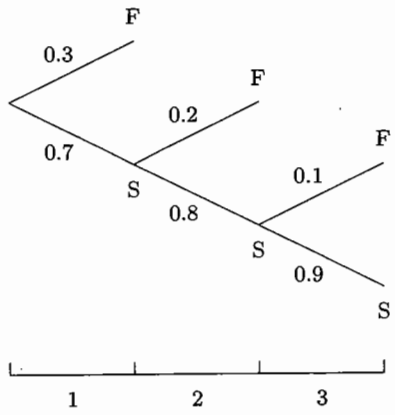


Fig. 4.2. Illustrative values for the conditional probabilities.

These probabilities work out to be 0.3, 0.14, 0.056, and 0.504, and these add to 1, as they should, since there are no other possible outcomes. The probability of failing at *some* stage is

$$0.3 + 0.14 + 0.056 = 0.496.$$

More conveniently this probability can be found by subtracting from 1 the probability of surviving the three years without failing, giving

$$1 - 0.504 = 0.496.$$

The probabilities of surviving one, two, and three years without failing are called the *cumulative survival probabilities* for the cohort. They are calculated by multiplying the conditional probabilities of surviving each year, and in this case are:

$$0.7$$
$$0.7 \times 0.8$$
$$0.7 \times 0.8 \times 0.9.$$

which work out to be 0.7, 0.56, and 0.504.

Exercise 4.1. In a three-year follow-up study the conditional probabilities of failure during the first, second, and third years are 0.05, 0.09, and 0.12 respectively. Draw a probability tree for the possible outcomes for a new subject, and label the branches of the tree with the appropriate conditional probabilities. Calculate the probability of each of the outcomes, and the probabilities of surviving

5
Rates

We have shown how, by splitting the follow-up period into small enough bands, the importance of arbitrary assumptions about when the losses occur can be minimized. We now follow this argument to its logical conclusion and divide the follow-up into infinitely small time bands.

5.1 The probability rate

As the bands get shorter, the conditional probability that a subject fails during any one band gets smaller. When a band shrinks towards a single moment of time, the conditional probability of failure during the band shrinks towards zero, but the conditional probability of failure *per unit time* converges to a quantity called the *probability rate*. This quantity is sometimes called the *instantaneous* probability rate to emphasize the fact that it refers to a moment in time. Other names are *hazard* rate and *force of mortality*.

The probability rate refers to an *individual subject*. This is counter-intuitive to many epidemiologists, who think of a rate as an empirical summary of the frequency of failures in a group observed over time. We show in the next section that such a summary is, in fact, the most likely value of the common probability rate for the subjects in the group. It is general practice in epidemiology to refer to both the probability rate and its estimated value as the rate, even though this leads to many logical absurdities. We have tried to keep as close as possible to this tradition, while avoiding the logical contradictions, by referring to the probability rate as the rate parameter and its estimated value as the observed rate.

5.2 Estimating the rate parameter

Even though the rate parameter refers to a single individual it is not possible to estimate its value from the experience of that individual. The estimate must be based on the experience of a group of subjects assumed to have the same rate. Similarly, even though the rate parameter refers to a single moment of time, its estimated value is usually based on a period of follow-up over which the rate is assumed to be constant. The estimated rate for this period then refers to the constant value which the rate parameter

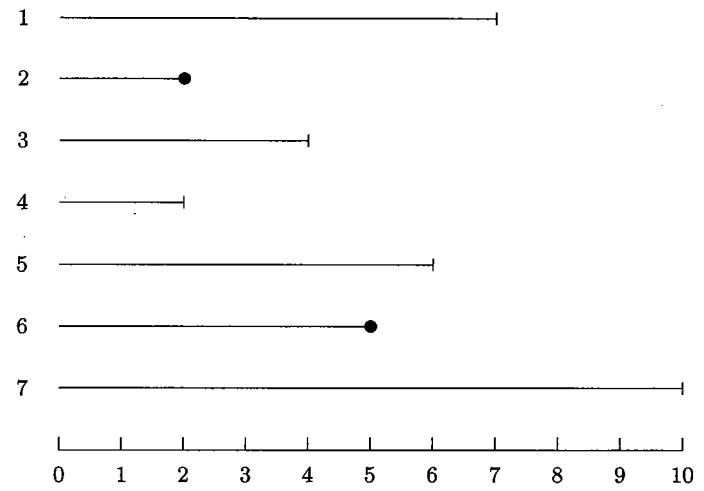


Fig. 5.1. The follow-up experience of 7 subjects.

takes at all time points during the period.

The rate parameter over a follow-up period is estimated by dividing the period into a number of small time bands of equal length and estimating the *common* probability of failure for each of the bands. This is divided by the length of a band to get the rate per unit time. The process is illustrated using the follow-up experience of 7 subjects shown in Fig. 5.1, in which the follow-up experience of the subjects is shown as lines which end when follow-up ends. The lines for those subjects who fail end with a •, while those whose observation time is censored end with a short bar. The follow-up period has been divided into 10 short bands and for the present we shall assume that follow-up always stops at the end of a short band. From the figure we see that the follow-up of subject 1 stops after 7 bands due to censoring. For subject 6 the follow-up stops after 5 bands when the subject fails, and so on.

Exercise 5.1. How many observations of one subject through one time band are observed? How many of these ended in failure?

Assuming that the rate parameter is constant over the follow-up period, the conditional probability of failure is the same for all bands and its most likely value is 2/36. The most likely value of the corresponding rate parameter is 2/36 divided by the length of the bands. Suppose for illustration that each band has length 0.05 years. The most likely value of the rate parameter is

then

$$\frac{2}{(36 \times 0.05)} = 1.11 \text{ per year.}$$

Note that 36×0.05, which equals 1.8 years, is the total observation time for the 7 subjects.

Now suppose that five times as many bands are used, so that each is 0.01 years in length. The most likely value of the probability of failure for these bands is 2/180, but the most likely value of the corresponding rate stays the same because there are now 180 bands of length 0.01 years and 180×0.01 is the same as 36×0.05, both being equal to the total observation time, added over subjects. In general, then, as the bands shrink to zero, the most likely value of the rate parameter is

$$\frac{\text{Total number of failures}}{\text{Total observation time}}.$$

Note that assumption that events occur at the end of bands is automatically true when the bands shrink to zero. This mathematical device of dividing the time scale into shorter and shorter bands is used frequently in this book, and we have found it useful to introduce the term *clicks* to describe these very short time bands.

Time can be measured in any convenient units, so that a rate of 1.11 per year is the same as a rate of 11.1 per 10 years, and so on. The total observation time added over subjects is known in epidemiology as the *person-time* of observation and is most commonly expressed as person-years. Because of the way they are calculated, estimates of rates are often given the units *per person-year* or *per 1000 person-years.*

The use of the general formula for the estimated value of a rate is now illustrated using data from a computer simulation of 30 subjects who are liable to only one disease (the failure) and the follow-up is indefinitely long, so that eventually all subjects develop the disease. The only variable in the outcome is how long it takes for the disease to develop, and these times are shown in Table 5.1.

Exercise 5.2. Using the time interval from the start of the study to the moment when the last subject develops the disease, find the total observation time for the 30 subjects and hence estimate the rate for this interval. Give your answer per 10^3 person-years as well.

Exercise 5.3. The previous exercise is rather unrealistic. Real follow-up studies are of limited duration and not all of the subjects will fail during the study period. Estimate the rate from a study in which the same subjects are observed only for the first five years.

Table 5.1. Time until the disease develops, for 30 subjects

Subject	Years	Subject	Years
1	19.6	16	0.6
2	10.8	17	2.1
3	14.1	18	0.8
4	3.5	19	8.9
5	4.8	20	11.6
6	4.6	21	1.3
7	12.2	22	3.4
8	14.0	23	15.3
9	3.8	24	8.5
10	12.6	25	21.5
11	12.8	26	8.3
12	12.1	27	0.4
13	4.7	28	36.5
14	3.2	29	1.1
15	7.3	30	1.5

5.3 The likelihood for a rate

The argument of the last section, although leading to the most likely value of the rate parameter, does not allow us to explore the support for other values. In this section we shall obtain a formula for the likelihood for a rate parameter.

Consider a more general example in which D failures are observed for a total of N clicks of time, each of duration h years, where h is very small and N is very large. The total observation time in years is $Y = Nh$. Let π be the conditional probability of failure during a click. Then the likelihood for π is

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

Let the corresponding rate parameter be λ, where, because h is small,

$$\lambda = \pi/h.$$

The likelihood for λ follows by replacing π by λh, and is

$$(\lambda h)^D(1-\lambda h)^{N-D}.$$

The log likelihood for λ is therefore

$$D\log(\lambda) + D\log(h) + (N-D)\log(1-\lambda h).$$