2
Conditional probability models

In this chapter we introduce the idea of *conditional probability*, which allows us to extend the binary model so that the probability of failure can depend on earlier events. The natural way of thinking about conditional probabilities is in terms of a tree diagram. These diagrams are used extensively throughout the book.

2.1 Conditional probability

Suppose a binary probability model assigns a probability to a subject's death during some future time period. It may be that this prediction would be better if we knew the subject's smoking habits. This would be the case if the probability of death for a smoker were 0.015 but only 0.005 for a non-smoker. These probabilities are called *conditional* probabilities; they are the probabilities of death conditional on being a smoker and a non-smoker respectively. Epidemiology is mainly concerned with conditional probability models that relate occurrence of some disease event, which we call failure, to events which precede it. These include potential causes, which we call *exposures.*

When subjects are classified as either exposed (E+) or not exposed (E−), the conditional probability model can be represented as a tree with 6 branches. The first two branches refer to E+ and E−; then there are two referring to failure and survival if the subject is exposed, and two referring to failure and survival if the subject is not exposed. An example is shown in Fig. 2.1. The tips of the tree correspond to the four possible combinations of exposure and outcome for any subject.

The probabilities on the first two branches of the tree refer to the probability that a subject is exposed and the probability that a subject is not exposed. Using the smoking example we have taken these to be 0.4 and 0.6. The probabilities in the next two pairs of branches are conditional probabilities. These are 0.015 (F) and 0.985 (S) if a subject is exposed (smokes), and 0.005 (F) and 0.995 (S) if a subject is not exposed (does not smoke).

The probability of any combination of exposure and outcome is obtained by multiplying the probabilities along the branches leading to the tip which corresponds to that combination. For example, the probability that a subject is exposed and fails is

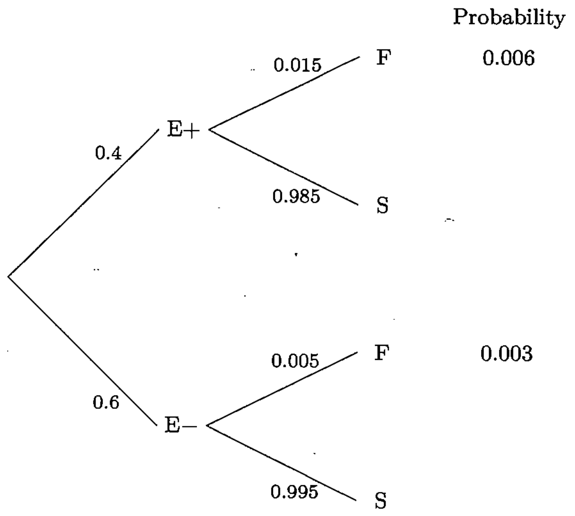

Fig. 2.1. A conditional probability tree.

$$0.4 \times 0.015 = 0.006,$$

and the probability that a subject is not exposed and fails is

$$0.6 \times 0.005 = 0.003.$$

This is called the multiplicative rule.

Exercise 2.1. Calculate the probabilities for each of the remaining 2 possibilities. What is the overall probability of failure regardless of exposure?

This overall probability is usually called the *marginal* probability of failure.

STATISTICAL DEPENDENCE AND INDEPENDENCE

Fig. 2.1 illustrates a model in which the probability of failure differs according to whether an individual was exposed or not. In this case, exposure and failure are said to be *statistically dependent.* If the probability of failure is the same, whether or not the subject is exposed, then exposure and failure are said to be *statistically independent.*

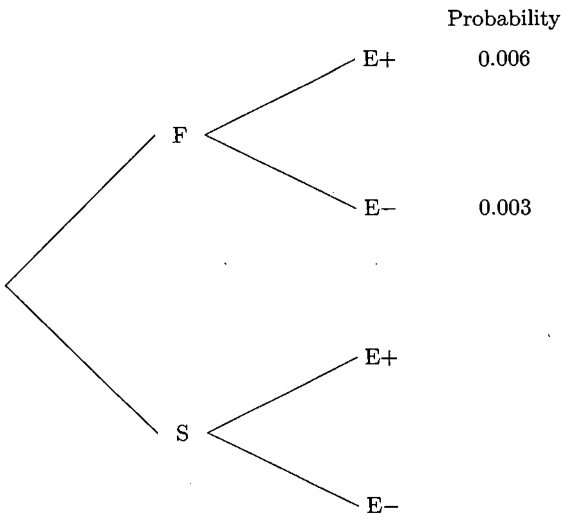

Fig. 2.2. Predicting exposure from the outcome.

2.2 Changing the conditioning: Bayes' rule

The additive and multiplicative rules are the basic building blocks of probability models. A simple application of these rules allows us to change the direction of prediction so that, for example, a model for the probability of failure given exposure can be transformed into a model for the probability of exposure given failure.

We shall demonstrate this by using the tree in Fig. 2.1, where the first level of branching refers to exposure and the second to outcome. This is turned round in Fig. 2.2, so that the first level of branching now refers to outcome and the second to exposure. The probabilities of the different combinations of exposure and outcome are the same whichever way the tree is written; our problem is to fill in the probabilities on the branches of this new tree.

Working backwards from the tips of the tree, the probability of failure regardless of exposure is $0.006 + 0.003 = 0.009$. This is the probability for the first branch of the tree to F. Since the probability corresponding to any tip of the tree is obtained by multiplying the probabilities in the branches that lead to the tip, it follows that the probability in the branch from F to E+, for example, is $0.006/0.009 = 0.667$. This is the conditional probability of being exposed given the outcome was failure. This process of reversing the order of the conditioning is called Bayes' rule, after Thomas Bayes.

Exercise 2.2. Calculate the remaining conditional probabilities.

The following exercise, inspired by problems in screening, demonstrates one of the many uses of Bayes' rule.

Exercise 2.3. A screening test has a probability of 0.90 of being positive in true cases of a disease (the *sensitivity*) and a probability of 0.995 of being negative in people without the disease (the *specificity*). The prevalence of the disease is 0.001 so before carrying out the test, the probability that a person has the disease is 0.001.
(a) Draw a probability tree in which the first level of branching refers to having the disease or not, and the second level to being positive or negative on the screening test. Fill in the probabilities for each of the branches and calculate the probabilities for the four possible combinations of disease and test.
(b) Draw the tree the other way, so that the first level of branching refers to being positive or negative on the screening test and the second level to having the disease or not. Fill in the probabilities for the branches of this tree. What is the probability of a person having the disease given that they have a positive test result? (This is called the *positive predictive value.*)

2.3 An example from genetics ☆

Our next exercises illustrate a problem in genetic epidemiology. For a specified genetic system (such as the HLA system), each person's *genotype* consists of two *haplotypes*,* one inherited from the mother and one from the father. If a mother has haplotypes (a,b), then one of these is passed to the offspring with probability 0.5. Likewise for a father's haplotypes, (c,d) say. Fig. 2.3 shows the probability tree for the genotype of the offspring. The presence of haplotype (a) carries a probability of disease of 0.05 while, in its absence, the probability is only 0.01.

Exercise 2.4. Work out the probabilities for the four tips of the probability tree which end in disease (F). Hence work out the probabilities of the four possible genotypes conditional on the fact that the offspring is affected by disease (Fig. 2.4).

Exercise 2.5. In practice the probabilities of disease conditional upon genotype are not known constants but unknown parameters. Repeat the previous exercise *algebraically*, replacing the probabilities 0.01 and 0.05 by π and $\theta\pi$ respectively. How are the conditional probabilities changed if the subject's father has genotype (c,c)?

The parameter θ, described in Exercise 2.5, is a *risk ratio*,

$$\theta = \frac{\text{Risk of disease if haplotype (a) present}}{\text{Risk of disease if haplotype (a) absent}}.$$

*The word haplotype refers to a group of genetic loci which are closely linked and therefore inherited together.

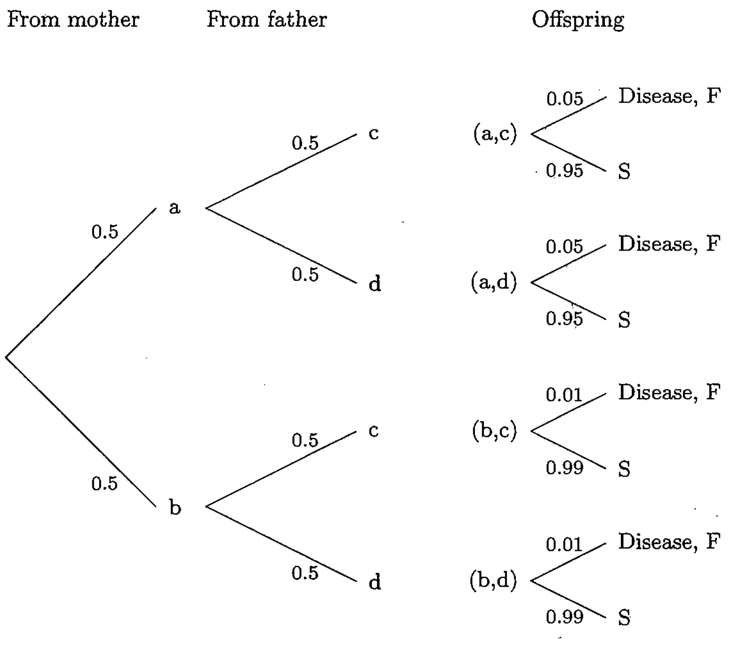

Fig. 2.3. Disease conditional upon inheritance.

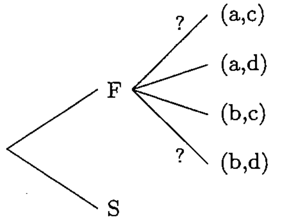

Fig. 2.4. Inheritance conditional upon disease.

It measures the strength of statistical dependence (or *association*) between the presence of haplotype (a) and occurrence of disease. The above exercise shows that the conditional probability of genotype given the presence of disease and parental genotypes depends only on this risk ratio.

Solutions to the exercises

2.1

$$\begin{aligned}\Pr(\text{E+ and S}) &= 0.4 \times 0.985 = 0.394\\ \Pr(\text{E− and S}) &= 0.6 \times 0.995 = 0.597\end{aligned}$$

The overall probability of failure is $0.006 + 0.003 = 0.009$.

2.2 See Fig. 2.5. The conditional probabilities of E+ and E− given survival are

$$\frac{0.394}{0.991} = 0.3976, \qquad \frac{0.597}{0.991} = 0.6024.$$

2.3 (a) See Fig. 2.6.
(b) See Fig. 2.7. The probability of disease given a positive test result is

$$\frac{0.0009}{0.005895} = 0.1527.$$

Note that this is much lower than 0.90, the sensitivity of the test. The remaining conditional probabilities are calculated in a similar manner.

2.4 The probabilities for each of the four tips are obtained by multiplying along the branches of the tree. The sum of the four probabilities is 0.0300. The *conditional* probabilities sum to 1.0.

Genotype	Disease	Probability	Conditional prob.
(a,c)	F	$0.5 \times 0.5 \times 0.05 = 0.0125$	$0.0125/0.03 = 0.417$
(a,d)	F	$0.5 \times 0.5 \times 0.05 = 0.0125$	0.417
(b,c)	F	$0.5 \times 0.5 \times 0.01 = 0.0025$	$0.0025/0.03 = 0.083$
(b,d)	F	$0.5 \times 0.5 \times 0.01 = 0.0025$	0.083
Total		0.0300	1.0

2.5 Repeating the above calculations algebraically yields:

Genotype	Disease	Probability	Conditional Prob.
(a,c)	F	$0.5 \times 0.5 \times \theta\pi = 0.25\theta\pi$	$\theta/(2\theta + 2)$
(a,d)	F	$0.5 \times 0.5 \times \theta\pi = 0.25\theta\pi$	$\theta/(2\theta + 2)$
(b,c)	F	$0.5 \times 0.5 \times \pi = 0.25\pi$	$1/(2\theta + 2)$
(b,d)	F	$0.5 \times 0.5 \times \pi = 0.25\pi$	$1/(2\theta + 2)$
Total		$0.25\pi(2\theta + 2)$	1.0

If the father has genotype (c,c) then he can only pass on (c) and the possible genotypes of offspring are (a,c) and (b,c). Prior to observation of disease presence, these both have probabilities 0.5. Thus, for a subject known to have disease, we have

Genotype	Disease	Probability	Conditional Prob.
(a,c)	F	$0.5 \times \theta\pi = 0.5\theta\pi$	$\theta/(\theta+1)$
(b,c)	F	$0.5 \times \pi = 0.5\pi$	$1/(\theta+1)$
Total		$0.5\pi(\theta+1)$	1.0

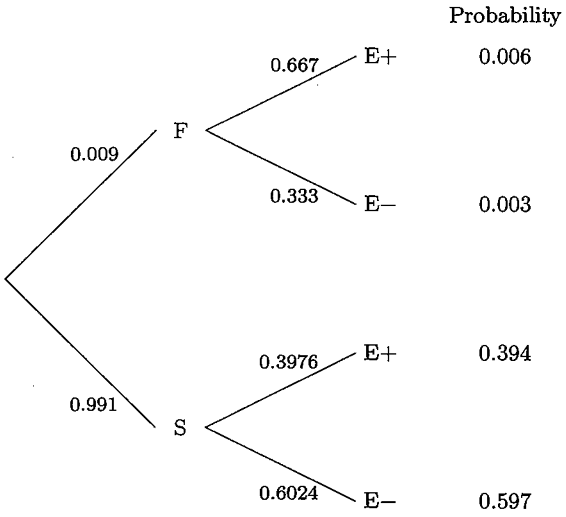

Fig. 2.5. Probability tree for exposure given outcome.

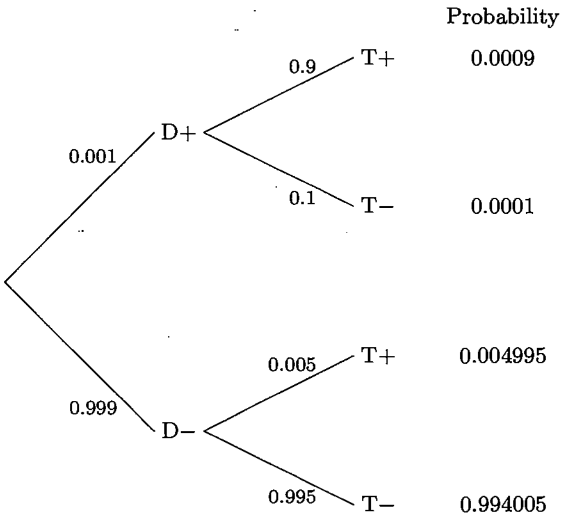

Fig. 2.6. Test results, T, given disease status, D.

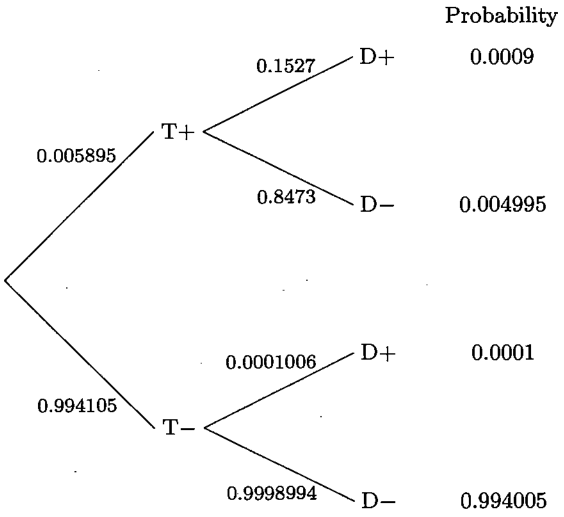

Fig. 2.7. Disease status given test results.