learning.

We begin by considering distributions for discrete variables before exploring the Gaussian distribution for continuous variables. These are specific examples of *parametric* distributions, so called because they are governed by a relatively small number of adjustable parameters, such as the mean and variance of a Gaussian. To apply such models to the problem of density estimation, we need a procedure for determining suitable values for the parameters, given an observed data set, and our main focus will be on maximizing the likelihood function. In this chapter, we will assume that the data observations are independent and identically distributed (i.i.d.), whereas in future chapters we will explore more complex scenarios involving *structured data* where this assumption no longer holds.

One limitation of the parametric approach is that it assumes a specific functional form for the distribution, which may turn out to be inappropriate for a particular application. An alternative approach is given by *nonparametric* density estimation methods in which the form of the distribution typically depends on the size of the data set. Such models still contain parameters, but these control the model complexity rather than the form of the distribution. We end this chapter by briefly considering three nonparametric methods based respectively on histograms, nearest neighbours, and kernels. A major limitation of nonparametric techniques such as these is that they involve storing all the training data. In other words, the number of parameters grows with the size of the data set, so that the method become very inefficient for large data sets. Deep learning combines the efficiency of parametric models with the generality of nonparametric methods by considering flexible distributions based on neural networks having a large, but fixed, number of parameters.

3.1. Discrete Variables

We begin by considering simple distributions for discrete variables, starting with binary variables and then moving on to multi-state variables.

3.1.1 Bernoulli distribution

Consider a single binary random variable $x \in \{0, 1\}$. For example, x might describe the outcome of flipping a coin, with $x = 1$ representing 'heads' and $x = 0$ representing 'tails'. If this were a damaged coin, such as the one shown in Figure 2.2, the probability of landing heads is not necessarily the same as that of landing tails. The probability of $x = 1$ will be denoted by the parameter μ so that

$$p(x = 1|\mu) = \mu \tag{3.1}$$

where $0 \leqslant \mu \leqslant 1$, from which it follows that $p(x = 0|\mu) = 1 - \mu$. The probability distribution over x can therefore be written in the form

$$\text{Bern}(x|\mu) = \mu^x(1 - \mu)^{1-x}, \tag{3.2}$$

Exercise 3.1 which is known as the *Bernoulli* distribution. It is easily verified that this distribution