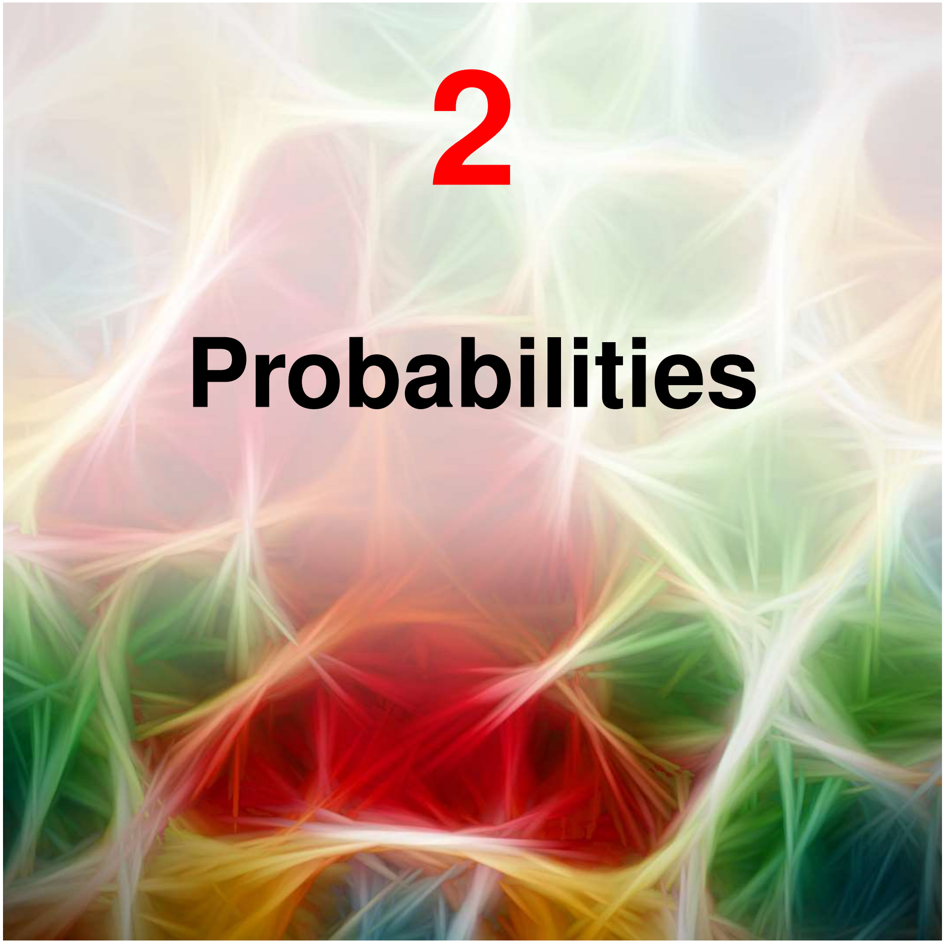
2 Probabilities



In almost every application of machine learning we have to deal with uncertainty. For example, a system that classifies images of skin lesions as benign or malignant can never in practice achieve perfect accuracy. We can distinguish between two kinds of uncertainty. The first is *epistemic uncertainty* (derived from the Greek word episteme meaning knowledge), sometimes called *systematic uncertainty*. It arises because we only get to see data sets of finite size. As we observe more data, for instance more examples of benign and malignant skin lesion images, we are better able to predict the class of a new example. However, even with an infinitely large data set, we would still not be able to achieve perfect accuracy due to the second kind of uncertainty known as *aleatoric uncertainty*, also called *intrinsic* or *stochastic* uncertainty, or sometimes simply called *noise*. Generally speaking, the noise arises because we are able to observe only partial information about the world, and therefore, one way to reduce this source of uncertainty is to gather different kinds of data. This is illustrated

C. M. Bishop, H. Bishop, *Deep Learning*, https://doi.org/10.1007/978-3-031-45468-4_2

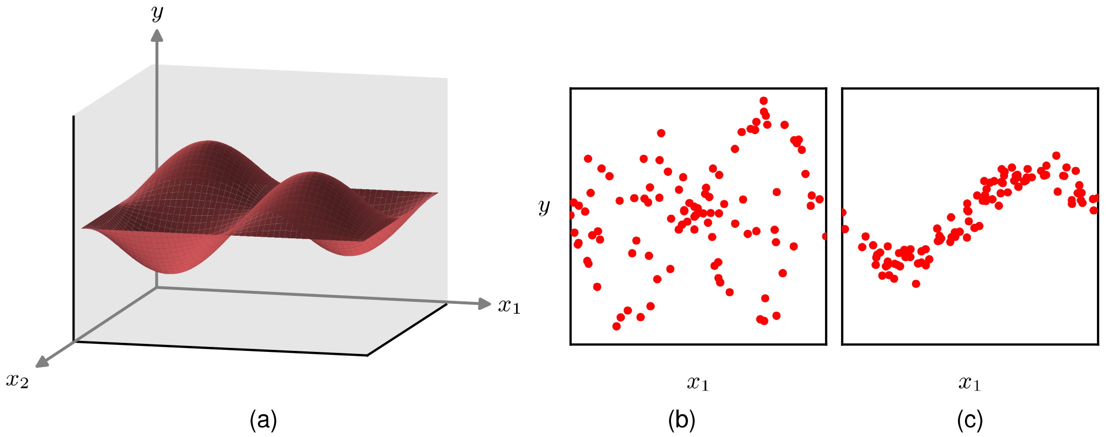


Figure 2.1 An extension of the simple sine curve regression problem to two dimensions. (a) A plot of the function $y(x_1, x_2) = \sin(2\pi x_1)\sin(2\pi x_2)$. Data is generated by selecting values for x_1 and x_2, computing the corresponding value of $y(x_1, x_2)$, and then adding Gaussian noise. (b) Plot of 100 data points in which x_2 is unobserved showing high levels of noise. (c) Plot of 100 data points in which x_2 is fixed to the value $x_2 = \frac{\pi}{2}$, simulating the effect of being able to measure x_2 as well as x_1, showing much lower levels of noise.

Section 1.2 using an extension of the sine curve example to two dimensions in Figure 2.1.

As a practical example of this, a biopsy sample of the skin lesion is much more informative than the image alone and might greatly improve the accuracy with which we can determine if a new lesion is malignant. Given both the image and the biopsy data, the intrinsic uncertainty might be very small, and by collecting a large training data set, we may be able to reduce the systematic uncertainty to a low level and thereby make predictions of the class of the lesion with high accuracy.

Both kinds of uncertainty can be handled using the framework of *probability theory*, which provides a consistent paradigm for the quantification and manipulation of uncertainty and therefore forms one of the central foundations for machine learning. We will see that probabilities are governed by two simple formulae known as the *sum rule* and the *product rule*. When coupled with *decision theory*, these rules allow us, at least in principle, to make optimal predictions given all the information available to us, even though that information may be incomplete or ambiguous.

Section 2.1

Section 5.2

The concept of probability is often introduced in terms of frequencies of repeatable events. Consider, for example, the bent coin shown in Figure 2.2, and suppose that the shape of the coin is such that if it is flipped a large number of times, it lands concave side up 60% of the time, and therefore lands convex side up 40% of the time. We say that the *probability* of landing concave side up is 60% or 0.6. Strictly, the probability is defined in the limit of an infinite number of 'trials' or coin flips in this case. Because the coin must land either concave side up or convex side up, these probabilities add to 100% or 1.0. This definition of probability in terms of the frequency of repeatable events is the basis for the *frequentist* view of statistics.

Now suppose that, although we know that the probability that the coin will land concave side up is 0.6, we are not allowed to look at the coin itself and we do not

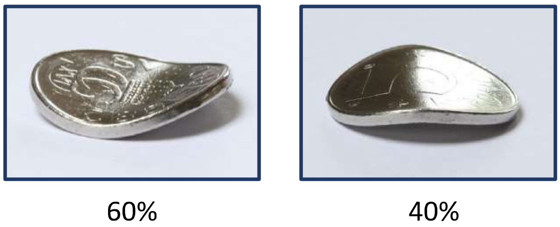


Figure 2.2 Probability can be viewed either as a frequency associated with a repeatable event or as a quantification of uncertainty. A bent coin can be used to illustrate the difference, as discussed in the text.

know which side is heads and which is tails. If asked to take a bet on whether the coin will land heads or tails when flipped, then symmetry suggests that our bet should be based on the assumption that the probability of seeing heads is 0.5, and indeed a more careful analysis shows that, in the absence of any additional information, this is indeed the rational choice. Here we are using probabilities in a more general sense than simply the frequency of events. Whether the convex side of the coin is heads or tails is not itself a repeatable event, it is simply unknown. The use of probability as a quantification of uncertainty is the *Bayesian* perspective and is more general in that it includes frequentist probability as a special case. We can learn about which side of the coin is heads if we are given results from a sequence of coin flips by making use of Bayesian reasoning. The more results we observe, the lower our uncertainty as to which side of the coin is which.

Section 2.6

Exercise 2.40

Having introduced the concept of probability informally, we turn now to a more detailed exploration of probabilities and discuss how to use them quantitatively. Concepts developed in the remainder of this chapter will form a core foundation for many of the topics discussed throughout the book.

2.1. The Rules of Probability

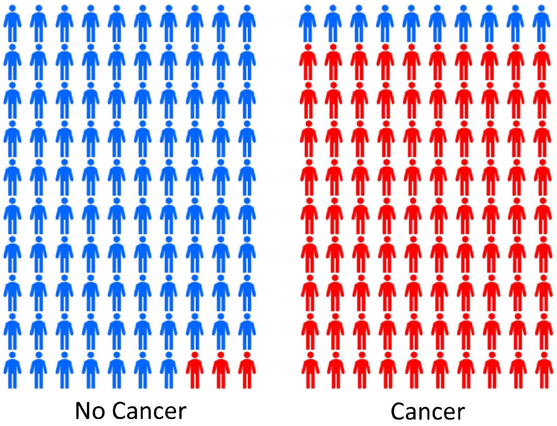
In this section we will derive two simple rules that govern the behaviour of probabilities. However, in spite of their apparent simplicity, these rules will prove to be very powerful and widely applicable. We will motivate the rules of probability by first introducing a simple example.

2.1.1 A medical screening example

Consider the problem of screening a population in order to provide early detection of cancer, and let us suppose that 1% of the population actually have cancer. Ideally our test for cancer would give a positive result for anyone who has cancer and a negative result for anyone who does not. However, tests are not perfect, so we will suppose that when the test is given to people who are free of cancer, 3% of them will test positive. These are known as *false positives*. Similarly, when the test is given to people who do have cancer, 10% of them will test negative. These are called *false negatives*. The various error rates are illustrated in Figure 2.3.

Given this information, we might ask the following questions: (1) ‘If we screen the population, what is the probability that someone will test positive?’, (2) ‘If some-

Figure 2.3 Illustration of the accuracy of a cancer test. Out of every hundred people taking the test who do not have cancer, shown on the left, on average three will test positive. For those who have cancer, shown on the right, out of every hundred people taking the test, on average 90 will test positive.



one receives a positive test result, what is the probability that they actually have cancer?'. We could answer such questions by working through the cancer screening case in detail. Instead, however, we will pause our discussion of this specific example and first derive the general rules of probability, known as the *sum rule of probability* and the *product rule*. We will then illustrate the use of these rules by answering our two questions.

2.1.2 The sum and product rules

To derive the rules of probability, consider the slightly more general example shown in Figure 2.4 involving two variables X and Y. In our cancer example, X could represent the presence or absence of cancer, and Y could be a variable denoting the outcome of the test. Because the values of these variables can vary from one person to another in a way that is generally unknown, they are called *random variables* or *stochastic variables*. We will suppose that X can take any of the values x_i where $i = 1, \ldots, L$ and that Y can take the values y_j where $j = 1, \ldots, M$. Consider a total of N trials in which we sample both of the variables X and Y, and let the number of such trials in which $X = x_i$ and $Y = y_j$ be n_{ij}. Also, let the number of trials in which X takes the value x_i (irrespective of the value that Y takes) be denoted by c_i, and similarly let the number of trials in which Y takes the value y_j be denoted by r_j.

The probability that X will take the value x_i and Y will take the value y_j is written $p(X = x_i, Y = y_j)$ and is called the *joint* probability of $X = x_i$ and $Y = y_j$. It is given by the number of points falling in the cell i,j as a fraction of the total number of points, and hence

$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N}. \tag{2.1}$$

Here we are implicitly considering the limit $N \to \infty$. Similarly, the probability that X takes the value x_i irrespective of the value of Y is written as $p(X = x_i)$ and is

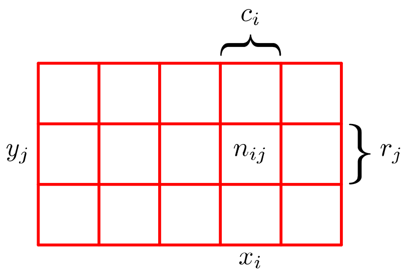


Figure 2.4 We can derive the sum and product rules of probability by considering a random variable X, which takes the values $\{x_i\}$ where $i = 1, \ldots, L$, and a second random variable Y, which takes the values $\{y_j\}$ where $j = 1, \ldots, M$. In this illustration, we have $L = 5$ and $M = 3$. If we consider the total number N of instances of these variables, then we denote the number of instances where $X = x_i$ and $Y = y_j$ by n_{ij}, which is the number of instances in the corresponding cell of the array. The number of instances in column i, corresponding to $X = x_i$, is denoted by c_i, and the number of instances in row j, corresponding to $Y = y_j$, is denoted by r_j.

given by the fraction of the total number of points that fall in column i, so that

$$p(X = x_i) = \frac{c_i}{N}. \tag{2.2}$$

Since $\sum_i c_i = N$, we see that

$$\sum_{i=1}^{L} p(X = x_i) = 1 \tag{2.3}$$

and, hence, the probabilities sum to one as required. Because the number of instances in column i in Figure 2.4 is just the sum of the number of instances in each cell of that column, we have $c_i = \sum_j n_{ij}$ and therefore, from (2.1) and (2.2), we have

$$p(X = x_i) = \sum_{j=1}^{M} p(X = x_i, Y = y_j), \tag{2.4}$$

which is the *sum rule* of probability. Note that $p(X = x_i)$ is sometimes called the *marginal* probability and is obtained by marginalizing, or summing out, the other variables (in this case Y).

If we consider only those instances for which $X = x_i$, then the fraction of such instances for which $Y = y_j$ is written $p(Y = y_j | X = x_i)$ and is called the *conditional* probability of $Y = y_j$ given $X = x_i$. It is obtained by finding the fraction of those points in column i that fall in cell i,j and, hence, is given by

$$p(Y = y_j | X = x_i) = \frac{n_{ij}}{c_i}. \tag{2.5}$$

Summing both sides over j and using $\sum_j n_{ij} = c_i$, we obtain

$$\sum_{j=1}^{M} p(Y = y_j | X = x_i) = 1 \tag{2.6}$$

showing that the conditional probabilities are correctly normalized. From (2.1), (2.2), and (2.5), we can then derive the following relationship:

$$\begin{aligned} p(X = x_i, Y = y_j) &= \frac{n_{ij}}{N} = \frac{n_{ij}}{c_i} \cdot \frac{c_i}{N} \\ &= p(Y = y_j | X = x_i) p(X = x_i), \end{aligned} \tag{2.7}$$

which is the *product rule* of probability.

So far, we have been quite careful to make a distinction between a random variable, such as X, and the values that the random variable can take, for example x_i. Thus, the probability that X takes the value x_i is denoted $p(X = x_i)$. Although this helps to avoid ambiguity, it leads to a rather cumbersome notation, and in many cases there will be no need for such pedantry. Instead, we may simply write $p(X)$ to denote a distribution over the random variable X, or $p(x_i)$ to denote the distribution evaluated for the particular value x_i, provided that the interpretation is clear from the context.

With this more compact notation, we can write the two fundamental rules of probability theory in the following form:

$$\textbf{sum rule} \qquad p(X) = \sum_Y p(X, Y) \tag{2.8}$$

$$\textbf{product rule} \qquad p(X, Y) = p(Y|X)p(X). \tag{2.9}$$

Here $p(X, Y)$ is a joint probability and is verbalized as ‘the probability of X *and* Y’. Similarly, the quantity $p(Y|X)$ is a conditional probability and is verbalized as ‘the probability of Y *given* X’. Finally, the quantity $p(X)$ is a marginal probability and is simply ‘the probability of X’. These two simple rules form the basis for all of the probabilistic machinery that we will use throughout this book.

2.1.3 Bayes’ theorem

From the product rule, together with the symmetry property $p(X, Y) = p(Y, X)$, we immediately obtain the following relationship between conditional probabilities:

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}, \tag{2.10}$$

which is called *Bayes’ theorem* and which plays an important role in machine learning. Note how Bayes’ theorem relates the conditional distribution $p(Y|X)$ on the left-hand side of the equation, to the ‘reversed’ conditional distribution $p(X|Y)$ on the right-hand side. Using the sum rule, the denominator in Bayes’ theorem can be expressed in terms of the quantities appearing in the numerator:

$$p(X) = \sum_Y p(X|Y)p(Y). \tag{2.11}$$

Thus, we can view the denominator in Bayes’ theorem as being the normalization constant required to ensure that the sum over the conditional probability distribution on the left-hand side of (2.10) over all values of Y equals one.

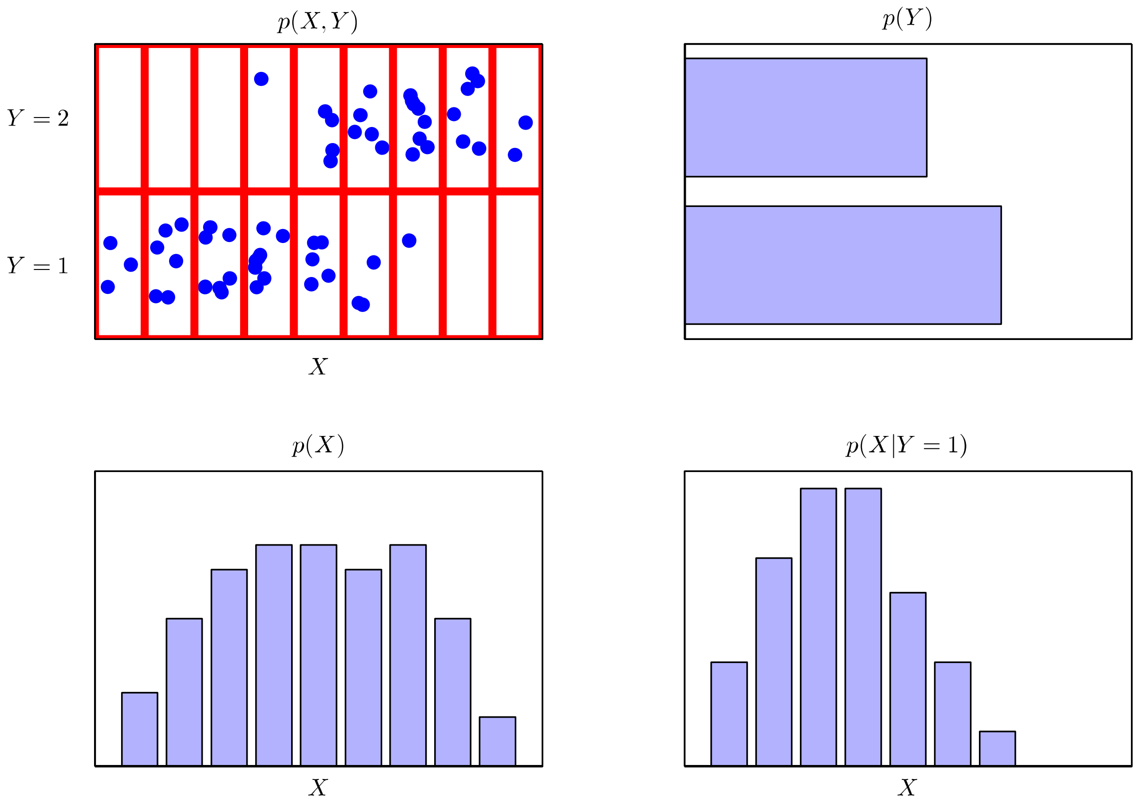


Figure 2.5 An illustration of a distribution over two variables, X, which takes nine possible values, and Y, which takes two possible values. The top left figure shows a sample of 60 points drawn from a joint probability distribution over these variables. The remaining figures show histogram estimates of the marginal distributions $p(X)$ and $p(Y)$, as well as the conditional distribution $p(X|Y=1)$ corresponding to the bottom row in the top left figure.

In Figure 2.5, we show a simple example involving a joint distribution over two variables to illustrate the concept of marginal and conditional distributions. Here a finite sample of $N = 60$ data points has been drawn from the joint distribution and is shown in the top left. In the top right is a histogram of the fractions of data points having each of the two values of Y. From the definition of probability, these fractions would equal the corresponding probabilities $p(Y)$ in the limit when the sample size $N \to \infty$. We can view the histogram as a simple way to model a probability distribution given only a finite number of points drawn from that distribution. The remaining two plots in Figure 2.5 show the corresponding histogram estimates of $p(X)$ and $p(X|Y=1)$.

Section 3.5.1

2.1.4 Medical screening revisited

Let us now return to our cancer screening example and apply the sum and product rules of probability to answer our two questions. For clarity, when working through this example, we will once again be explicit about distinguishing between the random variables and their instantiations. We will denote the presence or absence of cancer by the variable C, which can take two values: $C = 0$ corresponds to ‘no cancer’ and $C = 1$ corresponds to ‘cancer’. We have assumed that one person in a hundred in the population has cancer, and so we have

$$\begin{aligned} p(C=1) &= 1/100 & (2.12)\\ p(C=0) &= 99/100, & (2.13) \end{aligned}$$

respectively. Note that these satisfy $p(C = 0) + p(C = 1) = 1$.

Now let us introduce a second random variable T representing the outcome of a screening test, where $T = 1$ denotes a positive result, indicative of cancer, and $T = 0$ a negative result, indicative of the absence of cancer. As illustrated in Figure 2.3, we know that for those who have cancer the probability of a positive test result is 90%, while for those who do not have cancer the probability of a positive test result is 3%. We can therefore write out all four conditional probabilities:

$$\begin{aligned} p(T=1|C=1) &= 90/100 & (2.14)\\ p(T=0|C=1) &= 10/100 & (2.15)\\ p(T=1|C=0) &= 3/100 & (2.16)\\ p(T=0|C=0) &= 97/100. & (2.17) \end{aligned}$$

Again, note that these probabilities are normalized so that

$$p(T=1|C=1) + p(T=0|C=1) = 1 \tag{2.18}$$

and similarly

$$p(T=1|C=0) + p(T=0|C=0) = 1. \tag{2.19}$$

We can now use the sum and product rules of probability to answer our first question and evaluate the overall probability that someone who is tested at random will have a positive test result:

$$\begin{aligned} p(T=1) &= p(T=1|C=0)p(C=0) + p(T=1|C=1)p(C=1)\\ &= \frac{3}{100}\times\frac{99}{100} + \frac{90}{100}\times\frac{1}{100} = \frac{387}{10,000} = 0.0387. \end{aligned} \tag{2.20}$$

We see that if a person is tested at random there is a roughly 4% chance that the test will be positive even though there is a 1% chance that they actually have cancer. From this it follows, using the sum rule, that $p(T = 0) = 1 - 387/10,000 = 9613/10,000 = 0.9613$ and, hence, there is a roughly 96% chance that the do not have cancer.

Now consider our second question, which is the one that is of particular interest to a person being screened: if a test is positive, what is the probability that the person has cancer? This requires that we evaluate the probability of cancer conditional on the outcome of the test, whereas the probabilities in (2.14) to (2.17) give the probability distribution over the test outcome conditioned on whether the person has cancer. We can solve the problem of reversing the conditional probability by using Bayes' theorem (2.10) to give

$$\begin{aligned} p(C=1|T=1) &= \frac{p(T=1|C=1)p(C=1)}{p(T=1)} && (2.21) \\ &= \frac{90}{100} \times \frac{1}{100} \times \frac{10,000}{387} = \frac{90}{387} \simeq 0.23 && (2.22) \end{aligned}$$

so that if a person is tested at random and the test is positive, there is a 23% probability that they actually have cancer. From the sum rule, it then follows that $p(C = 0|T = 1) = 1 - 90/387 = 297/387 \simeq 0.77$, which is a 77% chance that they do not have cancer.

2.1.5 Prior and posterior probabilities

We can use the cancer screening example to provide an important interpretation of Bayes' theorem as follows. If we had been asked whether someone is likely to have cancer, before they have received a test, then the most complete information we have available is provided by the probability $p(C)$. We call this the *prior probability* because it is the probability available *before* we observe the result of the test. Once we are told that this person has received a positive test, we can then use Bayes' theorem to compute the probability $p(C|T)$, which we will call the *posterior probability* because it is the probability obtained *after* we have observed the test result T.

In this example, the prior probability of having cancer is 1%. However, once we have observed that the test result is positive, we find that the posterior probability of cancer is now 23%, which is a substantially higher probability of cancer, as we would intuitively expect. We note, however, that a person with a positive test still has only a 23% change of actually having cancer, even though the test appears, from Figure 2.3 to be reasonably 'accurate'. This conclusion seems counter-intuitive to many people. The reason has to do with the low prior probability of having cancer. Although the test provides strong evidence of cancer, this has to be combined with the prior probability using Bayes' theorem to arrive at the correct posterior probability.

Exercise 2.1

2.1.6 Independent variables

Finally, if the joint distribution of two variables factorizes into the product of the marginals, so that $p(X, Y) = p(X)p(Y)$, then X and Y are said to be *independent*. An example of independent events would be the successive flips of a coin. From the product rule, we see that $p(Y|X) = p(Y)$, and so the conditional distribution of Y given X is indeed independent of the value of X. In our cancer screening example, if the probability of a positive test is independent of whether the person has cancer, then $p(T|C) = p(T)$, which means that from Bayes' theorem (2.10) we have $p(C|T) = p(C)$, and therefore probability of cancer is not changed by observing the test outcome. Of course, such a test would be useless because the outcome of the test tells us nothing about whether the person has cancer.

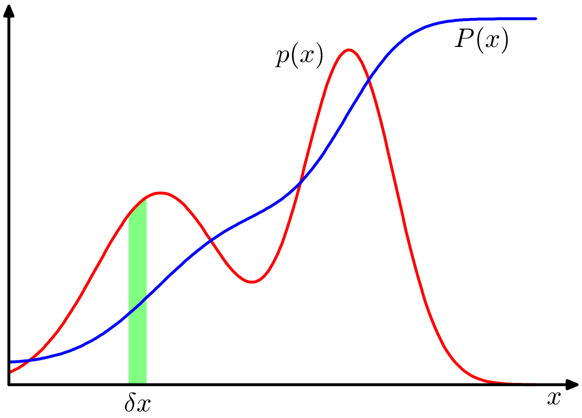


Figure 2.6 The concept of probability for discrete variables can be extended to that of a probability density $p(x)$ over a continuous variable x and is such that the probability of x lying in the interval $(x, x + \delta x)$ is given by $p(x)\delta x$ for $\delta x \to 0$. The probability density can be expressed as the derivative of a cumulative distribution function $P(x)$.

2.2. Probability Densities

As well as considering probabilities defined over discrete sets of values, we also wish to consider probabilities with respect to continuous variables. For instance, we might wish to predict what dose of drug to give to a patient. Since there will be uncertainty in this prediction, we want to quantify this uncertainty and again we can make use of probabilities. However, we cannot simply apply the concepts of probability discussed so far directly, since the probability of observing a specific value for a continuous variable, to infinite precision, will effectively be zero. Instead, we need to introduce the concept of a *probability density*. Here we will limit ourselves to a relatively informal discussion.

We define the probability density $p(x)$ over a continuous variable x to be such that the probability of x falling in the interval $(x, x + \delta x)$ is given by $p(x)\delta x$ for $\delta x \to 0$. This is illustrated in Figure 2.6. The probability that x will lie in an interval (a, b) is then given by

$$p(x \in (a, b)) = \int_a^b p(x)\,\mathrm{d}x. \tag{2.23}$$

Because probabilities are non-negative, and because the value of x must lie somewhere on the real axis, the probability density $p(x)$ must satisfy the two conditions

$$p(x) \geqslant 0 \tag{2.24}$$

$$\int_{-\infty}^{\infty} p(x)\,\mathrm{d}x = 1. \tag{2.25}$$

The probability that x lies in the interval $(-\infty, z)$ is given by the *cumulative distribution function* defined by

$$P(z) = \int_{-\infty}^{z} p(x)\,\mathrm{d}x, \tag{2.26}$$