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

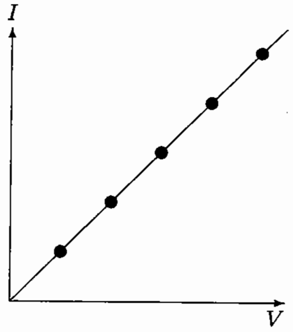

Fig. 1.1. A deterministic model: Ohm's law.

(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, 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.

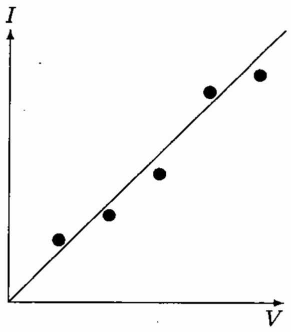

Fig. 1.2. Experimental/observational errors.

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).

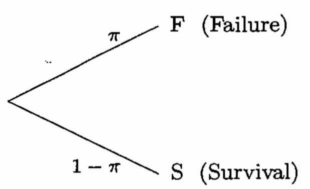

Fig. 1.3. The binary probability model.

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.