Exercise 2.12   Also, it is straightforward to show that the Gaussian is normalized, so that

$$\int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu, \sigma^2\right) \, \mathrm{d}x = 1. \tag{2.51}$$

Thus, (2.49) satisfies the two requirements for a valid probability density.

2.3.1 Mean and variance

We can readily find expectations of functions of x under the Gaussian distribu-
Exercise 2.13   tion. In particular, the average value of x is given by

$$\mathbb{E}[x] = \int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu, \sigma^2\right) x \, \mathrm{d}x = \mu. \tag{2.52}$$

Because the parameter μ represents the average value of x under the distribution, it
is referred to as the mean. The integral in (2.52) is known as the *first-order moment*
of the distribution because it is the expectation of x raised to the power one. We can
similarly evaluate the second-order moment given by

$$\mathbb{E}[x^2] = \int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu, \sigma^2\right) x^2 \, \mathrm{d}x = \mu^2 + \sigma^2. \tag{2.53}$$

From (2.52) and (2.53), it follows that the variance of x is given by

$$\mathrm{var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2 = \sigma^2 \tag{2.54}$$

and hence σ^2 is referred to as the variance parameter. The maximum of a distribution
Exercise 2.14   is known as its mode. For a Gaussian, the mode coincides with the mean.

2.3.2 Likelihood function

Suppose that we have a data set of observations represented as a row vector
$\mathbf{x} = (x_1, \ldots, x_N)$, representing N observations of the scalar variable x. Note that
we are using the typeface $\mathbf{x}$ to distinguish this from a single observation of a D-
dimensional vector-valued variable, which we represent by a column vector $\mathbf{x} = (x_1, \ldots, x_D)^{\mathrm{T}}$. We will suppose that the observations are drawn independently from
a Gaussian distribution whose mean μ and variance σ^2 are unknown, and we would
like to determine these parameters from the data set. The problem of estimating a
distribution, given a finite set of observations, is known as *density estimation*. It
should be emphasized that the problem of density estimation is fundamentally ill-
posed, because there are infinitely many probability distributions that could have
given rise to the observed finite data set. Indeed, any distribution $p(\mathbf{x})$ that is non-
zero at each of the data points $\mathbf{x}_1, \ldots, \mathbf{x}_N$ is a potential candidate. Here we constrain
the space of distributions to be Gaussians, which leads to a well-defined solution.

Data points that are drawn independently from the same distribution are said to
be *independent and identically distributed*, which is often abbreviated to i.i.d. or
IID. We have seen that the joint probability of two independent events is given by
the product of the marginal probabilities for each event separately. Because our data